# BUSINESS STATISTICS

## BLOCKS I & II



Dr. B. R. AMBEDKAR OPEN UNIVERSITY
HYDERABAD
1992

**COURSE TEAM**

Prof. A. Shankariah (Editor)

Dr. V . Gangadhar

Dr. R. Sudarshan

Sri. N . Hanumantha Rao

Sri. V.V. Subrahmanya Sarma

**ASSOCIATE EDITORS**

Prof. V. Nagaraja Naidu

Sri. P. Krishna Rao



Art
Chandra



DR. B. R. AMBEDKAR OPEN UNIVERSITY
Hyderabad

Frist Published in 1984
Revised in 1990
Re-Print - 1992 - 93.



This text forms part of Dr. B. R. Ambedkar Open University Course. The complete syllabus for the course appears at the end of the text.

Further information about the Dr. B. R. Ambedkar Open University Courses may be obtained from the Dir (Academic), Dr. B. R. Ambedkar Open University, Somajiguda, Hyderabad - 500 482 (A.P.).

Printed at M/s. Sruthi Graphics (P) Ltd. 8-3-231/G/4, Sri Krishna Nagar, Hyderabad - 500 045. Phone : 24

## CONTENTS

# PREFACE

This book deals with the topics in Bussiness Statistics included in the syllabus for the second year of the B.Com. course offered by the Andhra Pradesh Open University. These topics generally cover the "core" area of the subject to be studied in the second Year of the Three Year Degree Course in Commerce (B.Com.) . The syllabus for the sake of convenience is divided into Blocks, each of which comprises a number of units. Each Block generally covers a specific area of the subject. The units are prepared by specialists in accordance with a format so designed as to enable the student to read and understand them without much difficulty. Each unit begins with contents followed by Aims and objecitives and has at its end Model Examination Questions intended to test the student's comprehension of its subject matter. Technical terms with which the student may not generally be familiar are given at the end of each unit under the head, "Glossary".

This book is concerned with the study of statistics as applicable to business in which decisions may have to be taken with regard to complex transactions and strategies in the interest of its furtherence and development. Statistics as a subject uses its own tools and techniques to analyse complex phenomena and presents its conclusions in the form of graphs and numerals. It is, therefore, of practical importance as much to businessmen as to students of commerce.

The University hoped that this meterial will help the student to get acquainted with the principal issues in Business Statistics which make for its distinctiveness and significance.

# BLOCK-I : STATISTICS-COLLECTION-CLASSIFICATION AND PRESENTATION

## UNIT-1 : ORIGIN, GROWTH AND DEFINITION OF STATISTICS

**Contents**





## 1.0 AIMS AND OBJECTIVES

This unit aims at acquainting the students with the nature and growth of statistics.

After reading this unit you would be able to :

i explain the origin and growth of statistics

ii define the term statistics and

iii distinguish between statistical data and statistical methods.

## 1.1 INTRODUCTION

In the modern age, Statistics has become an integral part of human activity. Every aspect of human activity is counted or measured and interpreted with the help of statistics. Statistics refers to information about an activity or a process whether it be production, sales, population, national income, etc., which is described with the help of numbers. Statistics is a body of methods; it helps in taking decisions. Statistics provides necessary methods and techniques for the collection and simplification of large mass of data. It also provides a means for presenting the data in a systematic and meaningful form.

In business, Statistics serves as an informational segment of decision making process. Management at all levels is guided by facts obtained through the analysis of past records. A systematic analysis of past records helps the management in predicting the future trends. Statistical methods also help business management in measuring and evaluating the current accomplishments.

## 1.2 ORIGIN AND GROWTH OF STATISTICS

The word 'Statistics' has been derived from the La word "STATUS", the Italian word "STATO" and the German word "STATISTIK". The meaning of all these words is 'Political State' or the 'Statesman's art'. Thus, the origin of Statistics was due to administrative requirements of the state. In all the countries of ancient culture, there is evidence to show that they had some system of collecting statistics. In ancient Egypt, the State prepared registration lists of all the heads of families. In ancient Judea, a census of the population was taken on several occasions. In 2030 B.C., the population of Judea was estimated as 38,00,000. In Rome, the first census was taken in 435 B.C. In the middle ages, surveys in respect of taxation, military service, tithes (a tenth part of a person's income paid as church tax), imports and exports were conducted in various countries. All these surveys were conducted for political purposes such as raising new taxes to meet war requirements and to assess the relative strengths and weaknesses of military services. Soon after William the Conquerer became the king of England, he ordered a survey of all the lands of England for the purposes of taxation and military service. The results of this survey were recorded in "Domesday Book of 1086 A.D." In the beginning of the sixteenth century, a number of statistical book-lets were brought out mainly containing descriptive statistics.

In the beginning of the sixteenth century, statistical methods were used in physical sciences only. Astronomers used to record the movements of stars and planets to foretell their position and to make forecasts of eclipses. It is on the basis of these statistics that Sir Isac Newton formulated his famous theory of Gravitation. Subsequently statistical methods were also used in social sciences like Politics, Economics and Sociology.

During the seventeenth century, statistical methods were used under the banner of political arithmetic. Captain John Graunt of London (1620-1674) known as the father of vital statistics, studied statistics of births and deaths. In the year 1662, he published his work on "Observations on the London Bills of Mortality", which was the first work on social statistics. In this book, emphasis was laid on collection and study of statistics. Inspired by the works of Graunt, his French friend, William Petty (1623-1687) wrote a book entitled "*Essays on Political Arithmetic*". In his book, Petty wanted to achieve accuracy by means of calculation, even in the absence of empirical facts. The development of political arithmetic led to the birth of the subject of Economics. During the same period, preachers of the protestant churches collected figures in respect of births, deaths and marriages with a view to checking illegitimacy. During this period, Edmund Hally prepared the first life table giving the expectation of life at each age. His work was based on the data collected by Casper Newman in 1691, relating to death records of Breslau. Later on, James Dodson, Thomas Simpson, Dr. Price and others also prepared mortality tables. It

was during this period, that the idea of Life Insurance was developed.

J.P. Sussmilch (1707-1767), a Prussian Clergyman, statistically explained the theory of Natural Order of Physiocratic School which, according to him, is a kind of natural law. He propounded a doctrine in which he had stated that the ratio of births and deaths remains more or less constant.

Bernoulli (1654-1705) was the first person to state the law of large numbers. L.A.J. Quetlet (1796-1874) made a significant contribution to the modern theory of statistics. he propounded the concept of ' average man' . He stated that the actions of an average man conform to the average results obtained from society. Any deviations from the theoretical average were capable of being treated by the method of errors and probability. He also emphasised the importance of the 'Law of Large numbers' which was founded by Bernoulli.

Bernoulli and his nephew Daniel Bernoulli (1700-1782) laid a solid foundation for the theory of probability and put forward the idea of normal expectation. In fact, G.Cardano (1501-1536), who was a great mathematician as well as big gambler, wrote a valuable treatise on the hazards of the game of chances , in which he had formulated certain rules for minimising the risks of gambling. Subsequently, Laplace (1749-1827) published his works on the theory of probability. His works gained recognition as one of the best ever done on the subject of probabililty. The present body of statistical methods, particularly those concerned with drawing inferences about a population from the set of observations that make up a sample, is based on the theory of mathematical probability. Some of the renowned mathematicians like Galileo, Blaise Pascal, De Moivre and Farmet also helped in the development of modern statistics.

In Germany, systematic collection of statistics by the state started during the end of the eighteenth century. In India also the systematic collection of statistics had been an age old tradition. Megasthanese had given an account of method of collecting data in respect of revenue and expenditure, births and deaths, military, land, etc., during the Chandragupta's regime. Similarly Koutilya's "Arthasastra" expounds the principles of collection of data relating to revenues and expenses. During Mughal period, Statistics were collected and the system of collection was described in '*Tuzuk-i-Babari*' and '*Ain-i-Akbari*'. During Akbar's time, Raja Todar Mal collected land statistics for fixing land revenue. Such statistics were purely descriptive in nature.

Early in the nineteenth century, statistical records were prepared which described the economic and social problems of that era. Statistics collected in those days were mostly descriptive in nature and had limited sphere of activity. In the present century, statistics are used to solve problems and to determine courses of action.

Till the second half of the nineteenth century, the use of statistics was limited either to the requirements of governments or to social problems. But now there is hardly any field or branch of knowledge which does not make use of statistical methods. This is because of the increasing data needs of business, government and science. The technological revolution that has taken place in the field of data handling also necessitated scientific revolution in statistical theories and techniques.

During the last three centuries, the most notable among the scholars who have contributed to the development of the Science of statistics were Carl Friedrich Gauss (1777-1855), Francis Galton (1822-1911), Adolphe Quetelet (1796-1874).

Karl Pearson (1857-1937) was instrumental in developing correlation and regression formulae that are widely used even today. William S.Gosset (1876-1937) who was a student of Karl Pearson, formulated a number of statistical formulae under the name of 'Student'. R.A.Fisher is another great scholar who propounded many concepts which are useful to draw conclusions from the statistical data. Although, many of the scholars contributed to the science of modern statistics, Fisher and Karl Pearson are considered to be the real giants in the development of the theory of statistics.

In the field of Economics, Augustin Cournot (1801-1877), Leon Walras (1834-1910) Vilfredo Pareto (1848-1923), Alfred Marshall (1842-1924), Edgeworth and A.L.Bowley are some of the notable names who did a commendable job in developing the science of statistics. They gave an applied form to the description of statistics in developing quantitative approach to economic problems.

In India, Prof. P.C.Mahalanobis, Dr. V.K.R.V.Rao, R.C.Desai and Dr.P.V.Sukhatme are notable among the economists who have made considerable contribution to the applied field of statistics.

A cursory glance of the preceding pages reveals that the science of statistics is said to have developed from three main sources, viz., government records, mathematics and political economy.

Although statistics orginated as the science of kings, there has been a remarkable and sustained growth in the theoretical and applied field of statistics . Statistics is now regarded as one of the most important tools for taking decisions in the midst of uncertainty. The following two factors are considered to be responsible for the development of the modern statistics :

i. Increased demand for statistics.

ii. Decreasing cost of Statistics.

*i. Increased demand for statistics*: In the ancient culture of various countries, the use of statistics was limited to the conditions of society only. Maintenance of law and order was the primary task of the government. Government did not interfere in the economic matters of the country. Today, there is hardly any field in which government does not interfere. With the enlargement of government functions, the demand for statistics has also increased. Extensive research work is also being undertaken more now than what was a century ago. Since statistics is considered a tool of research, the demand for statistics has increased considerably.

*ii. Decreasing cost of statistics* : The cost in terms of time and money involved in the collection of data is the limiting factor in the use of statistics. With the development of sampling techniques and the electronic devices, such as calculators, computers, etc., the cost of collecting,

processing, analysing and interpreting the data has gone down considerably. This has facilitated the increasing use of statistics in solving various problems.

## 1.3 MEANING OF STATISTICS

The word 'Statistics' has many meanings. Usually Statistics is regarded as data or the numerical measurements of a phenomenon. Some people regard it as a study of figures, while others view it as a diagrammatic or graphic presentation of facts. Some people also regard statistics as an analysis of figures for drawing conclusions which are useful for forecasting. In brief, the following three forms of statistics are widely in use:

i. Statistics as a product

ii. Statistics as a process

iii. Statistics as an application.

*i. Statistics as a product :* Statistics as a product referes to the statistical data. It is the aggregate of numerical facts of a phenomenon. For example, we talk about statistics of national income, sales, production, births, deaths,etc. In this context, the term statistics is understood in a plural sense.

*ii. Statistics as a process :* Statistics as a process refers to statistical methods. Statistical methods are tools and techniques which aid the statistical investigation. The main phases of a statistical investigation are collection, organisation, presentation, analysis and interpretation of data. Here statistics is understood in a singular sense. Statistical methods also refer to the study and research of statistical principles which would enlarge and englighten the knowledge of all those who use those methods to take decisions in the face of uncertainty.

*iii. Statistics as an application:* Statistics in singular sense is a measure of the sample. For example, mean of a sample is known as statistics. The corresponding measure of the Universe is called parameter. Thus mean of the universe is a parameter. Measurements of sample like mean, median and mode viz., statistics are used to estimate the parameters of the universe.



## 1.4 DEFINITION OF STATISTICS

Many statisticians have defined 'Statistics' in different ways. While some of them defined statistics as 'Statistical Data' (in a plural sense), others defined it as 'Statistical Methods' (in a singular sense). Some of the important definitions are given below:

**A) Statistics defined as 'Statistical Data:**

i. Bowley defined statistics as "numerical statements of facts in any department of enquiry placed in relation to each other".

According to this definition, statistics are numerical statements of facts and are concerned with an enquiry. Such numerical statements must be placed in relation to each other in order to facilitate comparison. In the opinion of Bowley, all statistics are numerical facts, but all numerical facts are not statistics. His definition just covers the term enquiry, but it does not speak about

the other statistical methods. It also does not specify the nature of enquiry and nature of facts.

ii. According to *Yule and Kendall,* " Statistics are quantitative data affected to a marked extent by a multiplicity of causes".

As per this definition, quantitative data is called statistics. Such quantitative data is subject to the influence of various inter-related and inter-dependent operating forces. But the definition does not specify the field to which statistics relate.

iii. Tuttle defined Statistics as "measurements, enumerations or estimates of natural or social phenomena, usually systematically arranged, analysed and presented as to exhibit important relationships among them".

According to Tuttle, statistics relate to natural or social phenomena. Such phenomena will have to be measured and measurments can be either estimates or enumeration. Statistics are systematically arranged, analysed, and presented in such a way as to facilitate comparisons.

iv. Connor defined Statistics as "measurements, enumerations or estimates of natural or social phenomena, systematically arranged so as to exhibit their interrelations".

This definition resembles the definition of Tuttle. The only difference between these two definitions is that the latter does not include the words analysis and presentation of data.

All the above definitions are incomplete as their coverage is confined to certain aspects only. The definition given by Professor Horace Secrist is considered to be the comprehensive one. According to him " Statistics are aggregates of facts affected to a marked extent by multiplicity of causess, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a pre-determined purpose and placed in relation to each other".

A close examination of this definition reveals the following characteristics that statistics should possess:

*a) Statistics are aggregates of facts*

Aggregates of facts relating to a phenomenon which are comparable with other related aspects can only be called statistics. Single and isolated figures relating to purchases, sales, production, etc., are not statistics.

*b) Statistics are numerical statements*

Statements which do not contain numbers are not called statistics. Even qualitative statements which contain numbers can be called statistics but statements like, " India is a poor country", "India is attaining self-sufficiency in the production of food items" are not called statistics. On the other hand, " the output of foodgrains in 1983-84 increased to 175 million tonnes from 150 million tonnes in 1982-83" is called a statistical statement.

*c) Statistics are affected to a marked extent by multiplicity of causes*

To be statistics, facts and figures relating to a phenomenon must be subject to the influence of various operating forces. It is very difficult to isolate the impact of any factor on a given event. For example, wheat production in a given year is influenced by rainfal,fertility of soil, quality of seeds used, manures used and the method of cultivation. It is very difficult to segregate and study the effect of each of these factors individually on the production of wheat.

*d) Statistics are collected, arranged, analysed and presented systematically*

Statistics are collected systematically according to a detailed plan formulated before the data is collected. Statistics collected in a hasty and haphazard manner give misleading conclusions and mislead the statistical investigator. Data collected should be processed, analysed and arranged in suitable form to facilitate the decision-making process.

*e) Statistics are collected for a pre-determined purpose*

The purpose of collecting statistics should be defined clearly before they are collected. Statistics collected without any pre-determined purpose do not serve any useful purpose. For example, if the investigator plans to collect statistics on prices, he must clearly define the purpose of collecting such statistics so that the type of prices, whole sale prices or retail prices and the commodities to be included can be decided.

*f) Statistics are enumerated or estimated according to reasonable standards of accuracy*

Statistics about a given phenomenon can be collected in two ways, viz.,enumeration and estimation. Enumeration refers to the method of survey according to which data are collected for each and every unit of the universe by actual counting. Estimation refers to the prediction of values for a universe on the basis of general surveys. Estimates are not as precise and accurate as enumeration studies. However, in either case 100% accuracy is seldom possible. The degree of accuracy desired depends upon the nature and purpose of study. Reasonable standards of accuracy must, however, be attained, otherwise data may be altogether misleading.



*g) Statistics are placed in relation to each other*

Statistics must permit comparison of two or more related aspects. Comparison is made either period-wise or region-wise. Data relating to a phenomenon of a period can be compared with the data of the same phenomenon for different time periods. Similarly, data relating to a phenomenon of a region can be compared with the data of the same phenomenon of a different region for the same period. However, while comparing things, *uniformity* or *homogeneity* of data must be ensured. It is meaningless and futile to compare two or more unrelated and heterogeneous aspects.

**B) Statistics defined as Statistical Methods**

i. According to Bowley "Statistics is the science of the measurements of social organism regarded as a whole in all its manifestations."

As per this definition, Statistics is a method of measuring social phenomena. Thus the measurement is confined to social sciences only. Further, this definition includes only one aspect of statistical method i.e., measurement. It does not speak about the other methods such as analysis, presentation and interpretation of data.

ii. Boddington defined Statistics as " the science of estimates and probabilties".

According to him, statistics is confined to estimates and probabilities which constitute only a part of statistical methods.

iii. W.I.King defined Statistics as "the method of judging collective natural or social phenomena from the results obtained by the analysis of an enumeration or collection of estimates".

According to him, Statistics is a method of interpreting the data. Such data may relate to natural or social phenomena. Data might have been collected either by the method of estimation or by enumeration. This defintion gives more importance to interpretation of data when compared to other statistical methods such as collection, analysis and presentation.

iv. Professor M.G.Kendall defined statistics as "the brach of scientific method which deals with the data obtained by counting or measuring properties of population of natural phenomena".

According to him, the use of Statistics is restricted to the field of natural sciences. Further, it deals with only one of the statistical methods i.e., collection of data. He does not foresee the use of statistics to the field of social sciences.

All the above definitions are incomplete as some of them deal with social phenomena, whereas others deal with natural sciences. Further, none of them deals with the statistical methods. The following definitions are considered to be comprehensive because of their coverage of field and methods.

a) Professor Lovitt States that "Statistics deals with the collection, classification and tabulation of numerical facts as the basis for explanation, descripition and comparison of phenomena."

b) F.E.Croxton and D.J.Cowden defined statistics or statistical methods as " the collection, presentation, analysis and interpretation of numerical data."

c) According to R.H.Wessel, E.R.Willett and A.J.Simone "Statistics is the science that deals with the analysis of masses of quantitative data. It includes the collection, classification, summarisation, presentation and interpretation of such data."

d) Neter and Wasserman defined statistics as "the body of techniques or methodology which has been developed for the collection, presentation and analysis of quantitative data and for the use of such data in decision making".

A cursory glance at these definitions reveals that Statistics is undoubtedly a science. It deals with a body of methods or techniques such as collection, presentation, analysis and interpretation of data. However, none of the definitions focus on one of the important statistical methods i.e. organisation of data. If organisation of data is also considered, Statistics can be defined as the science of collection, organisation, presentation, analysis and interpretation of numerical data.

According to this definition, statistical methods include five stages:

(i) *Collection:* Collection of data is the first stage of any statistical investigation. Before proceeding for the collection of data, certain aspects like purpose of the study, sources of data, methods of collection and the degree of accuracy desired will have to be planned and specified. After this, data are to be collected systematically. If data are collected haphazardly without any plan, conclusions drawn on the basis of such data will be unreliable and misleading. Sometimes, the data required may be readily available from the existing published and unpublished sources. The investigator must make use of such facility rather than undertake the unnecessary trouble of collecting data afresh which invloves much time and money.

(ii) *Organisation:* Data collected from published sources are usually in an organised form. Mass data collected from a survey is usually in an unorganised form and requires organisation. The first step in the organisation of data is editing, i.e., careful scrutiny of data. After the scrutiny, data must be classified according to some common characteristics of the items. After completing the scrutiny and classification, data are to be tabulated. In this stage data are arranged in suitable columns and rows. The basic object of tabulation is to ensure clarity in the data presented.

(iii) *Presentation:* Presentation of data in a suitable form with the help of appropriate tables is an important stage of statistical investigation. Diagrams and graphs can also be used for presenting the data. Presentation of data through diagrams and graphs attracts the attention of the readers. The basic object of presentation of data is to help statistical analysis.

(iv) *Analysis:* The basic object of analysis of data is to establish relationships among the related variables. For the purpose of analysis of data, simple statistical techniques such as measures of central tendency, measures of despersion, skewness, etc., can be used. Complicated techniques such as statistical decision theory, correlation and regression analysis can also be used for analysing the data.

(v) *Interpretation:* The last stage in the statistical investigation is interpretation of data. This involves drawing conculsions impartially and objectively from the data collected. Interpretation of data requires a high degree of skill and experience. Wrong interpretation of data results in misleading conclusions which defeat the very purpose of statistical investigation.

**Cheack your progress - 1**

Explain the term 'Statistics'

**Check your progress - 2**

Explain the various stages of statistical methods briefly

## 1.5 STATISTICAL DATA VS. STATISTICAL METHODS

It is observed eariler that the word statistics in plural means data and in singular it means methods. Therefore Statistical data refer to the numerical facts of a phenomenon and aggregates of facts. Statistical methods refer to the tools and techniques applied to process the statistical data which are raw in nature. Some of the points of distinction between statistical data and statistical methods are given below:

| | *Statistical Data* | *Statistics as Methods* |
|---|---|---|
| (i) | Used in plural sense | Used in singular sense |
| (ii) | Refer to numerical Statements of facts. | Refers to various phases of a statistical investigation. |
| (iii) | Descriptive in nature. | Analytical and operational in nature. |
| (iv) | Provide data which are raw in nature | provides necessary tools and techniques to process the statistical data. |
| (v) | Statistical data, as such, do not serve any useful purpose unless it is further processed. | Statistical methods without the availabilty of appropriate data are useless and do not serve any useful purpose. |
| (vi) | The nature and purpose of Collected data dictate the application and choice of Statistical methods. | Collection of data is guided by the availability of suitable and appropriate statistical methods to process the data. Otherwise desired results cannot be attained with the data collected. |

From the above distinction, it is clear that mere collection of data does not serve any purpose.In order to draw valid conclusions, collected data must be organised, presented and analysed systematically. This is referred to as the processing of data which involves the application of statistical methods. Thus statistical data and statistical methods, considered individually, do not serve any purpose. They are complementary to each other.

## 1.6 SUMMING UP

The subject of statistics emerged to meet the administrative requirements of the State. We have evidence to show that the rulers of ancient time had some system of collecting statistics in respect of births, deaths, crops, imports and exports. All these statistics were descriptive in nature and were used mainly for political and administrative purposes. Till the second half of the nineteenth century, the use of statistics was limited either to meet the requirements of the governments or to the study of social problems. But now, there is hardly any field or branch of knowledge which does not make use of statistics. The ever increasing demand for statistics and the decreasing cost of statistics are responsible for the development of modern statistics. Some Statisticians have defined the term 'statistics' as statistical data (in plural sense) and as a body of statistical methods (in singular sense). Statistics defined as statistical data refer to the aggregates of facts affected by various operating forces, expressed in numerical form and collected systematically for pre-determined purposes with reasonable standards of accuracy. statistics, defined as statistical methods, refers to a body of methods or techniques such as collection, organisation, analysis, presentation and interpretation of data.

## 1.7 CHECK YOUR PROGRESS:MODEL ANSWERS

1. The term statistics can be explained both in singular and plural senses. In singular sense it refers to statistical methods and in plural sense it refers to statistical data.
2. List out the number of stages as shown below, and give a brief description on each of these stages.
   i) Collection
   ii) Organisation
   iii) Presentation
   iv) Analysis
   v) Interpretation.



## 1.8 MODEL EXAMINATION QUESTIONS

**A: Short Questions**

1. Comment on the following:
   (i) Statistics are aggregates of facts
   (ii) Statistics is a science of estimates
   (iii) Statistics is a scientific method
   (iv) statistics is a science of averages
   (v) Statistics is a science of kings
   (vi) Statistics is not an exact science

## 2.2 SCOPE OF STATISTICS

Statistics which originated as statesman's art, has developed in such a way that, today, its principles and methods have become indispensable to almost all branches of knowledge, Physical Sciences, Natural Sciences, and Social Sciences like Commerce or Economics. The use of statistics is ever increasing on account of its principles and methods which are appropriate for handling data that are subject to variations.

In order to find solutions in the areas of physical, natural and social sciences, statistics is widely used.

Statistics as a body of scientific methods helps in planning and designing the various procedures involved in the setting up and testing of hypotheses. While the setting up of hypotheses involves the description of relations among the data, the testing of hypotheses involves judging the accuracy and relevancy of predictions that are based on statistical data. It means that the statistical methods and procedures help in establishing and extending knowledge though they are characterised by imperfection.

The observations of Croxton and Cowden help us in understanding the subject matter of statistics. According to them,"The methods of statistics are useful in ever widening range of human activities in any field of thought in which numerical data may be had". Further, they say "Today, there is hardly a phase of endeavour which does not find statistical devices atleast occasionally useful".

The use of statistical data and statistical methods have assumed greater importance in the present day context of human life as they have become indispensable in every phase of human activity. Statistical principles and methods deal with the quantitative characteristics of objects.

## 2.3 IS STATISTICS A SCIENCE OR ART?

Before deciding whether Statistics is a science or an art or both, it is necessary to understand the meaning and characteristics of science and Art. Webster's Dictionary defines Science as an "accumulated and accepted knowledge that has been systematised and formulated with reference to the discovery of general truths or the operation of general law". In brief, Science refers to a systematic body of knowledge and studies cause and effect relationship among various aspects. It enlarges and sharpens the knowledge of individuals and helps to draw generalisations on the basis of empirical evidence. Such generalisations are known as laws of science. According to Karl Pearson, science must have certain distinguishing features such as providing mental education to the citizens, throwing light on social problems, giving happiness in practical life and providing an opportunity of satisfaction to artistic faculties.

Statistics as a branch of knowledge throws light on various economics and social problems. It provides the necessary tools and techniques to solve day-to-day problems of man kind. Its methods are widely used in business and economics whose main aim is to promote the welfare of human beings. Statistical methods provide useful techniques to find out cause and effect relationship among various aspects. Without statistical data and statistical methods, it would be difficult for the government to formulate and implement various welfare programmes aimed at providing a happy and peaceful life to its citizens. Thus, statistics fulfils all the conditions prescribed for a science. Hence, statistics can be called a science. However, statistics is not an exact science like the Physical Sciences. Hence, it is appropriate to call it a science of scientific methods 'or a specialised branch of knowledge'.

Art refers to the skill of handling facts in achieving a given objective. It is concerned with the skills associated with the presentation and handling and drawing valid inferences. Art is a synthesis of activities directed towards the solving of a problem. Statistics fulfils all these conditions as statistics by themselves do not serve any purpose unless they are systematically

collected, classified, analysed and presented. Statistics provides necessary tools and techniques but it is upto the statistician to use them properly. Hence, statistics can be called an art.

Statistics not only provides the necessary principles and techniques, but also guides us to achieve the desired objectives with them. Statistics is regarded as an art of applying the science of scientific methods. Hence, we can conclude that statistics is both a science and an art. In the words of Tippett, "Statistics is both a science and an art. It is a science in that its methods are basically systematic and have general application and an art in that their successful application depends, to a considerable degree, on the skill and special experience of the statistician, and on his knowledge of the field of application".

## 2.4 FUNCTIONS OF STATISTICS

According to Robert W. Burgess, statistics replaces ignorance, prejudice, rule of thumb, arbitrary and premature decisions, tradition and dogmatism with quantitative analysis and scientific decision making. The science of statistics enlarges our knowledge and provides a scientific approach in solving problems. The evergrowing importance of statistics can be understood from the functions it performs. They are explained below:

*(i) Simplifies mass data*

*(ii) Presents facts in a definite form*

*(iii) Facilitates comparative study*

*(iv) Helps in predictions*

*(v) Helps in formulating and testing hypothesis*

*(vi) Helps formulation of suitable policies*

*(vii) Enlarges individual knowledge and experience*

*(viii) Tries to interpret condition*

*(ix) Provides numerical measurement*

*(i) Simplifies mass data*

It is very difficult to read, understand and remember mass data. Statistical data as such do not help the managers in decision-making. Statistical methods provide the necessary means to condense mass data and present them with the help of single figures such as averages, rartios, variations, skewness, coefficients, etc. The single figure represents the whole data and they at once become more intelligible and understandable. It avoids any ambiguity. It is easy to understand and simple to follow. Readers can understand quickly the significant characteristics of the numerical data. Diagrams and graphs are also used to present the data in a more appealing manner.

*(ii) Presents facts in a definite form*

One of the most important functions of statistics is to present data in a precise and defininite form. Numerical expression of facts is some times vague Facts presented in a definite quantitative form are more convincing and enable the readers to understand the phenomena without any difficulty. For example, "enrolment of unemployed graduates in 1984 is expected to be more than that in 1983". This statement does not convey any clear-cut idea about the magnitude of the problem. On the other hand, if the problem is quantified and stated in numerical form, it conveys definite information about the problem. For example, "the enrolment of unemployed graduated in 1984 is expected to increase to 1.5 lakhs from 1.25 lakhs in 1983".

*(iii) Facilitates comparative study*

Figures, by themselves, do not convey any meaning unless they are compared with some other related facts. Statistics provides necessary tools and techniques to compare the facts.

According to A.L. Bowley, the chief practical use of statistics is to show the relative importance of various aspects under study. Statistical devices, like averages, ratios, percentages, variations, standard error, coefficients, graphs and diagrams are some of the important tools and techniques used to compare the data. Comparison of data enables us to know the changes that have taken place over a period of time or changes between two geographical locations. Thus, the analysis of the past enables us to study the impact of changes on the future.

*(iv) Helps in predictions*

Statistics not only helps in analysing the past data, but also helps in estimating and forecasting the future. The functions of bussiness management start with planning which involves estimating various aspects that will take place in future. Statistics furnishes useful techniques to predict the future values scientifically. Analysis of time series, regression analysis, extrapolation, etc.,are some of the important statistical techniques that are helpful in forecasting future events. Forecasting of future events is followed by proper analysis and understanding of the behaviour of past data.

*(v) Helps in formulating and testing hypothesis*

This is the most important theoretical function of statistics. Statistical methods are not only useful in testing and formulating the hypothesis but also helpful in discovering new theory. According to Prof.J.M.Keynes, the basic function of statistics is to suggest empirical laws which may or may not be capable of subsequent deductive explanation. It also supplements the empirical laws with deductive reasoning by checking its results. Statistical methods help in testing the correctness of the laws of the different branches of knowledge.

*(vi) Helps formulation of suitable policies*

Statistical analysis is extremely useful in evaluating the efficiency of the existing policies and practices. It also helps in formulating suitable policies in different areas such as social, economic and business fields. Statistical data and statistical methods help the government in formulating suitable policies in respect of taxation, import-export and socio-economic welfare programmes.

*(vii) Provides numerical measurements*

Statistics provides necessary tools and techniques to measure the various characteristics of a phenomenon. Some of the characteristics can be quantified without any difficulty but some of the characteristics which are qualitative in nature are not subject to direct quantification. But statistical methods provide useful techniques with the help of which even qualitative data are measured and expressed in numerical terms.

*(viii) Enlarges individual knowledge and experience*

Statistical methods help in finding out solutions to day-to-day problems of mankind in every field. Statistics sharpens rational thinking and reasoning. It helps in formulating new theories and concepts. In fact, many fields of knowledge would have ever remained closed to mankind, without the efficient and useful techniques of the science of statistics. "The proper function of statistics", says A.L.Bowley,"is to enlarge individual experience".

*(ix) Tries to interpret conditions*

Statistics helps in interpreting conditions by identifying possible causes for the results described. In a manufacturing concern, if a machine is found not upto the mark in turning out articles, the production manager can find out the reasons with the help of statistical techniques. Such a condition may be due to some defect which is inherent in the machine itself or may be due to faulty handling of the machine by the operator.

**Check your progress - 1**

What are the statistical tools used to simplify the mass data ?

## 2.5 IMPORTANCE OF STATISTICS

Now-a-days statistical methods and principles are widely used in various fields of knowledge. Consciously or unconsciously, people use the science of statistics to solve their day-to-day problems. It is more so in fields like business, economics and public administration. H.G. Wells remarked that, "statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write". Today his remarks have become true. Some of the practical applications of statistical methods are mentioned below:

### 2.5.1. TO BUSINESS

Success of business depends upon efficient management of production and sales which in turn depend on location, size, marketing and quality control, etc.

*(i) Determination of Location:* Before taking a decision regarding location of a business firm, proper data in respect of the availability of raw-material and labour must be collected, analysed and interpreted systematically. Decisions taken on scientific lines yield good results and help the organization to flourish. Data in respect of local taxes, incentives offered by various state governments should also be collected. A camparative study of all these things guides the promoters in deciding the right choice of location.

*(ii) Size:* Statistical methods help in comparative study of cost-benefit analysis at various hypothetical levels of operation and guide the management in deciding the right choice of size. Sales is the major source of revenue to any business organization. Sales revenue depends upon two factors. viz.,(a) quantity sold, and (b) selling price per unit. Both these factors are guided by the size and scale of operations of the business enterprise. Cost of production forms the basis for fixing the selling price.

*(iii) Marketing decisions:* Marketing decisions include decisions in respect of demand forecasting, product planning, pricing, promotional methods, distribution methods and market survey and market research.

Statistical methods provide the scientific means to collect, analyse and interpret data relating to study of demand . This enables the management to ascertain accurate sales forecasts on the basis of which various business programmmes can be finalised. Study of time series analysis is useful in identifying the seasonal and cyclical variations in demand. The techniques of business forecasting such as extra-polation, regression analysis etc.,are useful guides to predict future economic events.

Products should be designed and produced only after undertaking a detailed consumer survey. Such surveys help the producers in finding out the tastes and preferences of consumers. Today consumers are playing a pivotal role in the determination of product design and type of products to be undertaken by the manufacturers. Statistical methods help in taking decisions in respect of adding or droppong certain products on the basis of their cost-benefit analysis.

To avoid wrong pricing policies, pricing decisions must be based on the analysis and study of cost data provided by the records of the organisation and also on data collected in respect of competitors. Pricing is an important area where managers should show their wisdom with the help of statistical methods.

Regression and correlation techniques are widely used in finding out the relationship between different promotional methods and sales. With the help of these techniques, the efficacy and justification of incurring expenditure towards advertisement can be judged.

A comparative study of various distribution methods helps the management in the following areas:

*(iv) Production planning:* Statistical methods help the management in taking appropriate decisions in respect of production planning and inventory control. Accurate sales forecast help the management in regulating production planning. Production planning involves the finalisation of production schedules with respect to various commodities. On the basis of production plans, raw-material requirements can be determined and procured on scientific lines. This enables the management to exercise proper control over inventories. Inventory control aims at maintaining optimum stock levels, and minimise the evil effects of over-stocking and under-stocking.

*(v) Quality control:* Statistical methods are useful tools in the hands of management to ensure quality control. Majority of moderately large organizations are maintaining their own statistical quality control department. Today, the business world is facing severe competition. In order to withstand competition, managements should try hard to provide quality goods and services at reasonable prices.

*(vi) Personnel Administration:* Statistical methods are useful in evaluating and appraising the performance of employees on the basis of which employees incentive schemes can be introduced. Statistical methods help in testing the usefulness of a person to the business in terms of his interest, capacity to work, etc. Statistical tests have been designed to assess such psychological aspects of the employees.

*(vii) Accounts and Auditing:* To assess the profitability and financial position of busines, income statement and balance sheet are prepared respectively. While preparing the income statement, items of various expenses and incomes are classified according to their nature and presented in a systematic manner. Similarly grouping of various assets and liabilities in the balance sheet under different headings is also attempted in accordance with the statistical principles of classification.

Financial statements are prepared on the basis of the historical data provided by the accounting records. Such data do not reflect the present values of various items of assets, liabilities, expenses and income. Hence, with the help of such statements, the real worth of the business cannot be ascertained. But with the help of statistical techniques, such as index numbers, the historical figures can be adjusted and restated enbling the readers to assess the true and fair value of the business. In accounting, financial and other ratios are calculated to understand the relationship among various related variables.

The basic object of auditing is to ensure the accuracy and reliability of accounting data. This is done through verification of accounting data. Verification of each and every business transaction is tedious and involves unnecessary costs. In this connection, the science of statistics provides useful devices i.e. sampling techniques. Auditors select only a few items at random and subject them to test audit. On the basis of the results of test audits, the reliability of all the transactions is determined.

*(viii) Operations Research:* Operations Research techniques are mainly concerned with profit-maximisation and cost minimization. These techniques are useful in determining the optimum stock levels and also in deciding the economic order quantities. The theory of probability is extensively used in these decision areas. Operations Research techniques such as linear programming , Games theory, Critical path method (CPM) and Programme Evaluation and Review Technique (PERT) which are widely used for decision making heavily depend on statistical methods.

## 2.5.2 TO BANKERS, BROKERS AND INSURANCE COMPANIES

Statistical techniques like ratios, percentages, dispersion, etc., are extensively used by the bankers for the evaluation of the credit worthiness of their customers. Knowledge of probability will be useful to the brokers who are engaged in the sale and purchase of stocks and shares. Probability theory helps them in appraising the expected return on various investment portfolios. With the help of statistical methods, insurance companies calculate finacial risk on current life policies, which would mature on future dates. This is done on the basis of life expectancy studies of the policy holders in different age groups. Statistical methods also help the insurance companies in constructing the mortality tables on the basis of which premium rates are fixed.

## 2.5.3 TO ECONOMISTS

The impartance of statistics in Economics will be clear from the observations of the renowned Economist, Prof.A. Marshall. He said "statistics are the straw out of which I, like every other economist, have to make bricks". The specific uses of statistical methods in economics are detailed below:

*(i) Study of Economic Problems:* One of the most important functions of economists is to collect data in respect of demand and supply conditions, income levels of different classes of society, consumption pattern of people, standard of living, inequalities in income, savings and investment levels, problems of unemployment, etc. These economic problems can be understood in their correct perspective as they are subjected to statistical analysis.

*(ii) Formulation of Economic Laws:* Economic laws are based on inductive reasoning which involves careful study of the economic behaviour of a large number of units. For example, Engel's law of consumption was based on the study of family budgets in a town. Similarly, the Law of Demand and the Elasticity of Demand are also based on inductive reasoning. Statistical methods are also useful in testing the hypothesis formulated with reference to deductive reasoning.

*(iii) Computation of National Income Accounts:* Computation of National Income Accounts needs the systematic collection, classification and presentation of facts with respect to certain macro variables like production, income, expenditure, savings, investment, etc. Data collected in these areas are classified into certain functional areas or geographical regions or sectors of the economy. This type of analytical presentation of data helps us understand the state of economy and also the degree and direction of change. National Income Accounts also give information on the value-added by different sectors of the economy.

*(iv) Economic Planning:* Statistics provides necessary data in respect of trends in population growth, employment level, extent of industrialisation and rate of capital formation, level of production capacity, scope for exploiting the natural resources, etc. On the basis of this data priorities of growth are determined and time-bound targets are laid down while chalking out the economic planning. This is followed by the evaluation of different plans. In the absence of statistics, efficient planning can not be thought of.

A statistical approach to an economic problem not only helps in identifying the nature, scope and magnitude of deficiencies but also provides necessary guidance to tackle the problems. As A.L. Bowley opines,"No student of political economy can pretend to know complete equipment unless he is master of the methods of statistics, knows its difficulties, can see where accurate figures are possible, can criticise the statistical evidence and has an almost instinctive perception of the reliance that he may place on the estimates given to him". According to C.E. Engeberg, "No economist would attempt to arrive at a conclusion concerning the production or distribution of wealth without an exhaustive study of statistical data. In India, non-availability of or inadequate and inaccurate statistics are considered to be responsible for the drawbacks in the planning process.

## 2.5.4 TO GOVERNMENT

In fact the origin and growth of statistics as a science of knowledge is considered to be the byproduct of administrative requirements of governments. Today, modern governments are concerned with the promotion of human welfare. Promotion of human welfare requires proper understanding of the hurdles that obstruct the development process. Statistics in respect of economic and social programmes help the governemnt in envisaging various programmes. Formulation, implementation and evaluation of plans requires collection of statistics in the relevant fields. As such, every ministry and department of government depends heavily on factual data for its efficient functioning.

## 2.6. LIMITATIONS OF STATISTICS

At present, the usefulness of statistics is not limited to any particular field of knowledge. It is widely used as a means to solve many complex problems. However, one should not feel that statistics is free from limitations. The usefulness of statistics depends upon the systematic statistical investigation. Unless data are properly collected, organised, analysed, presented and interpreted, there is every likelihood of drawing wrong inferences. Some of the important limitations of statistics are explained below:

*(i) Statistics deals only with quantitative aspects of the problems*

Statistics deals with only such phenomena whose characteristics can be measured or counted and expressed in quantitative terms, e.g., wages, prices, production, sales, etc. Qualitative aspects which can not be subject to direct quantification are incapable of statistical analysis, e.g., honesty, goodwill, efficiency, intelligence,etc., can not be expressed in numerical terms. For the indirect quantification of such things, statistics provides necessary tools, though they are not precise and definite. For instance, intelligence of students is judged on the basis of marks obtained by them in an examination.

*(ii) Statistics deals with aggregates of facts*

Statistics does not deal with individual measurements. It is concerned with the values of mass data. For example, the wages earned by an individual worker at any particular time itself does not constitute the subject matter of statistics. But the wages of workers of a factory can very well be the subjet matter of statistics. In view of this, statistics proves to be ineffective where individual problems are to be studied. Thus, the usefulness of statistical analysis is limited to only those problems where the study of group characteristics are important.

*(iii) Statistical results are true only on an average*

Statistics deals with averages only. Statistical results reveal only average behaviour of the data under study. These averages are made up of various individual items which may be different from each other. Hence, statistical results can not be indiscriminately applied to individual cases. For example, the per capita income figure of Indians can not be made applicable to any single individual unless we ascertain the extent of dispersion of income.Dispersion of income shows the degree of variability in the incomes of individuals.

*(iv) Statistics is only a means and not an end*

Statistics is only one of the methods of studying a problem. Many a time it is necessary to study a problem in the light of a country's culture, religion and philosophy. Statistics willnot be of much use in studying such problems. In order to arrive at useful conclusions, the results obtained through statistical methods will have to be supplemented and substantiated with other evidences. Statistics analyses the facts and throws light on the real situation but statistical results will have to be interpreted by experts only. People, without having proper understanding and awareness of the limitations of statistical tools and techniques can not apply these results to

problem situations. Statistics provides necessary means and it is upto the investigator to use them properly. Thus, statistics is only a means in solving the problems but not an end in itself.

*(v) Statistics can not be indiscriminately applied to all situations*

Statistics are usually collected for a predetermined purpose. As such the data collected must be used for the specific purposes for which they have been collected. If statistics are applied indiscriminately to all situations, they are sure to give misleading conclusions. Data collected once can be preserved and used in subsequent studies but in such cases proper care should be taken to scrutinise the validity of such data.

*(vi) Statistics can be misused*

One of the important limitations of statistics is that they are often misused by many to prove their point of view. Statistical conclusions based on imcomplete information may lead to misleading conclusions. W.I.King pointed out, "one of the shortcomings of statistics is that they do not bear on their face the label of their quality". Moreover any inexpert can also deal with statistics.

Statistics can be interpreted according to the will and pleasure of the statistician. A statistician can try to establish wrong things as correct and correct things as wrong. Thus, statistics if they are misused, cause disaster. But if they are used meticulously, they work wonders.

**Check your progress - 2**

List out the limitations of statistics

## 2.7 DISTRUST OF STATISTICS

The usefulness of statistics depends upon the reliability of data collected and the correct interpretation of data. Figures are so convinvcing that any thing can be made easily belieavable. Mark Twain remarked that there are three types of lies, namely, lies, damned lies, and statistics. Thus, he considered statistics the superlative degree of lying. It is often said that statistics can prove any thing and that an ounce of truth can produce tonnes of statistics.

Statistics are considered to be the most dangerous tools in the hands of inexperts. Further, one can easily manipulate the figures to prove false things as correct. Thus due to ignorance and bias, people misuse the delicate tools of knowledge. This has created doubts and disrespect about the science of statistics among the common men. Statistics are like clay out of which one can make a God or a Devil as he wishes. Thus, the fault lies not with the science of statistics but with the users who misuse it. Statistical data and conclusions are often manipulated in the following ways:

i) Shifting definitions

ii) Inadequate data and small samples

iii) Bias

iv) Mis-interpretation of data

v) Misuse of statistical methods.

*(i) Shifting definitions:* Slight alteration in the definition of a term sometimes gives altogether a different conclusion. For example, while comparing the number of workers in two different factories, misleading and unwarranted inferences may be drawn if consistency with regard to the definition of the term 'worker' is not ensured. If in one factory 'workers' include casual labourer and in another factory casual labourer is excluded, comparability is lost. If dissimilar things are compared, the results obtained will be misleading.

*(ii) Inadequate data and small samples:* Conclusions drawn on the basis of inadequate data will be misleading. The small sample selected for the study may not be a true representative of significantly large population. For example , if we take out a sample of 50 students out of 5,000 students of a college and conclude that the students of a college are very intelligent, it would be quite mis-leading because the sample constitutes a very insignificant proportion of the Universe. Conclusions drawn on the basis of incomplete data will certainly mislead the people. In order to draw valid conclusions, data relating to all significant characteristics of the phenomena under study must be collected. Observe the following statement.

"The profits of Firm 'A' are Rs. 75,000 for 1982-83 and that of Firm 'B' Rs. 1,00,000 for the same period". On the basis of this information only one would infer that Firm 'B' is better than Firm 'A'. However, if we examine the amount of capital invested in both the firms we might reach at a different conclusion. Assuming that the Firms invested Rs. 10,00,000 and Rs. 20,00,000 respectively the rate of return on investments would be 7.5% and 5% respectively. Now one can conclude that Firm 'A' is decidedly more profitable than Firm 'B'.

*(iii) Bias:* While conducting a statistical investigation, the statistician should be impartial and unbiased. The investigator may be biased while selecting items to be included in the sample. While drawing conclusions also he must be free from bias and prejudice, otherwise conclusions drawn may be wrong and misleading. For example, while selecting sample units for conducting a survey on students' spending habits, if the investigator deliberately selects only those students who belong to well-to-do families, the conclusions drawn would be definitely misleading.

*(iv) Misinterpretatin of data:* While interpreting the results of statistical analysis, one must be aware of the limitations, characteristics and assumptions of statistical methods. Interpretations made by inexperienced persons may give altogether misleading generalisatins. If proper care is not taken there is every possibility of committing serious mistakes. For example, the profits of a firm in 1983 have come down from Rs. 2.0 lakhs in 1982 to Rs. 1.8 lakhs. The firm had stopped the production of commodity 'M' which was fetching Rs. 44,000 per annum because of the non-availability of raw-materials. This means that on the whole the firm has improved its performance but, by looking at the figures one would interpret that the firm's performance has deteriorated when compared to that of last year.

*(v) Misuse of statistical methods:* Statistics deals with averages and aggregates of facts. It does not deal with individual phenonmenon. Averages are always misleading. Unless the degree of dispersion is considered, generalisations made on the basis of averages willbe misleading. Similarly, if conclusions drawn on the basis of aggregates of facts are made applicable to individual cases, it amounts to the misuse of statistical methods. Similarly, use of graphs with two different bases, association or correlation between two or more unrelated attributes or variables will give a wrong picture about the facts. For example, if we try to cross a river on the basis of average depth which is 3'-6" we may face dire consequences. Because through out the distance of the river the depth will not be 3'-6". The average depth may be subject to higher degree of dispersion.

On account of the above deficiencies and also because of the dishonesty and ignorance of some statisticians, a sort of doubt is developed towards the science of statistics. In fact, statistics, as such, is not bad, but the people who manipulate the statistical results for their selfish ends are bad. Statistical methods are delicate tools like a blade. If the blade is not used with proper care, it can harm people. Hence, the fault lies not with the science of statistics but with those who misuse it.

## 2.8 SUMMING UP

Statistics helps in planning and designing the various procedures involved in the setting up and testing of hypothesis. Statistics not only provides necessary principles and techniques, but also guides us to achieve the desired objectives with them. Statistics is regarded as an art of applying the science of scientific methods.

Statistics replaces ignorance, prejudice, rule of thumb and arbitray decisions with quantitative

and scientific decision making. The importance of statistics is steadily increasing due to its wide range of functions, viz., simplification and presentation of mass data in a definite form. It helps comparative study and facilitates in making predictions. Knowledge of statistics helps the business managers to take appropriate decisions pertaining to the determination of location and size of plants, marketing, production planning, quality control,personnel administration, accounts and audinting. It also helps the government in planning and executing various economic and social programmes. However, statistics suffers from certain limitations as it deals with aggregates of facts. Its laws are not exact and its results are true only on an average. On account of these limitations and also due to ignorance and bias, people often misuse the delicate tools of statistics which has created distrust and disrespect towards statistics.

## 2.9 CHECK YOUR PROGRESS : MODEL ANSWERS

1. The statistical tools used to simplify the mass data are:
   Averages, ratios, measures of variation, coefficients etc.
2. Mention the limitations such as:
   - Dealing with quantitative aspects
   - Dealing with aggreagate of facts
   - Results being true on an average
   - Means to arrive at a conclusion rather than an end by itself
   - Can not be used for all situations
   - Scope for misuse.

## 2.10 MODEL EXAMINATION QUESTIONS

**A. Short Questions**

1. Comment on the following
   i Statistics deals only with quantitative aspects of the problem.
   ii Statistical results are true only on an average.
   iii Statistics is only a means and not an end.
   iv Statistics cannot be indiscriminately applied to all situations.
   v Statistics is the superlative degree of lying.
   vi The science of statistics throws light on social problems.
2. Give two examples of misuse of statistics.
3. List out the limitations of statistics.
4. How does statistics help in predictions?
5. How are statistical methods helpful in compiling National Income Accounts?
6. How does statistics enlarge individual knowledge and experience?

**B. Essay Questions**

7. "Statistics are like clay out of which you can make a God or a Devil, as you please." Comment.
8. Discuss the limitations of Statistics. What are the causes of distrust of statistics?
9. Explain the importance of statistics with special reference to business and industry.
10. Bring out the importance of statistics in arriving at the policy decisions in an economy.

11. "Statistics are straws out of which I, like every other economist, have to make bricks" (Marshall). Elucidate this statement and indicate the utility of statistics in economic planning of India.

## 2.11 RECOMMENDED BOOKS


1. Gupta, S.P. : "Statistical Methods", Sultan Chand & Company, New Delhi.
2. Gupta, B.N. : "Statistics", Sahitya Bhavan, Agra.
3. Gupta, S.C. : "Fundamentals of Statistics", Himalaya Pub. House, Bombay.
4. Simpson and Kafka : "Basic Statistics", Oxford and I.B.H. Publishing Company, Calcutta.


## 2.12 GLOSSARY

Cause and effect relationship : Sometimes, the occurrence of one event results in the happening of another event. The relationship between independent event and dependent event is known as cause and effect relationship.

Distrust of Statisttics : Disbelief of statistics

## UNIT - 3 : NATURE OF DATA

**Contents**





## 3.0 AIMS AND OBJECTIVES

This unit aims at giving the basic idea about the nature and types of data, statistical investigation and planning statistical investigation.

After reading this unit, you would be able to:

- explain the meaning and nature of data
- classify the types of data
- describe the meaning and types of statistical investigation
- identify the various stages of statistical investigation.

## 3.1 INTRODUCTION

The term 'data' refers to a set of numerical figures or facts. These figures and facts are collected through a statistical investigation. Since utility of the conclusions derived from the findings of the statistical investigation depends on the reliability of data, the statistical investigation must be planned carefully and executed properly. Further, the investigator has to select appropriate methods for collecting accurate data. In this unit, planning and execution of statistical investigation is explained in detail.

## 3.2 MEANING AND NATURE OF DATA

Data refers to a set of numerical facts and figures. These numerical facts may relate to units produced or sold, profits or losses, size of population, national income, etc. Statistical data may be treated as observations made on certain aspects for the purpose of attaining certain predetermined objectives. According to Levin data is 'collection of any number of related observations'. Data occupies the most important place in statistics as all statistical decisions are made on the basis of numerical facts and figures.

## 3.3 TYPES OF DATA

Data may be quantitative or qualitative. While data relating to quantitative aspects can be directly measured or counted in numerical terms, data in respect of qualitative aspects cannot be measured in direct and definite terms. It can only be quantified indirectly. Qualitative data come from the process of identification of the presence or absence of characteristics under study. Production, Sales, incomes, etc., are some of the examples for quantitative data which are called as variables. Rich, poor intelligence, honesty, etc., are some of the examples for qualitative data which are known as attributes. Quantitative data can be univariate, bivariate or multivariate. While univariate data describes a single variable, bivariate and multivariate data describe two and more than two variables respectively.



## 3.4 MEANING OF STATISTICAL INVESTIGATION

Usually data are collected with the help of statistical investigation. A statistical investigation is a systematic method of collecting information with respect to a variable or an attribute. This aims to achieve a particular objective or purpose. Statistical investigation is a search for knowledge conducted with the help of the application of appropriate statistical principles and techniques. It is a process wherein an individual or an agency collects relevant information in numbers rather than in words. In this process of statistical investigation, data in respect of an attribute is subject to indirect measurement. The person who conducts statistical investigation is called statistical investigator and the person or unit from where the information is collected is known as "respondent".

## 3.5 TYPES OF STATISTICAL INVESTIGATION

Before the investigator starts the work of data-collection, he must decide about the type of investigation, as the planning and organisation of data depend on the type of investigation. Further, the object and scope of investigation, the organisational arrangements, preparation of questionnaires and schedules, methods and techniques of data collection and processing of data depend to a large extent on the type of investigation. Statistical investigation may be of the following types:

3.5.1 Direct or Indirect

3.5.2 Initial or Repetitive

3.5.3 Regular or Adhoc

3.5.4 Confidential or open

3.5.5 Official or Non-official or Semi-official

3.5.6 Extensive or Limited

3.5.7 Pilot or Comprehensive

### 3.5.1 DIRECT OR INDIRECT

If the data are collected directly from the respondents concerned, it is called direct investigation. On the other hand, if the data are collected from third parties, other than the respondents, is called indirect investigation.

### 3.5.2 INITIAL OR REPETITIVE

An initial investigation is an original investigation conducted for the first time. This type of investigation requires careful planning and elaborate organisational arrangements. A repetitive investigation is conducted in continuation of an initial investigation. This investigation does not require separate planning since already a prepared plan exists. It may need only some changes to suit the repetition.

### 3.5.3 REGULAR OR ADHOC

A regular investigation is carried out regularly on a continuous basis. It is conducted at regular intervals of time period. An adhoc investigation is conducted at a particular time period aiming at a particular object and a specific purpose.

### 3.5.4 CONFIDENTIAL OR OPEN

Confidential investigation otherwise knwon as secret investigation is conducted confidentially or in camera. Confidential investigations are purely meant for private purpose. On the other hand, in nonconfidential or open investigations, the investigation is conducted publicly.

### 3.5.5 OFFICIAL OR NON-OFFICIAL OR SEMI-OFFICIAL

The investigations carried out by, and on behalf of the Government are known as official investigations. In these investigations, Government can use compulsion to provide information as certain laws include the compulsory provision of information. Since the Government can afford to spend huge amounts, the scope also will be wider in nature. On the other hand non-official investigations conducted by other than government agencies, will have limited scope due to problems related to extracting data and availability of resources. A semi-official investigation is carried out by certain organisations enjoying government patronage eg., Universities, Research Institutions, etc.

### 3.5.6 EXTENSIVE OR LIMITED

While an extensive investigation includes wider coverage of aspects, limited investigation usually includes very few aspects to be studied.

### 3.5.7 PILOT OR COMPREHENSIVE

A Pilot investigation precedes a comprehensive investigation to understand the problems likely to be encountered in the process of investigation, whereas a comprehensive investigation is an indepth study for attaining the object and purpose of investigation.

However, the types of investigations are not water tight compartments. Some of them are interchangeable depending upon the nature and scope of enquiry, availability of resources, etc.

## 3.6 STAGES OF STATISTICAL INVESTIGATION - PLANNING

A well organised statistical investigation contains the following stages:

a) Planning the Statistical Investigation

b) Collection of data

c) Organisation of Data

d) Presentation of Data

e) Analysis of Data

f) Interpretation of Data

While the planning of the statistical investigation is covered in this unit, other stages of statistical investigation are discussed in subsequent units.

***a) Planning the Statistical Investigation***

Planning is the first and most important stage of a statistical investigation. If the investigation is not well planned in the beginning itself, the investigator may not acquaint himself with the practical difficulties which he has to encounter in the actual execution of the plan. Thus, he may not be in a position to collect the appropriate data to fulfil the objectives of the study. Carefully planned statistical investigation helps the investigator to collect relevant data, which

can be systematically organised, presented, analysed and interpreted on scientific lines. Thus, meaningful conclusions can be drawn and the problematic situation can be understood in its right perspective. Planning of investigation needs the following:

i) Defining object and scope of investigation

ii) Specification of various terms

iii) Deciding the source of information

iv) Deciding the degree of accuracy desired in final results.

v) Deciding the methods of collecting the data.

*i) Defining object and scope of investigation*

At the outset, the investigator has to state the object of the investigation in clear and definite terms. It is also necessary to define the scope of investigation. Statement of the object involves specifying the purpose for which the investigation is to be carried out. The scope of an enquiry outlines the nature and coverage of investigation. This includes the determination of the type of information needed , number of respondents to be approached, geographical area to be covered, etc. For example, if the investigator intends to conduct an investigation in respect of a consumer survey, his objective may be to regulate the production in accordance with the consumers' tastes and preferences. To achieve this objective, his scope of investigation may determine the market area to be covered, present and prospective consumers to be approached, number of consumers to be included in the study, etc. To the extent possible, consumers opinion about the product and also data relating to the availability of substitutes and complimentaries should be collected. A clear statement of object and scope avoids collection of irrelevant data. This eliminates confusion and saves the precious time and money of the investigator. In the words of Wessel, Willet and Simone, "The purpose of the project should always be spelled out as precisely as possible. This will ensure the collection of the proper information and spare the expenses and trouble of handling irrelevant data". The specification of the objects helps classification, tabulation, analysis and interpretation of collected data. Further, the importance of defining the object and scope of a statistical investigation can be understood from the words of Prof. Ya-lun-chou who states that, "only after specifying the objects, it can be determined which data shall be relevant to the problem under enquiry. If he fails to do this, the data collected may be entirely irrelevant or may even tend to cloud rather than clarify the problem. It is well to remember the quality of statistical conclusion depends upon the appropriateness and accuracy of data, which in turn depends upon exactness in formulating the problem. Statistical techniques, no matter how refined and precise, cannot yield useful results for arriving at decisions on problems, if they are applied to inappropriate data".

The scope of investigation must be determined by keeping in mind the object of investigation, availability of resources such as money, time and experienced and trained personnel to carry out the investigation.

*ii) Specification of various terms*

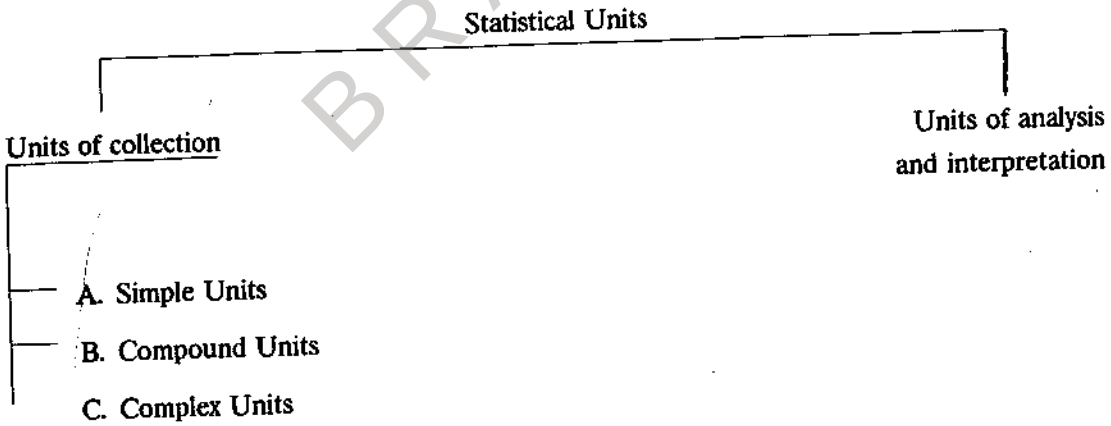
After determining the object and scope of the investigation, the investigator will have to state the various terms relating to the study. To avoid ambiguity and confusion, it is always desirable to define clearly and precisely all terms that are to be used in various phases of the statistical investigation. Such definitions should be followed uniformly throughout the investigation. For example, an investigation regarding wages of workers should clearly spell out the term 'worker'. At one phase of the survey if worker includes casual labour and in the subsequent stages if it excludes casual labour uniformity is not said to be ensured. This type of inconsistency will not serve the purpose of the investigation.

The investigator should also define the statistical unit in terms of which data is to be counted or measured. The unit in terms of which the investigator collects, organises, analyses and interprets the statistical results is known as a statistical unit. A clear definition of statistical unit avoids misunderstanding and ensures homogeneity in the data. This facilitates meaningful comparisons and observations. The need for defining the unit is very essential in case of qualitative data which lacks standard connotations.

Statistical units may be either conventional or arbitrary. While conventional units are universally accepted and do not require any specific definition, the definition or arbitrary units changes from investigation to investigation. Hence they require clear cut definition. Metre, litre, hour, year, ton, etc., are some of the examples of conventional units. On the other hand, workers, wages, capital employed, profits, etc., are some of the examples of arbitrary units.

Statistical units are classified into two categories viz., (a) units of collection, (b) units of analysis and interpretation. The classification of statistical units is depicted in the following chart:






Units of collection are specifications in terms of which data is collected, measured and recorded. Units of collection must be defined clearly before the commencement of the investigation to suit the object and scope of investigation.

Units of collection may be simple, compound and complex. A simple unit is a set of single condition or character. It is also known as conventional unit. A compound unit is a simple unit

with some qualifying conditions which limit the use of the unit. On the other hand, a complex unit is formed by adding two or more qualifications to a simple unit. A Rupee, a ton, a house, a month, a worker, etc., are examples of simple units. While a metric tonne, a skilled worker, a passenger kilometre, a machine hour etc., are examples of compound unit, wage per man hour, speed per kilometre hour, etc., are examples of complex units.

Units of analysis and interpretation are statistical units with the help of which statistical data is analysed, presented and interpreted. These units of analysis and interpretation differ from investigation to investigation depending upon the degree of accuracy desired. Examples of units of analysis and interpretation are rates, ratios, and percentages, co-efficients, etc. While rates are used to express the relationship among heterogeneous factors, ratios and percentages are used to describe the relationship among homogeneous factors. Usually rates are expressed per thousand, whereas, rate per unit is known as a co-efficient and a percentage is expressed per hundred. A ratio may be expressed in terms of a percentage or a co-efficient.

A good statistical unit must possess the following characteristics:

*(a) It must be simple and clear:* The statistical unit must be defined in simple and clear terms. It must be understandable to all connected with the investigation. It must be rigidly defined to avoid confusion and ambiguity.

*(b) It must be suitable to the object and scope:* The statistical unit must suit the object and scope of investigation. As Horace Secrist said, "the purpose of study and the unit of study are reciprocal to each other. Statistical units cannot be defined without regard to the purpose, and their purpose cannot be outlined with sufficient accuracy, without clear notion of units".

*(c) It must be specific:* The statistical unit used in the investigation must be specific. If the unit is not specific, it can be misunderstood and thus the data collected in terms of such units may suffer on account of inaccuracies. In case of arbitrary units, care must be taken to assign specific meaning.

*(d) It must be ascertainable:* Any unit of study must exist in the real world or population. It cannot be hypothetical or unreal. In the words of W.I.King, "Not only must the unit be defined with precision, but it must also be of such a nature that it may be correctly ascertainable".

*(e) It must be stable and standard:* The statistical unit must be stable and standard at different time periods and different places. If the value of statistical unit is subject to frequent changes the data collected in terms of such units would not be comparable and the utility of statistical investigation may be lost.

*(f) It must be homogeneous:* To ensure comparability, the statistical unit must be homogeneous and uniform. Heterogeneity of units will make the comparisons difficult, confusing and misleading.

*(g) It must be self explanatory:* The statistical unit must be self- explanatory and should not call for lengthy explanations. The self-explanatory nature of statistical unit helps the investigator to carry out the investigation without any confusion.

*iii) Deciding the source of information*

After determining the object, scope and statistical units, the next step in planning the statistical investigation is to decide the sources of data. The sources of data may be primary and secondary. Primary data consists of facts and figures collected for the first time to fulfil the objective of a particular statistical investigation. Secondary data consists of facts and figures originally collected for a particular statistical investigation but are used for a different statistical investigation. Primary sources include data collected in population census, data obtained through direct interviews with the respondents, etc. On the other hand, secondary sources include published or unpublished records. The distinction between the primary and the secondary data is only one of degree because what is primary data in the hands of a person may be secondary data to others. The decision about the choice of source of data depends upon the object, nature, scope and the degree of accuracy desired.

*iv) Degree of accuracy desired in final results*

Before the commencement of the collection of data, it is necessary to specify the degree of accuracy desired in clear terms. Since perfect accuracy is not possible, the investigator must aim at reasonable degree of accuracy. This is because of the following reasons.

a) Statistics is basically a science of estimates

b) The prevalence of unintentional bias

c) The problem of approximations and simplifications

d) Errors in collection, organisation, analysis, presentation and interpretation of data.

e) Limitations of resources.

According to W.I.King, "For every statistical problem, there should be determined in advance a definite standard of accuracy for each item and every endeavour should be made to bring each recorded instance upto this standard,but this standard by no means needs to correspond to the highest degree of accuracy attainable".

The degree of accuracy desired for a particular enquiry will depend mainly on the object and scope of investigation, the time and cost. To achieve the highest degree of accuracy it is undesirable to incur high costs. But at the same time, accuracy should not be sacrificed to minimise the costs. In the words of Riggleman and Frisbee, "The necessary degree of accuracy in counting or measuring depends upon the practical value of the accuracy in relation to its cost". Statistical errors cannot be totally eliminated but can be minimised to attain a reasonable degree of accuracy.

*v) Methods of data collection*

The decision regarding the methods of data collection is considered to be the last stage in planning the investigation. At this stage, the investigator has to decide about the methods through which necessary data is to be collected. Broadly, there are two methods of data collection, viz., Census method and sample method.

A census method, otherwise known as complete enumeration, implies collection of data from each and every unit of the universe or population. According to Humburg, "The universe or population consists of the total collection of items or elements that fall within the scope of a statistical investigation". On the other hand, sample method refers to the study of few items selected from a universe. These units are considered to be the representatives of the universe. Thus the results obtained through the sample method are made applicable to the universe from which they are drawn.

After planning the statistical investigation, the investigator has to execute the statistical investigation. A well-planned statistical investigation helps the investigator to acquaint himself with the plan of operation and execute it. The execution of statistical investigation involves the following steps:

a) Making necessary organisational arrangements

b) Designing various forms

c) Appointment of necessary field staff

d) Processing and analysis of data

e) Preparation of final report

*(a) Making necessary organisational arrangements:* The statistical investigations which have smaller coverage do not require elaborate organisational arrangements. But those investigations which are wider in scope require elaborate organisational arrangements. Often the investigator has to take the help of field investigators for conducting the investigation. In certain countries, specialised agencies are engaged to conduct the investigation. The organisational arrangements depend upon the nature and scope of investigation.

*(b) Designing various forms:*The investigator has to use a variety of forms such as questionnaires, schedules, data sheets, etc., in the process of investigation, organisation and presentation of data. Such forms must be designed properly and systematically before hand to help the work of the investigator. This avoids overlapping and confusion in the execution of the statistical investigation.

*(c) Appointment of necessary field staff:*Investigation with wider scope requires a number of field staff for the collection of data. If efficient and well trained investigators are not readily available in the organisation , they are to be selected, trained and supervised during the process of data collection. Since collection of data requires intelligence, honesty, hard work and scientific knowledge, field investigators have to be selected carefully. After the selection of investigators, they have to be properly trained to equip themselves with the selection of sample, advanced techniques and methods of data collection, various terms and references used in the questionnaire, and statistical techniques. Once the investigators are equipped with requisite training, they can be sent for enumeration. If necessary, field checks and field audit may be introduced to avoid collection of irrelevant data. All precautions should be taken to convince the non- respondents to give necessary information.

*(d) Processing and Analysis of data:* Mass data collected through enumeration may contain heterogeneous elements which are in raw form. Such data must be processed and analysed to make it suitable for further statistical treatment. The data so processed and analysed will become intelligible and easy to understand. For processing the data, both manual and mechanical techniques can be used. Modern data processing machines such as calculators, computers, punch cards, etc., may be used to speed up the work with reasonable accuracy. For analysis of data, simple techniques such as averages, dispersion, skewness, etc., and advanced and sophisticated techniques like correlation, regression, analysis of time series, probability, etc., can be used.

*(e) Preparation of final report:* After the data is processed and analysed, the results must be presented in a suitable form or a report. A report may be general or technical. While a general report contains description of data and the results of investigation, a technical report provides additional information such as sample design, statistical units, methods of calculations, etc. A good report must cover the following aspects:

i) Statement of the purpose of survey

ii) Description of the scope

iii) Specifications of frames and design of investigation

iv) Nature and methods of collection of information

v) Statistical analysis and computational procedures

vi) Computation of results

vii) Degree of accuracy

viii) Personnel and equipment used

ix) Comparisons

x) References, footnotes, index, etc.

**Check your progress - 1**

List the characteristics of a good statistical unit.

## 3.7 SUMMING UP

Statistical data constitutes the basis for statistical analysis. The term 'data' is generally applied to numerical statement of facts. Data may be qualitative and quantitative. Qualitative data is indirectly measured by classification of characteristics. Quantitative data measured in definite numbers can be uni-variate, bi-variate or multi-variate. While the uni-variate data describes single variable, the bi-variate data describes two variables. Multi-variate data describes more than two variables. Organising the collection of data requires a detailed planning. A Statistical plan requires appropriate organisational arrangement, careful selection and training of field enumerators.

## 3.8 CHECK YOUR PROGRESS: MODEL ANSWERS

1. Mention the following points.

i) It must be simple and clear.

ii) It must be suitable to object and scope

iii) It must be specific

iv) It must be ascertainable

v) It must be stable and standard

vi) It must be homogeneous

vii) It must be self explanatory.



## 3.9 MODEL EXAMINATION QUESTIONS

A. Short Questions

1. Define statistical data.
2. What is bivariate data? Give an example.
3. What is univariate data?
4. What is confidential investigation?
5. Define statistical unit.
6. How do you classify the statistical unit?
7. What is meant by "conventional statistical units"? Give an example.
8. What do you know simple statistical unit.
9. What is meant by homogenity of statistical unit?
10. What is compound statistical unit?
11. What is an arbitrary statistical unit?
12. Distinguish between extensive and limited investigations.

13. List out the stages of statistical investigation.

14. Distinguish between direct and indirect investigation

15. Distinguish between official and non-official investigations.

16. What do you understand by units of analysis and interpretation?

17. What are the contents of a good report?

18. How do you proceed to classify data?

**B. Essay questions**

19. "Defining the object and scope of investigation occupies most important place". Explain the statement with a suitable example.

20. Briefly explain different stages of statistical investigation.

21. Define the term "statistical investigation". Enumerate the various types of statistical investigations.

22. "Execution of the statistical plan is utmost important". Explain with reference to various steps involved in the execution of statistical plan of investigation.

23. What are the essential characteristics of a good statistical unit?

## 3.10 RECOMMENDED BOOKS


1. Gupta, S.P. : "Statistical Methods", Sultan Chand & Company, New Delhi.
2. Gupta, B.N. : "Statistics", Sahitya Bhavan, Agra.
3. Gupta, S.C. : "Fundamentals of Statistics", Himalaya Pub. House, Bombay.
4. Simpson and Kafka : "Basic Statistics", Oxford and I.B.H. Publishing Company, Calcutta.

## 3.11 GLOSSARY

| | | |
|---|---|---|
| 1. Attribute | : | The qualitative characteristic of an object, is called attribute. For example, illiteracy, unemployment, etc. |
| 2. Primary data | : | Data which is original in character. |
| 3. Secondary data | : | If any one else makes use of data which had already been collected by some agency, it is termed as secondary data. |
| 4. Statistical investigation | : | Statistical investigation is a process of collecting data from existing population units with no particular control over factors that may affect the population characteristic of interest in the study. |
| 5. Statistical unit | : | It is the unit with which the statistical investigator collects, organises, analyses and interprets the statistical data. |
| 6. Variable or variate | : | The quantitative phenomenon like rainfall in different years, price of a commodity at different points of time, heights of students in a class etc., is termed as variable or variate. |

# UNIT-4 COLLECTION OF DATA

**Contents**





## 4.0 AIMS AND OBJECTIVES

The aim of this unit is to describe the various methods of collecting primary data, sources of secondary data and precautions to be taken in using the secondary data.

After studying this unit, you would be able to :

i) explain the meaning of various types of data

ii) explain the various methods of collecting primary data

iii) identify the sources of secondary data ; and

iv) describe the editing of data

## 4.1 INTRODUCTION

In unit 3 we have discussed the first stage of statistical investigation i.e., planning the statistical investigation . Now we attempt the second stage i.e., collection of data. To draw any conclusion of a problem, collecting the quantitive facts or figures are the basis for final results. Before collecting the data, the statistical investigator must identify the sources of such data. If the data required was not collected previously by any agency or individual, the methods to be followed for collecting such data would be; direct personal interviews; indirect oral interviews; information from correspondents, questionnaires and schedules. If such data, were collected previously by some agency or individual, whether published or not, the approach to be followed would be different. Let us get into the details.

## 4.2 MEANING AND SIGNIFICANCE OF COLLECTION OF DATA

Collection of data implies a systematic and meaningful assembling of information for the accomplishment of the objectives of a statistical investigation. It refers to the purpose of gathering of information relevant to the subject matter of the study from the units under investigation. According to Crum, patton and Tebbutt, collection of data means, "The assembling, for the purpose of a particular investigation, of entirely new data, presumably not already available in published sources". This definition confines the term " collection" to the primary sources of data. However, in a broader sense, the term is generally applied to both primary and secondary sources of data. Collection of data also refers to the methods used in getting the required information from the units under investigation.

Collection of data is the most important stage in the process of statistical investigation. The final results of the investigation are ultimately based on the data collected from various sources. Any lapse, negligence or bias on the part of the investigator makes the final result faulty and worthless. Therefore, collection of data must be done carefully and properly.

## 4.3 SOURCES OF DATA

The data for the purpose of statistical investigation may be collected from primary and secondary sources. Here, a 'source' may be referred to a person or an organisation having the data, or the reports and publications in which the data is published. A primary source may be understood as any organisation , person or publication, which gathers the data hitherto not collected elsewhere through first hand investigation. In the words of Neiswanger, "Primary source is a publication in which the data are published by the authority which has gathered and analysed them". A secondary source means, a source which gives the information that has already been gathered, analysed and presented by other agencies for their own purpose.

Data generated from a primary source is called primary data and data collected from a secondary source may be referred to as secondary data. In other words, data orginally collected in the process of investigation is known as primary data and that collected by other persons is called secondary data.

## 4.4 DIFFERENCE BETWEEN PRIMARY AND SECONDARY DATA

It is difficult to clearly draw a line of demarcation between primary and secondary data, because the difference between them is only relative. Prof. Horace Secrist rightly points out that, "The distinction between primary and secondary data is largely of degree. Data which are secondary in the hands of one party may be primary in the hands of another". To illustrate this by an example, if the Open University collects the details of students appearing for the B.Com. Degree Examination through direct enumeration, it is regarded as the primary data for the University. If the Government or other agency takes such information from the records of the Open University, it becomes secondary data. Though the difference between the primary and secondary data is releative in nature, the following points may help to spot the differences.

| | *Primary data* | *Secondary data* |
|---|---|---|
| i) | The data is collected from its orginal source. | The data which is already collected from the original source is made use of for the purpose of statistical investigations. |
| ii) | The data is original in its form. | The data is refined in its form. |
| iii) | The definition of the statistical unit can be made use of to suit the objective of investigation. | The definition of units already used in the collection of data has to be adopted for the purpose of statistical investigation. |
| iv) | A new plan of investigation and execution has to be carefully prepared for the purpose of collection. | No fresh plan of investigation is required since the data is already collected and processed. |

**check your progress - 1**

Distinguish between primary and secondary data.

## 4.5 ADVANTAGES AND LIMITATIONS OF PRIMARY AND SECONDARY DATA

Before taking a decision to use primary or secondary data, it is important to consider their advantages and limitations. They are discussed below:

**Primary Data**

**Advantages**

*(i)* The primary data gives more reliable, accurate and adequate information which is suitable to the object and purpose of investigation.

*(ii)* The primary data has the advantages of avoiding the errors arising from the copying of figures from the books as in the case of secondary data.

*(iii)* While collecting the primary data, the definitions for various terms are given to suit the objective of investigation. Hence the data collected may be utilised completely and meaningfully for the purpose of investigation.

**Limitations**

*(i)* Collection of data from a primary source needs detailed planning and execution which is a time consuming process.

*(ii)* Collection of primary data requires field investigators who are efficient, trained, intelligent, sincere and tactful. Selection of such investigators involves many difficulties including heavy financial burden on the collecting agency.

*(iii)* The reliability of primary data depends upon the honesty, sincerity and integrity of the investigator on one hand and the degree of co-operation given by the respondents on the other. Often, the primary data gives misleading information due to lack of integrity of investigators and non-cooperation of respondents in giving answers to certain delicate questions.

**Secondary Data**

**Advantages**

*(i)* It is convenient to gather secondary data since it is already collected and analysed.

*(ii)* Collection of data on certain aspects like cropping pattern, characteristics of population, etc., through primary sources is not feasible for individual or private organisations due to time and money constraints. For such information, only secondary data has to be used.

*(iii)* Collection of secondary data is economical in terns of time., money and effort.

**Limitaions**

*(i)* Since the data has been already collected for a specific purpose, it may not suit the requirements of the present statistical investigation.

*(ii)* The degree of accuracy aimed at the time of collection of data may be different from

the degree of accuracy desired for the present investigation.

*(iii)* Secondary data may not be reliable, since the agency collecting the data may be possibly let in some biased or unbiased erorrs.

*(iv)* Various terms and references used in the secondary data may not always be useful in the present investigation.

## 4.6 CHOICE BETWEEN PRIMARY AND SECONDARY DATA

After considering the relative merits and limitations of primary and secondary data, the investigator faces the task of choosing the appropriate data. Primary data is generally used where the secondary data available does not suit the purpose and analysis. Now-a-days, there is increasing trend towards using secondary data, because it is available in different areas of research. However, the uses of primary and secondary data are not mutually exclusive. In most of the investigations, these two sources of data are used as complementary to each other. Selection of a particular source of data or determining the proper blend of the sources depends upon the following factors.

*i. Nature, object and scope of investigation*

The choice between the primary and secondary data largely depends upon the nature of investigation. If the nature of the investigation requires collection of information from each unit by census method, primary source of data should be selected. The object and scope of investigation also influence the source, because the source selected must be suitable, adequate and appropriate to the predetermined purpose of investigation. If the scope of investigation covers a wider area, then restrictions must be placed on the primary data.

*ii. Availability of finances*

Availability of finances is a an important factor which influences the selection of the source of the data. Collection of primary data requires large amount of money for planning, preparation of forms, training of investigators and organising the programmes. As the secondary data is readily available, it does not involve much expenditure. Therefore, secondary data is preferred,if the financial resources are meagre. Where finances are in abundance, primary data is generally preferred for its reliability and accuracy.

*iii. Availability of time*

Time is the most crucial factor in all statistical investigations. The source of data has to be selected by keeping in view the time available at the disposal of the investigator to complete the investigation and the time required to collect the data. Often, in certain statistical investigations, time elapses before the final results are processed and presented. For example,if a company wants to know the demand for air-coolers in summer, the data has to be collected from respondents

all over the country. But by the time the data are collected from primary source, the season may change and the purpose of the investigation may not be served. Therefore, selection of primary source has to be done keeping in view the time available at the disposal of the investigator. Where time is not adequate, secondary data may be used.

*iv. Degree of accuracy desired*

The extent of precision required in the data will determine the use of primary data or secondary data. Although the utmost degree of accuracy may not be practicable in statistics, the data must provide a reasonable degree of accuracy. Due to various limitations, secondary data cannot exactly provide a high degree of accuracy required for a particular investigation. Primary data, if systematically collected, provides more accurate information. Thus, in statistical investigations where a higher degree of accuracy is expected, the primary source of data can serve the purpose.

*v. Agency collecting the data*

The collecting agency has considerable influence on the choice of the source. If the Government body collects the data, it can naturally afford to bear the heavy costs and can go in for the primary data. Where an individual or private organisation conducts the investigation and if the data is to be collected from a wider geographical area, secondary sources are normally relied upon because of the limited finances and time available for them. However, if the data to be collected is confined to a small area and sample, primary sources may be used for such investigations.

No source of data will, however, serve the entire objective of the investigation. While collecting the data, a balance has to be maintained in deploying a particular source, considering its limitations.

## 4.7 METHODS OF COLLECTING PRIMARY DATA

primary data may be collected by the following methods:

i) Direct personal interviews.

ii) Indirect oral interviews.

iii) Information from correspondents.

iv) Questionnaires.

v) Schedules.

*i) Direct Personal Interviews*

In direct personal interview method, the data has to be collected by the investigator through personal conctact with the respondents. This method is particularly useful in collecting information regarding various qualitative and behavioural aspects. The investigator can personally

observe the reactions of the respondents and assess their attitudes. This method ensures correct and reliable information, if the investigator does not exercise any bias. In addition, for the success of this method, the investigator must possess the following qualities.

(a) The investigator must be skilful in extracting the reliable data from the respondents. Sometimes the data relating to personal matters may not be revaled directly and easily. In such cases, the investigator must use his intelligence and skill to place the respondents in good humour and collect the relevant information.

(b) Respondents, when cross examined or flattered, may depart from the actual data. Under such circumstances, the investigator should be tactful in getting the correct and actual data. In the absence of tactfulness, he may tend to collect irrelevant information.

(c) The investigator should not have rigid beliefs regarding religion, caste , color, sex, etc.

(d) The investigator should be polite, courteous and must be capable of adjusting himself to the temperament of the respondents. He must also acquaint himself with the conditions prevailing in the area of the respondents.

(e) The investigator should not possess bias or pre-determined notions. This will effect the collection of accurate data.

Direct personal interview method has the following advantages:

(a) This method helps the investigator to collect original, accurate, relevant and adequate information, because of the face to face interaction with the respondents.

(b) Respondents are likely to be affected by the personal nature of the questions and therefore, have a tendency to supply biased information. The interview method enables the investigator to check this inaccuracy because he can act skilfully and intelligently and explain the object of the investigation and get correct information.

(c) Under this method, the investigator can adopt the language suited to the respondents in order to get relevant data.

(d) This method ensures flexibility. The investigator can make necessary adjustements in the process of collection of data in order to get desired information.

The limitations of this method are given below:

(a) This method is suitable only if the scope of the investigation is limited. A wider scope of investigation restricts the use of this method as it is time consuming and costly.

(b) The use of interview method is limited by the personal bias of the investigator. In this regard, Prof. W.I.King says, "This type of enquiry, while admirable because of additional accuracy due to personal supervision, must not cover too narrow a field to be representative and is also liable to be too large, an injection of the personal element. The prejudices and desires of the investigators become too often unconsciously woven into the fabric of his conclusions".

(c) If the investigators are not adequately trained before the collection of data and if they

do not posses the neccessary qualities like skill, intelligence, honesty, courage and diplomacy, they may fail to collect the relevant data in the face to face situations.

This method is suitable to (a) investigations covering limited scope of area, and (b) investigations which are supported by huge resources of money, time and personnel.

*(ii) Indirect Oral Interviews*

Statistical data is collected with the help of Indirect Oral Interviews when the nature of data involves complexity and the respondents are relevant and unwilling to provide the information directly. Under this method, the investigator conducts interviews with several third party informants, who are supposed to have the knowledge about the problem under investigation. This method can be successfully used under the following conditions:

(a) The respondents approached indirectly must have full knowledge about the matter of investigation.

(b) The investigator must possess the neccessary talent, intelligence, skill, honesty and courage in eliciting the required responses.

(c) The respodents should not be influenced by the tactics used by the investigator to get biased information.

This method of collecting data offers the following advantages:

(a) Collection of data in a face to face personal contact situtation will help the investigator to collect relevant, adequate and accurate information.

(b) This method will cover a wide geographical area and involves less time and money for the investigation.

(c) Neccessary expert opinion can be collected to carryout the investigation more effectively and correctly.

This method suffers from the following limitations:

(a) The data collected through this method is indirect in nature and hence the accuracy of the data depends upon the personal qualities of the investigator and the respondents.

(b) It is possible to inject the personal bias of the respondent and investigator into the data since no direct supervision is made.

This method of collecting data is relevant and suitable (a) where the original source of getting the data is not available, and (b) when the nature of data such that, the investigator cannot ensure adequate data if the respondents are contacted directly.

*(iii) Information From Correspondents*

In this method, persons known as correspondents or local agents are appointed at different places to collect data. These correspondents gather information regularly and send the same to

their central office for futher processing and analysis. Correspondents are normally paid on a fixed salary basis.

This method has the following advantages:

(a) Collection of data is economical and covers a wide area.

(b) The data is collected at regular intervals and ensures continuity.

This method has the following limitations:

(a) Personal bias of the correspondents may effect the accuracy and adequacy of the data.

(b) Correspondents may neglect to report the data.

(c) Uniformity in the collection may not be possible, due to the adoption of different methods of collection of data by different correspondents.

This method can be suitably adopted in cases (a) where regular flow of information is required, and (b) where a high degree of accuracy is not needed and rough estimates are sufficient for the purpose of enquiry.

*(iv) Questionnaires*

Collection of data through questionnaires is an important and popular method. Under this method, the data is collected with the help of a particular form containing a number of questions which are designed to collect the neccessary data relating to the object of the investigation. This form is called 'questionnaire'. Thus, a questionnaire is a printed form, containing a list of important and pertinent questions relating to a problem. The questionnaires are sent through post to the respondents with a request to fill them up and return them to the investigator. If neccessary, the respondents are to be assured that the data collected from them will be kept confidential.



The questionnaire method has the following advantages:

(a) This method is useful for covering a wider geographical area.

(b) This method is relatively economical as the data can be collected with minimum use of resources such as money, time and personnel. In the words of Illersic, "This method possess the apparent advantages that a very large field of enquiry may be covered at relatively low cost".

(c) It enables the investigator to collect original data from the respondents.

(d) Data can be collected on a continuous basis.

(e) It avoids personal bias of the investigator, since the questionnaires are filled by the respondents themselves.

(f) Data collected through questionnaire method ensures more accuracy due to its large coverage.

However this method has certain limitations as mentioned below:

(a) Questionnaires can be administered to only educated people. Therefore , its coverage is restricted to the literate group in the population.

(b) Most of the informants may not respond in filling up the questionnaires mailed to them. Many persons may not return the questionnaires after its completion. Some times, they may supply vague, incomplete and unintelligible data which may not serve the purpose of investigation.

(c) Some times, the respondents may give inaccurate information. There is no scope to check these inaccuracies personally in this method as the investigator has no personal contact with the respondents.

(d) This method lacks flexibility. Once the questionnaires are posted, the investigator loses control over the responses. There is no possibility of asking supplementary questions eliciting the required information for the unanswered or partially answered questions as in the case of personal interview method.

Questionnaire method is suitable to (a) the government agencies, which can statutorily compel the respondents to provide the data (b) to obtain the primary data, to enlarge background of a problem or for verifying the accuracy of secondary data and (c) when the respondents are well educated and know the value of information supplied for the purpose of investigation. This method is not suitable when the data required is complex and confidential in nature. For a model questionaire see at the end of the unit.

*(v) Schedules*

Collecting the data through schedules is considered complementary to the questionnaire method. Under this method a group of investigators, also known as enumerators are asked to collect data through schedules. A schedule is a form containing specific questions, relating to the problem under investigation. The schedules are personally served on the respondents and their answers are recorded by the enumerators.

Schedule method has the following advantages:

(a) It ensures reliability and accuaracy of data as the investigator can have personal contact with the respondents.

(b) This method can be conveniently applied to collect data from uneducated persons.

(c) The personal bias of the investigator may not have much influence on the data as the enumerator has to confine himself to the questions designed in the schedule.

(d) The rate of non-responses can be minimised as the enumerators can clarify the doubts,and convince the respondents to give the information.

This method has the following limitations:

(a) Direct approach to the respondents is a time consuming process.

(b) Schedule method is very expensive, since it requires selection and training of investigators.

(c) This method can be succesful if the investigators are skilled,well trained and intelligent.

Schedule method is suitable (a) particularly in collecting data from illiterate masses and when there are vast resources at the disposal of the investigator, and (b) this method is considered to be more practicable and is widely accepted because of the high rate of responses due to personal interaction of the investigator.

Data can also be collected through telephone and correspondents.

## 4.8 DIFFERENCES BETWEEN QUESTIONNAIRE AND SCHEDULE

Though marked differences do not exist between questionnaire and schedule, the following points are worth mentioning.

| | *Questionnaire* | *Schedule* |
|---|---|---|
| (a) | Administered through post. | Administered personally by investigators. |
| (b) | Answers are recorded by respondents. | Answers are recorded by investigators. |
| (c) | This is an indirect method of collecting information. | This is a direct method of collecting information |
| (d) | The cost of collecting information is low | The cost of collecting information is high |
| (e) | It is suitable where a wider geographical area is to be covered. | It is suitable if a limited geographical area is to be covered. |
| (f) | It can be applied to educated respondents only. | It can be applied to both educated and uneducated respondents. |

The following precautions are to be taken to get a high rate of response through questionnaires or schedules.

(a) The questionnaire must be carefully designed and printed.

(b) Proper selection, training, supervision of personnel or investigators and test-checking of questionnaires/schedules ensures collection of accurate, adequate and reliable data.

(c) Respondents must be offered some incentives such as supplying a free copy of the results of investigation, free gifts, concession coupons,etc.

(d) As far as possible, the sample size must be attached to ensure the return of the questionnaire.

(e) Pre-paid postage and envelop must be attached to ensure the return of the questionnaire.

(f) If necessary, follow-up letters or personal visits must be made to reduce the high rate of non-responses.

## 4.9 CONSTRUCTION OF A QUESTIONNAIRE OR SCHEDULE

Questionnaire or schedules occupy an important place in the collection of primary data, since the success of the collection of relevant, adequate and accurate data, depends upon the careful and cautious preparation of the questionnaire or schedule. While designing the questionnaire or schedule the investigator may face certain problems. Arthur Kornhauser has classified the

problems in drafting a questionnaire as:

(a)Decision regarding question content

(b)Decision regarding question wording

(c)Decision regarding the form of response to the question

(d) Decision about the place of the question in the sequence.

Although no hard and fast rules exists for the designing of the questionnaire the following essential general principles can be followed:

(a) The number of questions in a questionnaire must, as far as possible, be minmum, and at the same time, they must fully cover the object and purpose of the investigation. Answering a lengthy questionnaire may be viewed as time consuming, boring and a tedious exercise.

(b) Questions should be prepared in simple, clear and straight terms, so that, the meaning may be understood in its proper perspective. A clear and straight question which is easily understood by the respondents, may help in getting relevant and accurate data. When technical terms are used in the questionnaire, they must be clearly defined, so that respondents get a clear meaning of the terms. As far as possible, questions should be courteous and non-offending.

(c)Proper wording and placement of questions would help to understand them clearly and unmistakably by the respondents. This ensures validity of the answers

(d) The questionnaire should be designed in such a manner that the questions fall into a logical sequence. This will enable the respondent to understand its purpose and also improves the quality of his answers. As far as possible, identical group of questions should be arranged at one place. This will facilitate the tabulation and leaves no chance for ommission or duplication.

(e) Answers to the questions should not call for lengthy explanations or calculations.

(f)As far as possible, simple and multiple choice questions are to be asked. Simple questions may be framed by suggesting possible answers and the respondents may be requested to choose one among the alternatives. These questions may have either two alternative answers, eg., 'yes or no' or multiple alternative answers.

(g) When the answers cannot be indicated in alternatives, open questions may be framed. For these questions, the answers will be open and respondents are free to give any lengthy answer. Here no restrictions are imposed on the respondents. They are free to express themselves in their own language.

(h) When specific information is required, the questions must be simple and direct. They should be asked only when the respondents are capable of giving correct answers.

(i) Personal questions which affect the pride and sentiments of the people should be avoided. As Prof.Horace Secrist said, "If difficult and unfamiliar question or questions which in any way incite distrust or suspicion, are asked, answers are likely to be either incomplete, brief, non-committal, general or puposely evasive".

(j) Necessary instructions for filling up the questionnaire must be given to guide the respondents.

(k) The investigators must enclose a covering letter, which should mention the following aspects:

(i) Name, address and other details of the investigator

(ii) Objectives and scope of enquiry

(iii) Definitions of various terms and

(iv) An assurance that the information will be kept confidential

(l) Certain corroborative or cross questions may be asked to check the correctness and consistency of answers.

(m) While designing the questionnaire, method of tabulation and processing should also be kept in mind.

(n) Before administering the questionnaire to the target respondents, it is necessary to carry out a pilot study. This helps the investigator to find out the draw-backs of the questionnaire and to correct them. The pilot study also helps in getting an idea about the extent of non-response and securing greater co-operation by redesigning the questionnaire. Pre-testing requires skill, caution, and care, otherwise the designing of the questionnaire may be affected. For good results, proper testing, revising and retesting are desirable.

## 4.10 SELECTION OF APPROPRIATE METHOD OF COLLECTING PRIMARY DATA

All the methods of collecting primary data have certain advantages and limitations. No single method may be suitable to any type of investigation. Therefore, while choosing a particular method of collecting data, the factors like the object and scope of investigation, nature of investigation, availability of resources, degree of accuracy, etc., are to be carefully considered. Though these factors guide the selection of appropriate method, personal qualities like skill, experience and commonsense of the investigator willl help to a greater extent in collecting the data. As Prof. A.L.Bowley has righly pointed out, "In collection ....commonsense is the chief requisite and experience the chief teacher".

## 4.11 COLLECTION OF SECONDARY DATA

Due to the constraints of money and time, collection of primary data may not be possible under all the circumstances. Therefore, at times, the investigator has to depend upon the data collected from secondary sources. For collecting secondary data, no specific methods or techniques are presecribed except that it is collected from a source. The sources of secondary data may broadly be classified into two categories, namely ;

4.11.1 Published sources and

4.11.2 Unpublised sources.

## 4.11.1 PUBLISHED SOURCES

Statistical data collected by different organisations or agencies is mostly in published form. In statistical investigations of a general and non-confidential nature, the results are usually published and kept for the use of public. Such published information may become secondary data to those who use them for their investions. Published source of secondary data may be obtained from the following:

*(a) Government Publications:* The government and its official organs collect and publish statistical data pertaining to various fields. These publications are known as official publications. The Office of the Registrar General and Census Commissioner of India, New Delhi, Labour Bureau, Directorate of Economics and Statistics, National Sample Survey Organisation and Central Statistical Organisation, are some of the official agencies which are collecting and publishing data regularly and periodically. Some of the important official publications are, Monthly Statistics of Production, Annual Survey of Industries, National Income Statistics, Vital Statistics Report, Administrative reports of various Departments, etc. Some times the Reports of the Committees appointed by the government constitute an important source of information. For example, Agricultural Price Commission, Pay Commission, Land Reforms Committee, etc.

*(b) Publications of semi-government organisations:* Certain orgnisations enjoying government patronage conduct statistical investigations and the results of such investigations are published in the form of reports. Such reports are known as publications of semi-government organisations. For example, certain organisations like the Statistics Department of Reserve Bank of India, The Institute of Economic Growth, The Institute of Foreign Trade, Municipal Corporations, various District Boards,etc., publish their reports which provide basic data for in-depth studies in their fields.

*(c) Publications of research Institutions:* Various research institutes publish reports on the projects undertaken by them. They may also publish data in the research journals which constitute an important source. Indian Statistical Institute, Institute of Applied Man Power Research, Institute of Labour Research, various Universities, etc., are some of the examples of research institutes bringing out publications.

*(d) Other institutional sources:* Various public and private, commercial and financial institutions such as the Institute of Chartered Accountants of India, Trade Unions, Stock Exchanges, Co-operative Societies, State Financial Corporations, Banks, etc., also publish reports on aspects related to their activities.

*(e) News papers and periodicals:* A variety of news papers and periodicals also report data concerning various fields. The Economic Times, The Financial Express, Commerce, Captial, Lok Udyog, etc., are some of the example of periodicals and journals that report statistical data relating to socio-economic, trade and financial aspects.

*(f) Reports of international institutions:* Certain international organisations publish statistical data relating to various global aspects of economics, trade, finance, etc. Such Institutions and publications include, I.L.O.(International Labour Organisation), I.M.F.(international Monetary Fund), W.H.O.(World Health Organisation), International Statistical Education Institute, etc. and U.N.O. Statistical Year Book, U.N. Demographic Year Book, etc.

### 4.11.2 UNPUBLISHED SOURCES

All the statistical data collected may not be available in published form but it can be used as secondary data. Such sources constitute the internal records of private organisations, results of research carried out by individual researchers, etc. The investigator has to visit their offices and take down such information from documents and records.

## 4.12 PRECAUTIONS IN USING SECONDARY DATA

While using secondary data, greater care must be exercised since already collected data may not be suitable to the present study because of inadequate sample size, errors in definitions of units, errors of substitution, arithmetical errors, etc. Therefore, secondary data should not be accepted at its face value. While using such data, the investigator must thoroughly scrutinise and satisfy himself regarding its reliability, adequacy and suitability. As pointed out by Wessel, Willett and Simone,"Greater care, nevertheless, should be exercised in using secondary data. The purpose of the investigation that led to the collection of the data, in the first place and the definition of terms employed, should be known, in order to ensure applicability to the problem at hand. It addition, the nature and repulation of the collecting agency should be considered. Needless to say, the investigator must have sufficient confidence in the integrity of his source to use their data without fear of undue bias or misrepresentation". The following precautions are necessary while using the secondary data:

(a) The investigator must carefully examine whether the secondary data is suited to the object and scope of present investigation.

(b) The reliability of the data must be examined by the investigator. Reliability means the extent to which the data can be substituted in the present investigation. Reliability of the data can be tested with reference to the collecting agency, type of enquiry, methods of collection, definitions of units, etc. Simon Kuznets also expressed the same opinion regarding the reliability of the secondary data. To quote, " The degree of reliability of a secondary source is to be assessed from the source, the compiler and his capacity to produce correct statistics and the users also, for the most part, tend to accept a series, particularly one issued by a government agency, at its face value without enquiring its reliability".

(c) The secondary data collected must be adequate and its coverage must suit the object and scope of enquiry. If the data is inadequate, the conclusions drawn on the basis of such data may be faulty and worthless.

## 4.13 EDITING THE DATA

Often, the data collected through primary or secondary sources may be irregular and unintelligible. Such data must be made suitable through editing. Editing is a process of refining the data for proper utilisation. This brings consistency , uniformity, completeness and accuracy in the data. Editing the data is necessary because during the process of collection , intentional or unintentional bias may influence the investigator to collect irrelevant and unnecessary information, or certain errors may creep into the collected data due to misunderstanding of questions, units, sample, etc. Editing requires careful insight into the data. As observed by Crum, Patton and Tebbutt, " The process of editing is by no means an unimportant and routine operation, latter, it requires marked ability, scrupulous care and rigid adherance to scientific objectivity". Data collected through various methods is edited for uniformity, reliability, adequacy and approporiateness. The process of editing for these aspects is explained below:

### (a) Homogeneity

Editing for homogeneity or uniformity implies, that data collected must be fit for uniform interpretations. Heterogeneous elements must be identified and eliminated to make the data homogeneous. Otherwise, conculsions drawn from such data may be incomaparble. In case of primary data, uniformity should be maintained in the answers of different questions. For example, different answers may appear for a question on wages of workers like weekly wage, monthly wage,etc.

### (b) Reliability

Editing for reliability is the most difficult task of the investigator. The data is considered reliable when it is accurate. Accuaracy of data depends on the skill, and personality of the investigator and the respondent. Conclusions based on unreliable and inaccurate data can never be correct. Certain types of inaccuracies like arithmetical errors can easily be found out and rectified, but it is difficult to identify and verify the faulty information collected or supplied. Hence, full care and caution must be exercised to remove unreliable elements and make the data more meaningful.

### (c) Adequacy

Adequacy implies that the data collected should be complete in all respects and suited to the object and scope of investigation. In the case of questionnaire or schedule method, care must be taken to see that the questions are answered completely. Unanswered questions must be tried again to get complete data. The data must be adequate to the objective and scope of the present investigation, when secondary data is expected to be used.

### (d) Consistency

There should be no contradiction or vagueness in the collected data. In many instances,

certain cross questions are asked to check the information. In such cases there should not be any contradiction. For example, if the answer for the question ' Do you live in your own house' is 'yes', and for another question 'How much rent are you paying', if the respondent mentions the figure Rs. 300/-, it conveys an absurd meaning. Such answers should be carefully editied to make the data consistent.

## 4.14 SUMMING UP

Data can be collected from a primary or a secondary source or both. A Primary source is the original source from which first hand information is collected. Data collected from the primary source ensures accuarcy and reliability, but requires huge manpower and financial resources. Among the methods of collecting data from a primary source are: direct personal interviews, indirect oral interviews, information from correspondents, questionnaire method and schedule method. A secondary source gives information which is already collected and processed by some other agency or individual. Secondary data may be collected from two sources, viz; published and unpublished. Collection of data through secondary source can save time, money and manpower resources of the investigator, but the reliability and accuracy of such data need to be carefully examined. As the data collection from various sources is unorganised and unintelligible, careful editing is necessary. Editing helps to eliminate hetrogeneous and unnecessary elements from the collected information.

## 4.15 CHECK YOUR PROGRESS : MODEL ANSWERS

| 1. *Primary Data* | *Secondary Data* |
|---|---|
| i) Collected from its original source. | Collected from already collected sources. |
| ii) Data is original in its form | It is refined in its form |
| iii) The definition of statistical unit is altered to suit the objective of investigation. | Already used definitions of units are adopted for investigation. |
| iv) A fresh plan is prepared for investigation. | No fresh plan of investigation is required |

## 4.16 MODEL EXAMINATION QUESTIONS

**A. Short Questions.**

1. Define the following:
   (a) Primary Data.
   (b) Collection of Data.
   (c) Secondary Data.
   (d) Editing the Data.
2. Outline the methods of collecting primary data.
3. What is a Questionnaire ?

4. What is a Schedule ?
5. What are the sources of secondary data?
6. What are the precautions to be taken while administering the questionnaire?
7. Explain the advantages and limitations of secondary data?
8. Distinguish between a Questionnaire and Schedule.
9. State the advantages of the Questionnaire method.
10. What are the precautions needed while editing the primary data?
11. What is meant by editing the data for homogeneity?

**B. Essay Questions**

12. Briefly explain the various methods of collecting primary data.
13. What is a secondary source? What are the cheif sources of Secondary data?
14. Explain the factors affecting the selection of primary data or secondary data
15. Critically examine the direct personal interview method of collection of data.
16. Discuss the advantages of the Questionnaire method. Under what Circumstances is this method suitable?
17. What are the essentials of a Questionnaire?
18. What precautions do you take while accepting the secondary data?
19. Why do you edit data ? Explain the process of editing data.
20. Draft a Questionnaire to elicit the responses regarding the working conditions of workers in a Cloth Mill.
21. Critically evalutate the questionnaire method of collecting information. Why do you prefer this method?

## 4.17 RECOMMENDED BOOKS


1. Gupta, S.P. : "Statistical Methods",Sultan chand & Company, New Delhi.
2. Gupta, B.N. : "Statistics", Sahitya Bhavan, Agra.
3. Gupta, S.C. : "Fundamentals of Statistics", Himalaya pub. House, Bombay.
4. Simpson and Kafka : "Basic Statistics", Oxfornd and I.B.H. publishing Company, Calcutta.


## 4.18 GLOSSARY

1. Census method : Under this method all the units of populations are observed.
2. Questionnaire : Aform containing a list of questions relating to the problem under study is called questionnarie. These questions are answered by the respondents in their own hand writing
3. Sample method : Under this method a few units of the population are observed.

4. Schedule : A form containing a list of questions which is filled by the enumertors in the face to face contact with the respondents.

## 4.19 MODEL QUESTIONNAIRE

"Questionnaire to farmers for collecting information about method of Sale and Price determination in Regulated Markets".

Village ..... Market Committee.....

**(A) PERSONAL BACKGROUND.**

1. Name :
2. Age :
3. Caste :..... FC/BC/ST/SC.
4. Educational Qualifications :
5. Occupation :
   (a) Primary
   (b) Secondary
6. Size of Family
   (A) i) Adults
   ii) Children
   (B) i) Number of earning members
   ii) Number of dependents.
7. Land Holdings (in Acres)

| | Dry | wet |
|---|---|---|
| Total(Dry Equivalent) | | |

a) Extend Owned

b) Extent Cultivated (including leased in and excluding leased out)

**(B) METHOD OF SALE AND PRICE DETERMINATION**

( please tick (√ ) appropriate altenative)

8. (a) Are you satisfied with the present method of sale in the market — Yes/No

   (b) If No what are the defects in the present system?

   (c) Suggest suitable method of sale

9. (a) Are you satisfied with the present timings for bidding? — Yes/No

   b) If No, mention reasons.

10. Do you listen to the announcement of bidding in the market? — Yes/No

11. Do you consult market officials regarding the price, after entering into the market ? — Yes/No

12. Do the traders immediately come to you for bidding ? — Yes/No

13. Does the bidding always take place in the presence of officials — Yes/No

14. (a) If price offered is not satisfactory, what do you do ?

    i) Sell at prevailing prices.

    ii) Take back the produce

    iii) Make use of Market godowns

    iv) Keep with commission agent.

15. (a) Do you feel that prices offered in the market are fair and reasonable ? — Yes/No

    (b) If no, is it

    ( i) Below cost of production

    ( ii) Equal to cost of production

    (iii) Slightly above cost of production but not remuerative.

16. (a) Are you aware of support prices fixed by the Government for important commodities. — Yes/No

17. (a) Do you suspect price rigging in the market? — Yes/No

18. (a) Did you ever sell produce to institutional purchasers? — Yes/No

19. (a) How do you rate the services of A.M.C. ( Agricultural Market Committee).

i) Very good

ii) Good

iii) Satisfactory

iv) Bad

v) Very bad

(b) Give reasons for your rating.

# UNIT - 5 : SAMPLING TECHNIQUES

**Contents**





## 5.0 AIMS AND OBJECTIVES

The aims of this unit are; to discuss the census and sampling methods of collecting data, their application, methods, merits and limitations. Further it also describes the nature and types of statistical errors.

After going through this unit, you should be able to:

i. Explain the meaning of census and sampling methods.
ii. Explain the theories of sampling
iii. Describe the essentials of sampling
iv. Identify the points to be considered while selecting sample

v. Categorise the techniques of sampling

vi. Classify the statistical errors.

## 5.1 INTRODUCTION

Statistical data can be collected either by census method or by sampling method. If it is needed to study all the objects of a universe, census method is followed. On the other hand, if it is proposed to study only a few objects of the universe and draw conclusion concerning the entire universe, sampling method is followed. If the census method is not possible or difficult to execute, sampling method is followed.

## 5.2 CENSUS METHOD

Census technique of collecting data means collection of information from each and every unit in the population or universe. Here, population or universe refers to the total observations which are included in the scope of a statistical investigation. Thus, according to Ya-lun-Chou, the word "population" in statistics refers to "the aggregate of individual items, whether composed of people or things, that are to be observed in a given problem situation". The individual observations are known as "elements" or "items". For example, if a study regarding the wage pattern of workers in the Singareni Colleries Company Limited includes all the workers of the organisation, such number is known as population.

Census method has certain merits and limitations.

*(a) Merits*

(i) As the information is collected from every unit of the population the results tend to be more accurate and reliable.

(ii) Census technique helps to conduct a detailed study of the universe on the basis of complete and detailed data.

*(b) Limitations*

(i) Census method of enquiry involves time, huge expenditure and a large number of trained personnel. As such small organisations cannot afford to use the census method. This technique is not suitable if the results of investigation are immediately required.

(ii) This technique can be applied only when the number of units in the population can be measured. In case of infinite population where the number of units cannot be measured, census method cannot be applied.

*Suitability*

The census method is suitable in the following circumstances:

(a) When the population or universe is small, and there is no alternative technique except the complete enumeration technique.

(b) Where accurate results are needed for investigation.

(c) Where the population contains heterogeneous elements.

(d) When the resources at the disposal of the investigator are large.

(e) When there is no urgency for results.

## 5.3 SAMPLE METHOD

This technique was first used by Dr. A.L.Bowley in 1912 to study the extent of the poverty of labourers. In 1934 the Indian Government also adopted the technique of sample survey on the basis of recommendations made by the Bowley Robertson Committee.

In contrast to the census method, sample method studies a part of the total elements of the universe and the results obtained on the basis of study are applied to the universe from which the sample is drawn. In sampling technique, the data are collected from a few representative units selected from the large population for making logical inferences about the parent population. Thus the sampling procedure involves (i) selection of representative sample, (ii) collection of information and analysis of data and (iii) application of appropriate statistical techniques for drawing inferences about the population.

**Merits and Limitations of Sampling**

The merits and limitations of sampling techniques are explained below:

*(a) Merits*

(i) As the data is collected from a part of the universe, it is economical. Though, the cost of collection of data per unit will be much higher the total cost of collection of data is lower than the cost involved in complete enumeration method. The data can be collected with a few investigators. The cost of organising the investigation is low in this method.

(ii) Under this method time can be saved not only in collection of data but also in processing and analysis. Thus, it does not require as much time as the census technique.

(iii) Despite the limitation of the fact that the data are collected from a few units, the existence of sampling and non-sampling errors can be reduced by employing trained investigators.

(iv) The sampling method has a wide scope in terms of coverage of information , since the number of units to be covered would be small.

(v) Often, the sampling method is used to test the accuracy of the results obtained by the complete enumeration method.

*(b) limitations*

(i) Sample technique does not yield reliable results without careful planning and design. If the selection of a sample is not based on scientific principles, it may not be representative of the population. Such samples give misleading and unreliable results.

(ii) Collection of data from a sample requires employment of trained personnel. If trained people are not employed, there is a possibility of bias affecting the data collected and the final results.

(iii) As there is no organisational set-up for sample investigations, it may involve undue time and expenditure.

(iv) The selection of sample and determination of sample size pose serious problems. This is particularly true in case of a large population consisting of different elements. Since every technique of sampling has its limitations, it is difficult to select the representative sample.

(v) Sampling technique cannot be applied if the information required is to be collected from all the units of the Universe.

In spite of these limitations, sampling method is widely accepted technique of collection of information. This produces accurate and reliable results when the representative sample is selected adequately at random.

## 5.4 THEORIES OF SAMPLING

In statistics, where complete enumeration is not possible, conclusions are drawn by studying the sample on the assumption that the sample units are representative of the population. This assumption is based on the following two theories of sampling.

(i) Law of statistical regularity; and

(ii) Law of inertia of large numbers

### 5.4.1 LAW OF STATISTICAL REGULARITY

This theory was developed from the theory of probability which states that the chances of inclusion or exclusion of a particular item are equal. The units in the population will be influenced by various forces. As such they differ from one another. But the variations in each individual unit are not wide. Hence they ensure the properties of the population, if they are selected at random.

According to W.I. King, "The Law of Statistical Regularity lays down that a moderately large number of items chosen at random from a very large group are almost sure on the average to have the characteristics of the large group". The forces that cause the variations in the characteristics of units in population are independent and related to each other. As such, the values will normally concentrate equally above and below the average. This process is called statistical uniformity. According to Ya-lun Chou, statistical uniformity or regularity refers to "the tendency of the measurable characteristics to cluster around some centre of gravity". Thus, "Because of statistical uniformity, if a large random sample is selected, characteristics of this sample will differ very little from those in the population. Because of diversity, if a number of random samples are taken, although quite similar in many respects, the samples will never agree completely with one another".

The law is based on the following assumptions, that

(a) the sample is selected on random basis.

(b) the number of items in the sample are many.

(c) the results are true on an average.

Here, random selection implies that every item will have an equal chance to be included in the sample. And such a sample would fairly represent the universe. The theory occupies the most important place in statistics because the inferences about the universe can be made on the basis of sample. Though this theory is not as definite as scientific principles, it ensures reasonable accuracy if the sample is selected on the assumptions of this theory.

## 5.4.2 LAW OF INERTIA OF LARGE NUMBERS

This law also known as 'Principle of stability of mass data' is developed from the law of statistical regularity. This law is based on the assumption that the larger the size of data, the lesser the fluctuations, because of the cancellation effect. It means that the moderately large number of items would balance the variations of small observations. Thus, the law of inertia of large numbers states that by increasing the size of the sample, the results can be made more accurate. In the words of H.M. Walker, "The tendency of distribution of random samples to resemble the distributions of their parent population more closely as sample size increases is called the law of large numbers". Hence, a fairly large number of observations are to be included in the sample for getting accurate results. But no statistical theory lays down the size of the sample. However, the size of the sample is determined by factors such as cost, time, degree of accuracy, etc.

## 5.5 ESSENTIALS OF SAMPLING

The essence of sampling is to draw inferences about the parent population. Such inferences would be true, if the sample has the following essentials.

*(i) Representativeness*

Representativeness implies that the sample should possess the characteristics of the population from which it is drawn. It ensures drawing valid inferences about the universe. This is possible only when the items are chosen on a random basis.

*(ii) Homogeneity*

Each sample drawn must be a homogeneous subset. It means that there should not be marked deviations in the characteristics of the elementary units. The results of different samples drawn from the same universe must be similar.

*(iii) Independent*

While selecting the sample, every unit must be independent of each other in order to be included in the sample. The selection of a particular item in the sample in a particular draw must not influence the selection of other items in the subsequent draws.

*(iv) Adequacy*

The size of the sample must be adequate to yield accurate results. If a large number of items are included in the sample, they will truly represent the universe.

## 5.6 SELECTION OF THE SAMPLE

Scientific selection of a sample is necessary for arriving at accurate results. While selecting the sample, the following points are considered.

*(a) Type of universe*

The nature and type of universe will decide the sample. A large universe requires a large sample to ensure representativeness and vice-versa.

*(b) Sampling Unit*

Sampling unit or statistical unit is the unit in terms of which the enumerator collects the data. This unit may be a geographical unit, a construction unit, or social groups or individuals. The selection of sample depends on the nature of the unit.

*(c) Source List*

Source list is the list containing particulars of various items in the universe. This enables the investigator to select and identify the samples. Such a list must be exhaustive, accurate and reliable and must give the relevant information. This list has to be prepared cautiously without repetition of units.

*(d) Size of the sample*

Since sample is the basis for drawing conclusions, it must be adequate in its size. The size of the sample denotes the number of observations that must be included in the sample. The selection of the size of the sample is of great importance since a bigger size of the sample is financially not feasible and a smaller size does not represent the universe. Hence, the sample size should neither be too big nor too small. An optimum size must always be determined. An optimum size is one which fulfils the requirements of effeciency, representativeness, reliability and flexibility. However, the following points should be remembered while selecting the size of the sample.

(i) The size of the sample is influenced by the nature and size of the population. A small sample is preferred when homogeneous elements are present in the population, whereas a large sample is preferred if the size of the universe is large and has heterogeneous elements.

(ii) The limited availability of time and finance acts as a hurdle for selecting a large sample.

(iii) If a greater degree of accuracy is needed, large size sample is to be selected.

(iv) The size of the sample must be small for intensive and technical studies. On the other hand, if the data needs to be classified by a large number of classes, the size of the sample must be large.

(v) If the data cannot be generated even from a large number of units, the sample size must be still larger.

(vi) A simple sampling technique requires a large sample. In case of other techniques, a small size also gives accurate results.

Though these factors guide the determination of sample size, it can also be determined by mathematical models.

## 5.7 TECHNIQUES OF SAMPLING

There are two methods of selecting the samples. They are :

5.7.1 Random sampling method

5.7.2 Non-Random sampling method

## 5.7.1 RANDOM SAMPLING METHOD

Random sample, according to W.M. Harper is "a sample selected in such a way that every item in the population has an equal chance of being included". Here, the investigator determines the sample not at his will but by chance. For this reason, random sampling is also called probability sampling. The selection of sample on the basis of random sampling technique provides unbiased and more representative units.

The chief limitation of the random sampling technique is that it is time consuming and tedious. It requires skill and intelligence in selecting the sample. Otherwise, the sample drawn may not be a representative one.

The cost may be heavy for collection of data if the units selected under random sample method are spread over a wide geographical area. Random sampling may be of two types:(i) Simple or Unrestricted Random sampling and (ii) Restricted Random Sampling.

**(i) Simple or Unrestricted Random Sampling**

It is a process by which the units to be included in the sample are decided purely by chance. Here all the units in the population are independent and have equal chance of being selected in the sample. Again, simple random sampling can be selected by two methods.

(a) Lottery or Slip System

(b) Using Table of Random Numbers

**(a) Lottery Method or Slip System**

Under this method, all the units in the population are assigned certain symbols, preferably numericals. The symbol assigned to each unit is written on a slip is rolled. All the slips are mixed thoroughly and the required number of slips are picked up by a blind-fold selection. The units corresponding to the symbols in the drawn slips constitute the sample. The slips must have uniformity in their size, shape, colour, etc.

The chief advantage of this method is that it is simple to understand and easy to practice. It does not require training and involves less cost, time and personnel. As such, it is popularly used in many cases. But the main limitation of the method is that since selection is independent of the charecteristics of the population, it is not practicable in case of large population.

**(b) Using Table of Random Numbers**

The lottery method may cause bias if the selection is not made scientifically. Hence, the appropriate method of selecting random sample is using the table of random numbers. Some times mechanically prepared random numbers can also be used. But in practice there are certain tables which provide random numbers. Some of them are:

*(i) Tippett's Random Number Table:* This was designed in the year 1927. It comprises 41,600 digits selected randomly from British census report. They are arranged in 4 digits of 10,400 sets.

*(ii) Kendall and Smith Table of Random Numbers:* It was prepared in the year 1939 with 25,000 sets of 4 digit numbers.

*(iii) Fisher and Yates Table of Random Numbers:* This was designed in 1938. It constitutes 1,500 sets of 10 digits.

Other standard tables of Random Numbers include Rand Corporation Table of Random Numbers, Rao, Mitra and Matthai Table of Random Numbers, Snedecor's Table of Random Numbers, etc.

The procedure for selecting the random sample by using Table of Random Numbers is explained below.

a) Identify the units with numbers.

b) Take a page from the Table of Random Numbers and take any digit in a row at random and identify the numbers serially upto the required sample size. The corresponding units in the population are included in the sample.

Simple random sampling method ensures unbiased and scientific selection of sample. Such samples are more representative of the population. This method can also save time, cost and personnel. The accuracy can be easily verified. However, the limitations of this method are that it requires sourcelist with up-to-date information which some times may not be available. When units selected are geographically scattered, it involves cost and time for collection of data. This method is not feasible in case of small universe. In such cases, the sample selected may not be a representative one. However this method of sampling is still widely used because of its simplicity and reliability.

**(ii) Restricted Random Sampling**

Under this method, the sample is selected by placing certain restrictions on the units to be selected. These restrictions may be regarding the grouping of characteristics or grouping of identical items of size, shape, colour, etc. The various methods of restricted random sampling are examined below.

**(A) Systematic Sampling**

This is also called Quasi-random sampling. Under this method, units to be drawn into the

sample are selected at evenly-spaced intervals. It means that the units in the population are to be arranged in some order, which may be geographical, chronological, alphabetical or numerical, etc. The first unit is selected by following any of the simple random techniques. Subsequent units are drawn at an equal sampling interval. The sampling interval refers to the absolute ratio of the population size to the total sample size. Thus symbolically

$$K = \frac{N}{n}$$

where,

K = Sampling interval
N = Total number of units in the population
n = Sample size

Though systematic random sampling is treated as simple random sampling, it is not so, since all the units in the sample are not independent of each other. Further, the units in the sample are not drawn on a strictly random basis. Only the first unit is drawn by following simple random technique. This method of sampling is suitable when the elements in the population are known, the list of units is arranged systematically and the selection of items is in the evenly spaced intervals is not deleberate. This is also suitable where there are variations in the population. The exclusive merit of the systematic sampling is that it is simple and convenient to adopt. It saves time and labour of the investigator, since selection of units can be made very easily. This method ensures a more representative sample, if the list of units in the universe is complete, up-to-date and unbiased.

However, this method is not suitable when the units in the universe have similar characteristics at similar intervals. In such a case, the sample selected will be very less representative and biased. Hence, while selecting a particular unit, it is necessary to examine the charecteristics of the specific population units.



**(B) Stratified Random Sampling**

under this methed, various units in the population are divided into certain groups so that each item within every group has common charecteristics. While each group is called a 'strata', the basis on which the grouping is made is called 'stratifying factor'. These stratifying factors may be geografical, sociological, economical, such as age, sex, income, marital status, skills, etc. Here the strata can be conveniently made only when the population consists of heterogeneous elements. For an effective use of the stratified random sampling technique, the following points must be kept in view:

i. The units included in a stratum must be homogeneous.
ii. Different strata should be independent of each other.
iii. There should not be any overlapping of the units and strata. It means that no item in the population should be included in more than one stratum.

For selecting the sample, first, group the items of the population into different strata and determine the size of the sample; then select at random the elements from each stratum to be included in the sample. Stratified random sampling may be;

a) Proportional stratified sampling; and

b) Disproportional stratified sampling

**(a) Proportional stratified sampling**

Under this method, the elements of each stratum to be included in the sample are selected as the ratio of the proportion of the stratum to the sample size. Thus according to Ya-Lun Chou, "In a proportional stratified sampling plan, the number of items drawn from each stratum is proportional to the size of the stratum". If, for example, a sample of 10% out of the total population of 1000 is to be drawn, and each stratum is in the proportion of 1:2:3:4, then the items of sample from each stratum will be as follows:

Total sample = 100

Strata = 100, 200, 300, 400

Each sample will be in the ratio of

$\frac{10}{100}$ $\frac{20}{100}$ $\frac{30}{100}$ and $\frac{40}{100}$

Hence, $\frac{100 \times 10}{100} = 10$. Thus 20, 30, and 40 units are selected from each stratum respectively.

**(b) Disproportional stratified sampling**

Contrary to the proportional sratification, the sample in this method represents various units from the strata equally. Hence, no weightage is given to the representatives of the strata in the total population. For example, if a sample of 100 units is to be drawn from the population which is divided into four strata, a sample of 25 units will be included in the sample from each stratum. As this method does not consider the variations in the strata, an alternative method called an optimum stratification is suggested. Under this stratification, samples are drawn by considering the variability and size of the each stratum .

Stratified random sample has certain merits and limitations which are explained below:

**Merits**

(i) A stratified random sample helps to select a representative sample from the population. It does not leave any strata to be unrepresentative in the sample. Thus, it ensures more representativeness of the sampling units and avoids bias in selection of samples.

(ii) The results of the sample will be more accurate.

(iii) There will be no bias due to non-response because, if any unit selected under the sample is unable to give information, it can be replaced from the same stratum. Thus this method is flexible.

(iv) It saves the time and expense of the investigator when the units selected in the sample are confined to a limited area. This is helpful to get the results in a short period of time.

**Limitations**

(i) The grouping of different units into a homogeneous stratum is a difficult task. Such division may not always be practicable.

(ii) The grouping of items involves much time and cost.

(iii) The information regarding the characteristics of the population and group is difficult to find out and understand.

(iv) A careless selection of strata and the sample may affect the degree of accuracy.

However, this method is suitable when the population consists of heterogeneous elements and greater variations are present in the population. This method can be useful when the characteristics of the population are precisely known. Greater representation of the sample can be ensured by dividing each stratum into different sub-sets.

**(C) Multistage Ramdom Sampling**

Under this method, the samples are drawn by selecting various substrata or clusters from the orginal population. The procedure for selecting the sample is: frist a group is selected by following simple random technique; then, a second stage sample is obtained from the sample of the frist stage, and this process will be continued till the required size of sample is obtained. It is necessary that the sample at each stage may be selected at random with out any personal bias. Multistage sampling is a flexible method of sampling. As the sample is selected by following different stages, it is likely to represent the characteristics of the Universe.It is simple to understand and easy to select the sample. It also saves time and money since the final sample is obtained from the sample originally selected.

But this method does not guarantee accuracy because in the process of selection, certain representative units may be eliminated. The sample selected under this method may not be a representative one. However, this method is suitable in case of extensive investigations. This is also suitable when the identification of each and every unit in the population is difficult.



**Check your progress - 1**

Explain the term proportional stratified sampling with an example.

_____

_____

_____

_____

## 5.7.2 NON RANDOM SAMPLING METHOD

In non-random sampling method, the sample is selected according to the discretion of the investigator. Various methods of non-random sampling are:

a) Judgement sampling

b) Convenience sampling

c) Quota sampling

**(a) Judgement Sampling**

This is also knwon as 'purposive or deliberate sampling'. Under this method the units to be included in the sample are selected by a deliberate judgement or choice of the investigator. The investigator will select those units in the sample, which in his opinion, are representative of the characteristics of the population. The selection of sample under this method requires the following precautions:

i) The investigator must be careful and intelligent in selecting the sample. The characteristics of the entire population have to be thoroughly understood by him.

ii) The sample must be representative of the population.

iii) The investigator must avoid any bias in the selection of sample.

This method has the following merits:

(i) The selection of sample is simple and easy.

(ii) It saves the time and money of the investigator, if a sample is selected from the place which is nearer to the investigator.

(iii) If the sample is carefully selected to represent the universe, the results are likely to be reliable and accurate.

However, this method suffers from the following limitations:

(a) As individual bias may affect the selection of the sample, it may not adequately represent the population. Therefore, the results tend to be unreliable.

(b) Since the selection is made on non-random basis, sampling error cannot be calculated exactly.

(c) The results of the samples drawn on the basis of judgement sampling cannot be compared. Thus sampling stability and representativeness cannot be exactly known.

Though this is not a scientific method, it is widely used in most of the studies due to its convenience.

This method is suitable under the following circumstances:

i) where the characteristics can be readily determined and understood in case of small and limited universe.

ii) When the items included in the sample are small.

**(b) Convenience Sampling**

Under this method, the items included in the sample are selected by mere convenience. Here a convenient group' may be selected as a sample. This group is technically called a 'chunk'. Thus a 'chunk' is a representative group of units selected on the basis of convenience to be included in the sample.

Though this method is very simple and convenient, it has certain limitations. They are:

(i) The sample selected by this method is not representative, since the selection is not based on random techniques. As such the results are also not accurate and precise.

(ii) This method of selection introduces bias in the selection of a sample. Thus there will be more sample errors.

However, this method is suitable and can be adopted where the population is not rigidly defined and the characteristics of the population are not known. This is useful in extensive studies where aproximate results are adequate for final decisions.

**(c) Quota Sampling**

Quota sampling technique is a combination of stratified sampling and judgement sampling. The procedure for selection of items under this method involves dividing the universe into certain groups known as quotas. These quotas are to be selected in such a way that each quota consists of units with specified characteristics. These quotas are assigned to various filed enumerators who are instructed to select a specified number of units within their quotas. While selecting the units, the investigator must apply his judgement, experience, skill and intution. Here, those items of the quota are selected as sample which are considered to be representative of the quota. This type of sampling technique saves time and cost. It is also useful because the strata are based on certain specific charateristics. This method of selection is more representative, if the final selection of the sample is made on random sample basis. This method is known for its suitability when there is a high rate of non-response.. Non-response can easily be eliminated since the field investigators will have the choice of selecting units in a particular quota.

Quota sampling gives accurate, reliable and representative samples if the selection is made by thoroughly studying the characteristics of the population.

This method of sampling has the following limitations:

i) It is affected by the personal bias of the enumerators.

ii) This method suffers from the error of substitution. A situation may arise when there is a high rate of non-reponse and substitute units are selected by the enumerators.

iii) The sampling error cannot be estimated accurately, as the selection is based on non-random technique.

But this method is sutiable in cases of extensive political and economic surveys.

**Selection of Appropriate Method**

Selection of samples by random and non-random methods have their own merits and limitations. No method gives a representative sample, as representativeness is a subjective element. However, the choice depends upon the size and characteristics of the population, size of the sample, nature and objects of the investigation, availability of finances, etc.

## 5.8 STATISTICAL ERRORS

A statistical error is the variation in the value of a sample and the corresponding value of the universe. In statistical investigations, many factors cause variations in the results.The errors arising out of these factors may be grouped into two categories, viz, Sampling errors and non-sampling errors.

## 5.8.1 SAMPLING ERRORS

According to Ya-lun Chou,"the sampling error is the difference between the sample results and that of the census, when both results are obtained by using the same procedure". Sampling error arises because of the fact that a samll part of the universe is examined to estimate the characteristics of the universe. Hence, there will be a variation in the results obtained through sample enumeration and the census enumeration. These errors are found both in random sampling and non-random sampling which occur due to sampling fluctuations.

The causes of sampling error are explained below:

*(i) Bias in selection*

This type of error arises when the selection of sample is influenced by bias rather than scientific reason. The use of inappropriate methods for the selection of sample results in sampling errors. For example, if a judgement sample is used in place of random sample, the sampling error occurs. However, this error can be minimised by adhering to the random sampling techniques.

*(ii) Bias of the Enumerator*

Some times the field investigator may substitute a convenient unit from the population, if the collection of information from the orginal units is not practicable. In this case the error arises when the unit substituted does not possess the characteristics of the unit orginally selected for enumeration. The error may also arise when enumerator resorts to convenient demarcation of the sample from the population and insufficient coverage of the sampling units.

*(iii) Bias of the respondent*

If questions affecting the dignity or sentiments of the respondents are asked during the collection of data, the respondents are likely to give wrong answers. In such cases the errors arise because of faulty answers given by the respondents. For instance, persons are likely to understate their incomes or may try to conceal their habits, etc.

*(iv) Bias in the Collection*

Errors may crop up due to the bias of the investigator in the collection of data. This bias may result due to faulty selection of the problem, ambiguity in the definition of nature and scope, wrong hypothesis, ill-designed questionnaire or schedule, lack of adequate training of the investigators, biased nature of the investigator, poorly designed plan and frame, etc. These errors can be minimised by adopting an appropriate statistical plan of investigation.

*(v) Bias in Analysis and Interpretation*

Biased error in statistics may also arise due to adoption of inappropriate technique of analysis. Since the population parameter is estimated by the sample statistic, error may arise due to wrong method of estimation. Here, the error may also arise due to wrong method of estimation, tabulation, formation of frequency districutions, wrong choice of average, dispersion, skewness, inappropriate technique of estimation, approximations, etc.

*(vi) Heterogeneous elements in the Universe*

Sampling error may also arise due to the existence of heterogeneous elements in the universe.

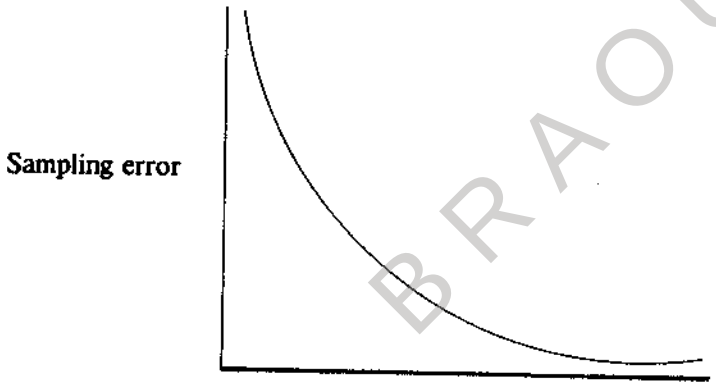
Though theoretically these biased errors are supposed to occur in sampling only, they may also creep into the census method.

Sampling error may also arise due to certain unbiased errors. Unbiased errors are also called compensatory errors. These errors do not increase with the increase in the sample size, but they tend to cancel each other, when they are averaged. These errors may include:

i) Errors due to chance differences between the units of population included in the sample and those not included.
ii) Errors due to approximations which cause over-estimation or under estimation.
iii) Errors in editing the data.
iv) Errors in analysis and presentation.

**Reduction of Sampling Errors**

Biased errors in sampling can be effectively reduced by minimising the bias. But unbiased errors which occur due to random causes cannot be completely eliminated. Hence such errors must be reduced to the extent possible to get the desired degree of accuracy. since the reduction of error is inversely proportional to the square root of the sample size, sampling errors can be reduced by studying large samples. This is depicted in the following diagram.






By including large number of items in the sample, the errors can be reduced upto a particular stage, where further increase in the size of the sample will not reduce the error, as the biased errors are directly proportional to the number of observations. An increase in the sample size may impose heavy burden on the finances and time. In view of these constraints, an optimum sample is to be chosen to achieve the required degree of accuracy.

## 5.8.2 NON-SAMPLING ERRORS

A distinct feature of the non-sampling errors is that they affect both the census and the sample technique of collecting the data. Such errors may arise at any stage of statistical

investigation. However, non-sampling errors are numerous. The following are some of the reasons for non-sampling errors.

i) Faulty description of objectives, scope and location of sample units.

ii) Faults in recording the responses and identifying the statistical unit.

iii) Wrong answers of the respondents.

iv) Inaccurate information resulting form high rate of non-response.

v) Inclusion or exclusion of sampling units in a haphazard manner by the enumerators.

vi) Errors in processing, analysis, coding,tabulation and compilation.

vii) errors in presentation such as printing and proof reading.

Since non-sampling errors affect both the census and the sample results, they have to be effectively controlled. The salient feature of non- sampling errors is that they tend to increase with the increase in the size of the observations. Hence, such errors must be controlled to achieve the desired degree of accuracy in the final results. These errors can be minimised by defining clearly the object and scope of investigation, recruiting efficient and well trained investigators, adopting field supervision and field checks, careful processing, analysis and interpretation of the data.

In statistics, total error constitutes sampling as well as non-sampling errors. According to Ya-lun Chou, the error of a statistical survey is "The square root of the sum of the squares of the sampling error and the non-sampling error.One major concern in sampling is to make the total error as small as possible". A small total error would make the results of the sample more unreliable. The accuracy of the sample results can be examined by (a) comparing the results of two or more sets of samples of same size drawn from the same universe, (b) comparing the results with that of samples where the results of population are known and (c) comparing the results of sub-samples and those of the original sample. However, the reliability of sampling results depends upon the personality of the investigator.



## 5.9 SUMMING UP

Data can be obtained by census method or sample method. Under census method, data is collected from every unit in the population. In sample method, a part of the total units is examined and the results are generalised to the entire population. The chief merit of this method is that it saves time and resources of the investigator and also gives fairly accurate and reliable results. For the selection of sample,random sampling and non-random sampling techniques can be applied. Under the ramdom sampling method, the sample is drawn in such a way that equal chance is given to every item in the population to be included in the sample. Again, in random sampling technique, two methods may be adopted. They are: unrestricted random sampling and restricted random sampling. In case of unrestricted random sampling, the sample is drawn by lottery method or table of random numbers. Restricted random sampling techniques include systematic sampling, stratified sampling and multistage sampling. Non-random sampling techniques include judgement sampling, convenience sampling and quota sampling.

When the sample results are used to estimate the population parameter, some variations may be reported due to sampling error or non-sampling error. while sampling errors occur in sample method of investigation, non- sampling errors occur in both census and sample method of investigation.

## 5.10. CHECK YOUR PROGRESS : MODEL ANSWERS

1. It is the method of drawing sample in proportion to the size of the stratum. Now you should give an example.

## 5.11. MODEL EXAMINATION QUESTIONS

**A. Short questions**

1. Define

(a) sample
(b) Population
(c) statistical error
(d) Total Error
(e) strata

2. Explain

(a) Chunk
(b) Cluster
(c) Quota
(d) sampling Interval

3. What is 'Law of statistical regularity ?
4. What do you understand by the 'Principle of Stability of of mass data' ?
5. What are the assumptions of 'Law of statistical Regularity' ?
6. What is a random sample ?
7. Define random sampling.
8. What is purposive sampling ?
9. How do you test reliability of a sample ?
10. What do you mean by unrestricted sampling ?
11. *What is*

(a) Systematic sampling ?
(b) Judgement sampling ?
(c) Representative sample ?
(d) Bias ?

12. What are the essentials of sampling ?
13. How do you control the non-sampling errors ?

14. What is a population? Distinguish between population and sample ?

15. How do you eliminate sampling errors ?

16. Explain the factors that are considered for determining the selection of the sample.

17. What is proportional stratified random sampling ?

18. Explain the merits and limitations of convenience sampling.

19. How do you select a quota under quota sampling ?

20. Distinguish between quota and strata.

**B. Essay Questions**

21. Explain the merits and limitations of census and sample methods.

22. "The theories of sampling ensures selection of representative sample".

23. Discuss the various methods of random sampling.

24. Distinguish between biased and unbiased errors. Mention the sources of biased errors.

25. Enumerate the various methods of non-random sampling

26. Distinguish between sampling and non-sampling errors. Expalin the causes of non-sampling errors.

27. Distinguish between stratified random sampling and multistage random sampling. How do you select a sample under stratified random sampling method ?

## 5.13 RECOMMENDED BOOKS


1. Gupta, S.p. : "Statistical Methods" Sultan Chand & Company, New Delhi
2. Gupta, B.N. : "Statistics", Sahitya Bhavan, Agra.
3. Gupta, S.C. : "Fundamentls of statistics" Himalaya Pub. House, Bombay.
4. Simpson and Kafka : "Basic statistics",Oxford and IBH Publishing Company, Calcutta.


## 5.14 GLOSSARY

1. Non-sampling errors : Owing to the human participation these errors may arise at any stage of investigation. These are present both in sample surveys and census surveys.
2. Population or Universe : Population or Universe refers to the composition of all conceivably (or hypothetically) possible observations relating to a given phenomenon. It is the totality of objects under consideration.
3. Random Sampling : A method of selecting sample where all the units in the population would have an equal chance of being included in sample.
4. Sample : A portion of the population which is studied to learn about population.
5. Sampling Errors : The difference between sample estimates and census estimates are referred to as sampling errors. They occur due to the observation of a part of population.

## UNIT -6 : CLASSIFICATION OF DATA

**contents**



## 6.0 AIMS AND OBJECTIVES

This unit aims at presenting the meaning, objectives and principles of classification. Further the bases adopted for the classification of data are also discussed.

After reading this unit, you should be able to :

i) explain the meaning of classification of data

ii) identify the objectives of classification of data

iii) describe the principles of classification

iv) list out the various bases of classification



## 6.1 INTRODUCATION

Statistical data collected through statistical investigation are huge and voluminous in mass of figures. The data which is in a raw form may not be understandable and useful for interpretation. The raw data is not suitable for statistical analysis and interpretation. To understand the whole mass of unorganised and complex data further processing is necessary. This is evident from the observation made by A.R.Ilersic who states that, "The statistician's first task is to reduce and simplify the details into such a condensed form that all the salient features may be brought out, while facilitating the interpretation of assembled data. This procedure is known as classifing and tabulating the data". Hence, the collected data is to be divided into different groups or classes. This process of dividing the mass of data into classes and sub-classes is referred to as "Classification of data". It also involves the determination of various class categories or group heads into which the data will be distributed.

## 6.2 MEANING AND DEFINITION OF CLASSIFICATION

Classification is the process of making the raw data more meaningful and useful. This process involves grouping of related items into classes and sub-classes. Grouping of the related items may be either on the basis of similarity of the data or dissimilarity of the data. According to Connor "Classification is the process of arranging things (either actually or notionally ) in the groups or classes according to their resemblances and affinities and gives expression to the unity of attributes that may subsist amongst a diversity of individuals".

Classification of data is a function which can be compared to the sorting of letters in the post office. In the Post Office, the letters collected from various places are sorted out into various lots on locational basis, i.e., Bombay, Calcutta, Hyderabad, Warngal, etc. The sorted out letters will be kept in different bags for different destinations.

Horace Secrist says that "Classification is the process of arranging data in sequences and groups according to their common characterstics or separating them into different but related parts".

Further, Professor A.M.Tuttle is of the view that " a classification is a scheme for breaking a category into a set of parts, called classes, according to some precisely defined differing charactersitics possessed by all the elements of the category".

## 6.3 OBJECTIVES OF CLASSIFICATION

The following are the main objectives of classification.

**i) Simplification of data**

Classification presents the complex data in a condensed form, eleminating the unnecessary details. It also eleminate the complexity of data as statistical data are an aggregate of facts. Classification makes the data simpler, more meaningful and comparable to that of other items. For example, the Government of India conducts population census every 10 years. During the process of data collection, large data are collected and the data collected will not serve any meaningful purpose, unless they are thoroughly processed. To make the data more meaningful and useful for analysis and interpretation, it should be Classified into groups and sub-groups. The grouping of data can be made in accordance with religion, sex, education, age, occupation, etc. This type of condensing the mass data into simple form makes the data more meaningful. Thus classification facilitates the presentation of data in simple, clear, definite, correct and concise manner. This helps to draw valid inferences for scientific decision.

**ii) Arrangement of data as per similarities and dissimilarities**

Classification is the process of arranging the data into groups according to their resemblances and affinities. Similar items are placed in one class and dissimilar items are placed in the other class. For example, the census data collected in respect of various aspects of socio-economic

conditions of the people will not be of much use unless the data are segregated and arranged in accordance with similarities and dissimilarities.

**iii) Facilitation of Comparisons**

The basic objective of statistics is that it should facilitate the comparison between two or more related variables or attributes. Comparability or similarity of objects to be compared must be ensured in the collected data. This can be done through a systematic arrangement of data interms of their basic common characteristics. It is unwise to compare dissimilar things. Classification avoids confusion, ambiguity and facilitates comparison. Comparative data will be more meaningful, understandable, appealing and useful for prompt decision. For example, the data on households, classified on the basis of occupation income, level of education, etc., can be used for making valid comparisons between two or more of these aspects. The comparison may be between level of education and income, level of education and occupation, etc.

**Iv) Establishment of relationships**

Classification of data helps establishment of relationships among various items that constitute the data. For example, the data classified on the basis of income levels and spending habits, or saving pattern enables the readers to understand the relationship between income levels and savings. With the help of these relationships, one can easily ascertain the nature, scope and importance of various items of data. Further, it is very convenient to establish cause and effect relationships of various items of the classified data.

**v) Facilitation of tabulation of data**

The classified data forms the basis for tabulation. It is very easy to present the classified data in precise and appropriate tables. On the other hand, it is very difficult and also meaningless to present the unclassified data in tables

**vi) Facilitation of the statistical treatment of data**

The basic object of classification is to ensure homogeneity and uniformity of data. It provides clarity and intelligibility. Classified data is more acceptable for further statistical treatment, such as tabulation, analysis, presentation, interpretation.

## 6.4 PRINCIPLES OF CLASSIFICATION

Classification is one of the most important stage of data processing. Even though there are no clear cut principles for classifying data, an appropriate method of classification has to be selected, keeping in mind the objectives and scope of the statistical enquiry. However, the following principles may be helpful in classification.

**i) Classification must be exhaustive**

While classifying the data, great care must be taken to see that each and every item of the data are included in appropriate class or group. Grouping of items must be based on their common characteristics. This avoids ambiguity and ensures homogeneity of data. For example, if data, relating to marital status of persons is classified into two classes only, i.e., married and unmarried, that cannot be exhaustive, as other classes such as married but divorced and widowed are not included. This type of classification covers only those persons who are married and unmarried, but it does not contain any classes for including those persons who are married but divorced or widowed. Hence, to remove any further confusion, classification must contain as many classes as possible to incorporate each and every item of data.

**ii) Classification must be mutually exclusive**

Each and every item of the data must be included in only one class. Otherwise the items included in more than one class or group defeat the purpose of classification because of its overlapping nature. Moreover, the decisions taken on this basis of classification will be misleading. For example, classification into literates, illiterates and females is not proper, because females also come under the category of either literates or illiterates. Hence, the proper classification in this regard is to group the population into males and femeals and further dividing the two groups into literate and illiterates. An alternative way for classifying this data is to group the population into literates and illiterates and further dividing the two groups into males and females.

**iii) Classification must be consistent**

The principles and techniques adopted in classification will have to be followed constantly throughout the statistical investigation. Further, care should be taken to see that frequent changes in the principles and techniques are avoided. Otherwise, data loses comparability and misleads the readers, For example, at one stage of the statistical investigation, data relating to incomes of people is classified into higher, middle and lower income groups by defining Rs. 1,000 per month and above as higher income group, Rs. 500 per month to Rs. 1,000 per month as middle income group and below Rs. 500 per month as lower income group. Subsequently, the basis of classification has been changed as Rs. 1,500 per month and above as higher income group, Rs. 1,000 per month to Rs. 1,500 per month as middle income group and below Rs. 1,000 per month as lower income group. If consistency with regard to the system of classification is not followed, it fails to conform to the objects of enquiry.

**iv) Classification must suit the requirements of the enquiry**

The methodology and procedures adopted for classification of data must suit the purpose and requirements of the enquiry. For example, if an enquiry is conducted to assess the performance of students in the examinations, it will be useless to classify the data on the basis of their caste and religion. It would be more appropriate if the data is classified on the basis of marks obtained by them in the examinations.

**v) Classification must be Flexible**

A good classification must be flexible and should adjust to the subsequent changing conditions and circumstances. For example, a detailed classification of data into various groups and further, each group into sub-groups must be quite adoptable to changing situations.

**vi) Classification must be based on homogeneity of items**

All the homogeneous items of the data must be included in one class only. In order to ensure homogeneity, the items included in various classes or groups should be further classified into sub-groups. For example, classification of labour force into employed and unemployed is not adequate to determine the effect of education. In order to be more meaningful, each of these classes should be further classified into literates and illiterates.

## 6.5 BASES OF CLASSIFICATION

classification of data may broadly be of four types.

6.5.1 Qualitative Classification

6.5.2 Quantitative Classification

6.5.3 Geographical Classification

6.5.4 Chronological Classification

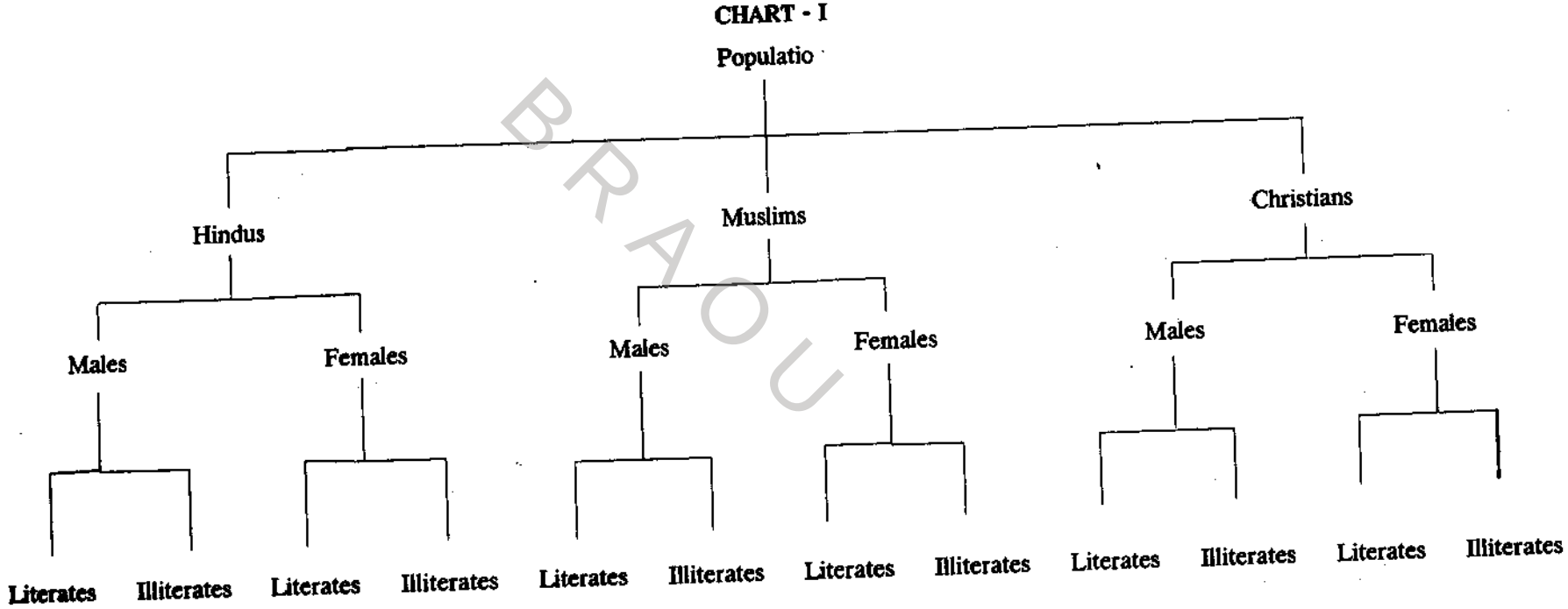
Statistical data are classified on the basis of the characteristics of various items. The characteristics of the data may be descriptive or numerical. Religion, caste, occupation, unemployment, sex and literacy are examples of descriptive characteristics of data. Income, age, weight and height are examples of numerical characteristics of data. Descriptive characteristics of data cannot be precisely quantified, though they can be indirectly measured by means of identifying the presence or absence of such characteristics.



## 6.5.1 QUALITATIVE CLASSIFICATION

Classification of qualitative data is referred to as qualitative classification. The basis of qualitative classification of data is the presence or absence of particular attribute. The Qualities of an attribute can be differentiated by means of some natural differences, which will become the basis for grouping the items. Qualitative classification can be either dichotomous classification or manifold classification. Grouping of qualitative data into two classes is called "dichotomous classification", e.g., classification of population on the basis of sex, into males and females, grouping of qualitative data into various classes and sub-classes is known as "manifold classification", e.g., classification of population on the basis of religion into Hindus, Muslims and Christians and a further classification of them on the basis of sex into males and females and males and females into literates and illiterates. The chart shown on page 82 will help the reader to understand the classification and sub-classification of data better.

**CHART - I**



## 6.5.2 QUANTITATIVE CLASSIFICAION

Classification of data which can be measured or counted in definite terms is called "quantitative classification". In case of quantitative classification , data can be arranged in terms of classes and if necessary, sub-classes. For example, classification of data with respect to income of employees into higher income group, middle income group and lower income group come under this category.

## 6.5.3 GEOGRAPHICAL CLASSIFICATION

In geographical classification, data are classified on the basis of geographical regions or locations. Usually, geographical regions are identified with the help of the existing political boundaries and expressed in terms of countries, states, districts and taluqs. For example, the population of Andhra Pradesh may be presented region-wise as given below.

**Example– 1**

REGION-WISE POPULATION OF ANDHRA PRADESH
(1981)

| Sl. No. | Region | Population (in crores) |
|---|---|---|
| 1. | Coastal Andhra | 2.37 |
| 2. | Rayalaseema | 0.96 |
| 3. | Telangana | 2.02 |
| | Andhra pradhesh | 5.35 |

Source : Census of India, Series -2,
Andhra Pradesh, Paper-1, 1982.

While classifying the data on locational basis, the data is listed according to alphabetical order usually for easy reference. Some times, the data is also expressed in accordance with the size to emphasize the importance of data .

## 6.5.4 CHRONOLOGICAL CLASSIFICATION

Data with respect to a variable relating to different time periods classified according to the time of their occurence is known as " chronological classification". This type of classification is more suitable in presentation of data in respect of sales of a company, population, imports and exports, etc. For example, the population figures of Andhra Pradhesh from 1911 to 1981 are presented below:

POPULATION OF ANDHRA PRADESH (1911-1981)

| Year | Population (in crores) |
|---|---|
| 1911 | 2.15 |
| 1921 | 2.14 |
| 1931 | 2.42 |
| 1941 | 2.73 |
| 1951 | 3.11 |
| 1961 | 3.60 |
| 1971 | 4.35 |
| 1981 | 5.35 |

Source: Census publications of relevant years.

**Check your progress - 1**

List out the bases of classification.

## 6.6 SUMMING UP

Classification is the process of arranging the mass data into classes and sub-classes according to their common characteristics. The main purpose of classification is to simplify the mass data and arrange it according to its resemblences and affinities. The data so classified facilitates the comparison and establishes a relationship among various items of the data. It is also amenable for further statistical treatment like tabulation, analysis, presentation and interpretation. While classifying the data, certain important points like inclusion of every item, avoidance of overlapping in the inclusion of items, consistency in adhering to the principles, adoption of procedure suitable to the purpose of enquiry, flexibility to adjust to the changing conditions and homogeneity as a basis for classification must be taken into account.

Broadly, classification of data may be of four types; namely, qualitaitive, quantitative, geographical and chronological. While the qualitative classification is made on the basis of the qualities of an attribute, quantitative classification is made on the basis of measurable attributes. Geographical classification refers to the grouping of data on the basis of geographical regions or loacations. Chronological classification is made on the basis of the time of occurrence of the data.

## 6.7 CHECK YOUR PROGRESS :MODEL ANSWERS

The following are some of the bases of classification:

i) Qualitative classification.
ii) Quantitative classification
iii) Geographical classification.
iv) Chronological classification.

## 6.8 MODEL EXAMINATION QUESTIONS

**A. Short Questions**

1. what do you mean by "Classification"?
2. what do you mean by qualitative classification ?
3. what do you mean by quantitative classification ?
4. what are the objectives of classification of data ?
5. state the various types of classification ?

**B. Essay Questions**

6. Explain the various bases of classification.
7. Explain the objectives and methods of classification of data giving suitable examples.
8. Explain the principles of classification of data.

## 6.9 RECOMMENDED BOOKS


1. Gupta, S.P. : "Statistical methods", Sultan Chand & Company, New Delhi.
2. Gupta,B.N. : "Statistics", Sahitya Bhavan, Agra.
3. Gupta, S.C. : "Fundamentals of Statistics", Himalaya Pub. House, Bombay.
4. Simpson and Kafka : "Basic Statistics", Oxford and I.B.H. publishing company, Calcutta.


## 6.10 GLOSSARY

1. Chronological classification : The classification according to different time periods.
2. Classification of data : The process of dividing the mass data into classes and sub-classes based on their resemblances and affinities.
3. Geographical classification : Classifying the data according to geographical regions.
4. Qualitative classification : Classifying the data, on the basis of presence or absence of a particular attribute.
5. Quantitative classification : The classification of data on the basis of size.
6. Tabulation of data : The process of condensing the data and arranging such data into columns and rows.

# UNIT - 7 SERIATION OF DATA

**contents**





## 7.0 AIMS AND OBJECTIVES.

This unit aims at explaining the meaning, types of seriation and the bases adopted for the seriation of data.

After going through this unit, you should be able to:

i. explain the meaning of seriation of data

ii. identify the types of series

iii. recognise the basic principles for forming frequency distribution

iv. descibe the methods of forming class intervals

## 7.1 INTRODUCTION

Having collected the data, it would be arranged in some order. The systematic arrangement of such data based on its general characteristics or based on frequency distribution is known as

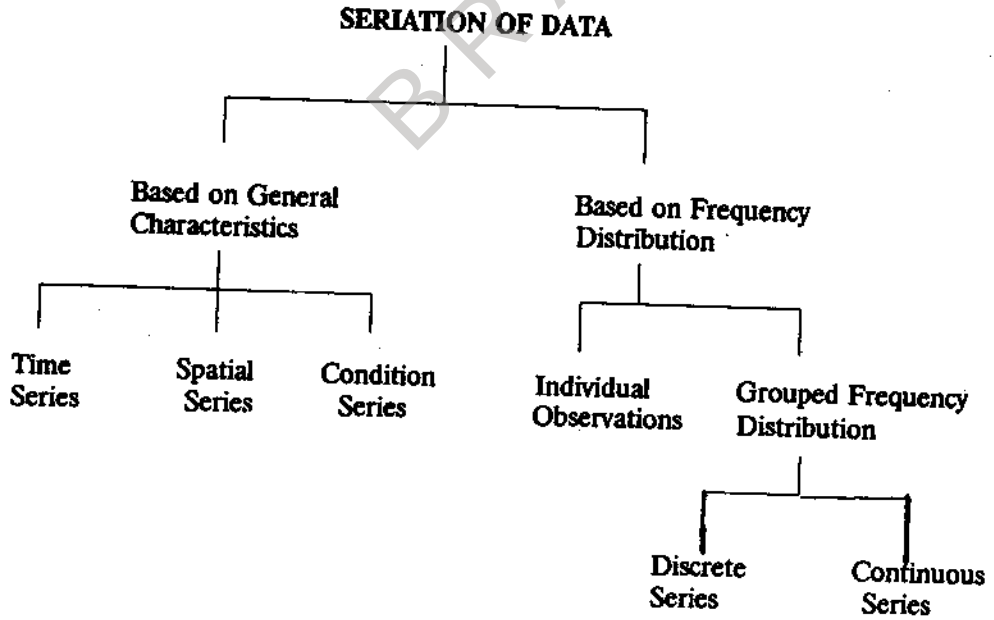
seriation of data. There are various methods of arranging such data. We shall also touch upon the basic principles for forming frequency distribution and methods of forming class intervals.

## 7.2 MEANING AND DEFINITION OF SERIATION

In order to make the statistical data more meaningful and useful, the classified data is to be arranged systematically. This systematic arrangement of data either on the basis of its general characteristics or on the basis of the methods of its construction is referred to as seriation of data always succeeds classification of data. According to Professor Connor, "If two variable quantities can be arranged side by side, so that measurable differences in the one, correspond with measurable differences in the other, the result is said to form a statistical series". As per this definition, seriation involves the arrangement of data relating to two variables. Such variables should be measurable in quantitative terms. This definition is incomplete as its scope is confined only to the data relating to variables. It does not include the arrangement of data which is qualitative in nature. The definition given by Professor Horace Secrist is considered to be a comprehensive one. According to him " A series, as used statistically, may be defined as a thing or attributes of things arranged according to some logical order". A close examination of this definition reveals that seriation is the arrangement of data relating to variables or attributes and the arrangement of data is done on the basis of some logical order.

## 7.3 TYPES OF SERIES

Series can be classified into two main categories, namely, (i) Series based on general characteristics and (ii) Series based on frequency distribution. The various types of series are explained with the help of the following chart.



## 7.3.1 SERIATION BASED ON GENERAL CHARACTERISTICS

Seriation based on general characteristics are further classified as (a) Time Series, (b) Spatial Series and (c) Condition Series.

**(a) Time Series.**

According to Morris Hambrug, " A time series is a set of statistical observations arranged in chronological order". In the words of Wessel and Wellet, when quantitative data are arranged in the order of their occurrence, the resulting statistical series is called a time series". Thus, seriation based on time involves the chronological arrangement of data collected at successive intervals of time. Time unit may be a year, month, week or day. The following is an example of time series.

**Example - 1**

**SALES OF A HYPOTHETICAL FIRM**

| Year | Sales (Rs. in Lakhs) |
|---|---|
| 1974 | 18 |
| 1975 | 23 |
| 1976 | 24 |
| 1977 | 25 |
| 1978 | 27 |
| 1979 | 31 |
| 1980 | 36 |
| 1981 | 43 |
| 1982 | 54 |
| 1983 | 62 |



Data arranged in chronological order helps to understand the past behaviour. This type of arrangement of data is convenient for statistical analysis. It provides a basis for a compartive study of variables over a period of time.

**(b) Spatial Series.**

Spatial series refer to the arrangement of data collected in relation to different geogrphical locations. Strictly speaking, spatial series is not a statistical series as the changes in different regions cannot be quantified directly. The following is an example of a spatial series.

**Example - 2**

**POPULATION OF SELECTED DISTRICTS OF ANDHRA PRADESH**

| District | Population (in Lakhs) |
|---|---|
| Adilabad | 16.4 |
| Ananthapur | 25.2 |
| Chittor | 27.4 |
| East Godavari | 37.0 |
| Guntur | 34.3 |
| Hyderabad | 22.6 |
| Karimnager | 24.4 |
| Khammam | 17.5 |
| Warangal | 23.0 |
| West Godavari | 28.7 |

(Source : Census of India, 1981, Part-2 Andhra Pradesh, Paper-1, 1982).

Spatial seriation of data helps to understand the relative importance of different regions. This may form a basis for future policy making and implementation of various programmes.

**(c) Condition Series**

Data arranged neither with reference to time nor in accordance with place, is arranged with reference to a condition. A condition may be income, expenditure, production, marks, weight, height, etc. The following is an example of condition series.

**Example -3**

**PRODUCTION OF PRINCIPAL CROPS OF ANDHRA PRADESH DURING THE YEAR 1980-81**

| Crops | Production ( in '000 tonnes) |
|---|---|
| Rice | 7,011 |
| Jowar | 1,082 |
| Sugarcane | 1,042 |
| Maize | 725 |
| Bajra | 366 |
| Ragi | 245 |

(Source : Census of India, 1981, Part. 2. Andhra Pradesh, Paper I 1982).

Condition series helps to understand the relative importance of variables relating to a

condition. For example, the above data shows the relative importance of Principal crops in Andhra Pradesh.

## 7.3.2 SERIATION BASED ON FREQUENCY DISTRIBUTION

Statistical data collected and classified is presented in the form of frequency distribution. In a series based on frequency, the observations with similar or closely related values are placed in specific groups or classes. In other words, the distribution shows the frequencies of occurrences of different items of phenomenon. Each group is arranged in order of its magnitude in the series. Erricker has defined it as a classification according to the number possessing the same value of the variable. According to Croxton and Cowden, "Frequency distribution is a statistical table which shows the set of all distinct values of the variable arranged in order of magnitude, either individually or in groups, with their corresponding frequencies side by side". The objectives of constructing a frequency distribution are as follows:

**(a) Analysis of data**

Frequency distribution facilitates analysis of data. The unorganised data which is raw in nature is not helpful for decision making. Hence, the raw data should be classified and arranged in accordance with the nature, objectives and scope of the enquiry. Such a classification and arrangement of data helps in further statistical analysis.

**(b) Estimation of unknown population parameter.**

Statistic is a statistical measure which is computed on the basis of a sample of observations drawn from the population. In a sample survey a part of the population is studied. On the basis of the results obtained from the sample, the corresponding values of population viz., parameters are estimated.

**(c) Computation of various statistical measures.**

Data collected, classified and arranged in a systematic order is more useful for further statistical treatment. From such data, various statistcal measures can be computed to establish relationship among two or more related phenomena. Such relationship can also be subjected to various statistical tests to determine their accuracy and validity.

Seriation on the basis of frequency distribution is of two types, viz., (a) individual series or observations and (b) grouped observations.

**(a) Individual series or observations**

In case of individual series, the values of various items are listed individually one after the other. They are not arranged in groups. Following is an individual series.

**Example - 4**

**WAGES OF 50 WORKERS OF A FACTORY**

80, 90, 70, 60, 64, 65, 78, 93,95, 96, 97, 102, 105, 104, 105, 58, 106, 118, 119, 120, 132, 134, 133, 135, 136, 120, 110, 112, 114, 113, 107, 109, 120, 132, 132, 133, 63, 58, 137, 138, 139, 126, 127, 162, 139, 164, 165, 170, 172, 171, 94.

The above data is not useful for statistical treatment unless it is arranged in a systematic manner. To be more useful, such data should be rearranged either in ascending order or in descending order which is known as 'array'. If the data is arranged in order of lowest value to the highest value, the arrangement is known as ascending order. The arrangement of data in the order of highest value to lowest value is refered to as descending order. The data given in example-4 is arranged in ascending order as below:

**WAGES OF 50 WORKERS OF A FACTORY**

(In ascending order )

58, 59, 60, 63, 64, 65, 70, 78, 80, 90, 93, 94, 95, 96, 97, 102, 104, 105, 105, 106, 107, 109, 110, 112, 113, 114, 118, 119, 120, 120, 120, 126, 127, 132, 132, 133, 133, 134, 135, 136, 137, 138, 139, 139, 162, 164, 165, 170, 171, 172.

The data given in example -4 can also be arranged in descending order as given below:

**WAGES OF 50 WORKERS OF A FACTORY**

( In descending order )

172, 171, 170, 165 164, 162, 139, 138, 137, 136, 135, 134, 133, 133, 132, 132, 127, 126, 120, 120, 120, 119, 118, 114, 113, 112, 110, 109, 107, 106, 105, 105, 104, 102, 97, 96, 95, 94, 93, 90, 80, 78, 70, 65, 64, 63, 60, 59,58.

Though arrangement of data in individual series is quite simple, its utility is limited, as it does not help the statistician to interpret the data. Further, it does not help to condense the mass data. Thus, it does not fulfil the basic principles of classification, such as comparability, condensing, placing the values in relation to each other, etc.

**(b) Grouped observation**

Raw data is not useful for analysis and interpretation unless it is processed. In grouped observations, the total number of items possessing certain number of values of the variable put together are stated as the frequency of those values. Data can be grouped on the basis of their occurrences for the purpose of condensing and simplifying, without losing its essentials. Grouped observations may be arranged in two ways, viz., discrete series and continuous series.

**i) Discrete Series:** In discrete series, various values of a variable are shown in discrete numbers. They are not grouped in terms of range, but there is definite difference between variable of different groups. Their corresponding occurrence known as frequency are also

indicated in the series . According to Boddington, "Discrete variable is one where the variates (individual values) differ from each other by definite amounts". An example of discrete frequency distribution is given below:

**Example - 5**

| No. of children Per family | No. of families |
|---|---|
| 0 | 8 |
| 1 | 10 |
| 2 | 15 |
| 3 | 32 |
| 4 | 7 |
| 5 | 2 |
| 6 | 1 |

Procedure to construct frequency distribution:

The method of constructing frequency distribution can be facilitated through the technique of tally marks or tally bars as explained below:

While classifying the data into discrete or continuous frequency distribution, it should be expressed in terms of tally marks or tally bars. The frequency of each variable is counted by marking a vertical bar (|) i.e., tally mark against each variable. While in the first column the possible values of a variable are placed, in the second column tally mark is put against the variable whenever it occurs. After a particular value has occurred four times, (IIII) the item is marked by a horizontal or a slanted line accross the vertical bars for the fifth occurrence (𝍸) to give a block of five. When the value occurs six times, a vertical bar (I) will be put and for tenth occurrence again a cross mark will be put against the four vertical bars(IIII) to make a group of another 5 (𝍸) and so on. This technique of grouping the frequency into 5 members facilitates easy counting of the occurrences of the values at the end. In the absence of such cross tally bars, we will get continuous tally bars (IIIIIII), and these tally bars may create confusion and may give scope for mistakes.

Now, the total of tally marks of each variable is placed in the third column called frequency. The procedure for the construction of frequency distribution will be clear by considering the following illustration.

**Illustration -1**

From the following data prepare a discrete frequency distribution of Marks obtained by 30 students in an examination.

50, 50, 50, 51, 51, 52, 52, 52, 53, 60, 60,
60, 60, 61, 61, 62, 62, 62, 65, 65, 66 ,66,
66, 66, 66, 67, 67, 68, 68, 68.

**FREQUENCY DISTRIBUTION OF MARKS OBTAINED BY 30 STUDENTS**

| Marks Obtained | Tally Marks | No. of students (frequency) |
|---|---|---|
| 50 | III | 3 |
| 51 | II | 2 |
| 52 | III | 3 |
| 53 | I | 1 |
| 60 | IIII | 4 |
| 61 | II | 2 |
| 62 | III | 3 |
| 65 | II | 2 |
| 66 | IIII | 5 |
| 67 | II | 2 |
| 68 | III | 3 |

Though discrete series is useful to arrange the data in condensed form its utility is limited to a frequency distribution which contains few items only. For a fairly good number of items, this type of seriation is not useful and does not help in condensing the data. Further, the variable whose values are not capable of exact measurement, does not suit the requirements of discrete series.

**ii) Continuous Series:** continuous series is suitable for arranging the data which can be presented in terms of approximations, but not in exact measurements. In this series the values of a variable are stated in terms of range which is known as class interval. According to Boddington, class interval is, "the variable which can take any intermediate value between the smallest and longest value in the distribution". Class intervals in the continuous series continue from the beginning of the frequency distribution to the end without any break. The difference between continous series and discrete series is that, while the former is expressed in terms of class intervals - the upper limit and the lower limit for each class, - the latter is expressed as a list of classified values in discrete numbers. The following is an example of continuous series:

**Example - 6**

FREQUENCY DISTRIBUTION OF MARKS OBTAINED BY 80 STUDENTS IN A CLASS

| Marks Obtained | No.of students (Frequency ) |
|---|---|
| 0-10 | 5 |
| 10-20 | 6 |
| 20-30 | 8 |
| 30-40 | 10 |
| 40-50 | 4 |
| 50-60 | 20 |
| 60-70 | 10 |
| 70-80 | 8 |
| 80-90 | 9 |

**Illustration - 2**

The following observations relate to monthly incomes of 50 persons. Prepare a frequency distribution in ascending order .

Income in Rupees:

211, 230, 225, 200, 206, 230, 231, 300, 305, 209
216, 225, 234, 217, 220, 229, 240, 235, 259, 260
261, 271, 280, 282, 274, 279, 280, 285, 289, 301
304, 289, 290, 291, 309, 310, 315, 314, 320, 324
317, 265, 247, 251, 231, 205, 210, 217, 246, 255

**Solution**

The lowest value is 200 and the highest value is 324. The difference between these two extreme values is 124. If we take a class interval of 20, 7 classes would be formed. The first class will be 200-220.

| Income in Rs. | Tally bars | No.of Persons |
|---|---|---|
| 200-220 | 𝍸 IIII | 9 |
| 220-240 | 𝍸 𝍸 | 10 |
| 240-260 | 𝍸 I | 6 |
| 260-280 | 𝍸 I | 6 |
| 280-300 | 𝍸 III | 8 |
| 300-320 | 𝍸 IIII | 9 |
| 320- 340 | II | 2 |

## 7.4 BASIC PRINCIPLES FOR FORMING FREQUENCY DISTRIBUTION

During the course of statistical enquiry, the data will be processed and classified into different homongeneous groups for meaningful interpretation. After grouping the values on the basis of different characteristics into classes, the question arises as to how to construct the frequency distribution. Beacuse of the growing importance of frequency distribution in the Science of Statistics, they must be prepared and presented systematically, but no clear cut rules are laid down for constructing frequency distribution. However, the statistician can use his discretion, experience, skill and intution to construct frequency distribution. However, the following guidelines may be useful for the construction of frequency distribution.

**a) Number of classes**

The number of classes that should be in a frequency distribution depends upon the total number of observations of the data , magnitude of the items, degree of accuracy desired and the case for further statistical treatment.

Even though these factors are considered to be more important for deciding the number of classes, there are some other factors such as the size of class intervals, method of analysis and interpretation of data. While fewer classes make the data more concentrated, more classes make the data more fragmented. Large number of classes will make the distribution unweildy to handle and quite tedious to work for computations of data at the time of analysis. Further, the classes give only a broad picture of the data. The choice with regard to number of classes depends on the number of observations and size of class intervals. However, a balance should be maintained between the two. Professor H.A. Sturges has formulated a principle by which the number of class intervals can be determined. His formula is,

$$n = 1 + 3.322 \; Log \; N.$$

where,

n=number of classes, N= Number of observations. Sturges forula minimises the number of classes in a frequency distribution. For, example, eventhough there are 25 and 25,000 items in a distribution, the classes will be 6 and 16 respectively. The number of classes can be ascertained through this formula.

| No.of items | Sturge's formula | No.of class intervals |
|---|---|---|
| 25 | n = 1+3.3(1.3979)= | 1+4.61 = 5.61 or 6 |
| 25,000 | n=1+3.3(4.3979) = | 1+14.51 = 15.51 or 16 |

**b) Size of class interval**

The size of the class interval is inversely proportional to the number of classes of a given

distribution. In deciding the size of the class interval, the classes must be arranged in such a way that each and every item of the data is included in the classification. Absolute accuracy can be obtained by having a class for every value represented in the original data. A systematic procedure for approximating a class interval will be of much use. The size of class interval can be decided by taking into consideration the largest value and the smallest value in the observations and the number of classes to be framed.

$$Class\ interval = \frac{L - S}{n}$$

where,

L = Largest value

S = Smallest value

n = Number of classes to be framed.

Besides taking the help of the above formula, as far as possible, one should avoid the use of odd values of class intervals e.g., 3, 7, 11, 26, 39 etc. It is preferable to have class intervals of either 5 or multiples of 5 like 10, 20, 25, 100 etc.

**c) Class limits**

The class limits must be designated in clear terms to avoid confusion. They must be precisely stated so that there will not be any ambiguity. For example, in a frequency distribution the classes are denoted as 10-20, 20-30, 30-40, and so on . In this instance, the class limits are not clear as the upper limit of one class is becoming the lower limit of another class. Further, in the course of classification, it will create some confusion in placing the variable in a particular class. If the value of a variable is 20, there is a scope for its inclusion in two classes, viz., 10-20,or 20-30. Hence, the classes may be denoted as 10-19, 20-29, 30-39 and so on, so that there will not be any confusion with regard to grouping of values of a variable.

**d) Beginning and ending of frequency distribution**

The frequency distribution should begin and end at pertinent points. For example, if the minimum age for employment of adults is fixed at 18 years and that of the retirement age at 55 years, then in a frequency distribution dealing with such data, the classes should not be below 18 years but above 55 years.

## 7.5 METHODS OF FORMING CLASS INTERVALS

The following methods are followed for forming class intervals.

### 7.5.1 EXCLUSIVE METHOD

A frequency distribution of continuous series, where the upper limit of one class is the lower limit of the next class is called 'exclusive method' of class. Exclusive method of classification

of data ensures continuity. The following data shows the classification on the basis of exclusive method.

**Example - 7**

**DISTRIBUTION OF MARKS OF 100 STUDENTS IN AN EXAMINATION**

| Marks Obtained | No. of students (Frequency) |
|---|---|
| 0-10 | 4 |
| 10-20 | 10 |
| 20-30 | 16 |
| 30-40 | 22 |
| 40-50 | 18 |
| 50-60 | 20 |
| 60-70 | 6 |
| 70-80 | 4 |

In the above example, the upper limit of the first class is 10, which is the lower limit of the next class. Similarly the upper limit of the second class is 20 which is the lower limit of the third class. In the first class, there are four persons whose marks are between 0 and 9.99. A student whose marks are 10, would be included in second class, i.e., 10-20 class interval. Though this method is widely used in practice, it is not understood by persons who have no knowledge of statisitcs. There is some confusion with regard to the placing of values like 10, 20, 30 and 40, etc., because they can be shown in two classes. It will be lower limit in one class and upper limit in another class. Hence, whenever this method is used, it is presumed that the upper limit of the class is exclusive, i.e., the item of that value is not kept in that method of class. In order to make the method more understandable, the exclusive method of class limits should be arranged as below. The data given in example -7 is arranged as under.

**DISTRIBUTION OF MARKS OF 100 STUDENTS IN AN EXAMINATION**

| Marks Obtained | No. of students (Frequency ) |
|---|---|
| 0 but under 10 | 4 |
| 10 but under 20 | 10 |
| 20 but under 30 | 16 |
| 30 but under 40 | 22 |
| 40 but under 50 | 18 |
| 50 but under 60 | 20 |
| 60 but under 70 | 6 |
| 70 but under 80 | 4 |

The above method of class intervals avoids ambiguity regarding the classes and the variables that should be placed in the classes. Hence, this type of class intervals should be preferred in practice.

**Illustration - 3**

Given below is the information relating to marks obtained by 40 students in an examination.

42. 18, 63, 38, 65, 68, 50, 17, 48, 25, 78, 76, 62, 34, 39, 41, 30, 28,
19, 52, 32, 23, 49, 54, 64, 73, 34, 58, 67, 43, 11, 14, 19, [illegible], 17, 16,
14, 18, 26, 37.

Arrange the data in a continuous series by exclusive method.

**Solution**

The lowest value is 11 and the highest value 78 . The difference between these two extreme values is 67. If we take a class interval of 10, 7 classes are to be formed. The first class is 10-20.

| Marks | Tally bars | No. of students |
|---|---|---|
| 10-20 | 𝍸 𝍸 | 10 |
| 20-30 | 𝍸 | 5 |
| 30-40 | 𝍸 II | 7 |
| 40-50 | 𝍸 | 5 |
| 50-60 | IIII | 4 |
| 60-70 | 𝍸 I | 6 |
| 70-80 | III | 3 |

## 7.5.1.1. OPENEND CLASSES

Open-end classes are those in which the lower limit of the first class and the upper limit of the last calss are not known exactly. An open-end class is one which includes all items smaller than some specified upper limit or larger than some specified lower limit. The following example shows the open-end classes.

**Example - 8**

**DISTRIBUTION OF WAGES OF 100 FACTORY WORKERS**

| Wages in Rupees | No. of Workers ( Frequency ) |
|---|---|
| Less than 1500 | 10 |
| 1500 - 2000 | 15 |
| 2000 - 2500 | 27 |
| 2500 - 3000 | 25 |
| 3000 - 3500 | 15 |
| More than 3500 | 8 |

In the above example, the first class interval " Less than 1500" does not denote the meaning of "1000-1500" and the last class interval " more than 3500" does not denote the meaning of "3500- 4000". This type of class intervals is known as open-end class intervals. The open-end classes are almost unavoidable as more number of class intervals are required for a large volume of data which will become unwieldy for analysis.

### 7.5.1.2 UNEQUAL CLASS INTERVALS

Sometimes the data classified into various classes, whose size of intervals may not be equal and uniform. This type of class interval is known as "Unequal" class intervals.

**Example -9**

**DISTRIBUTION OF INCOMES OF 100 HOUSEHOLDS IN A VILLAGE**

| Monthly Income in Ruppes | No. of Households (Frequency) |
|---|---|
| 100-200 | 6 |
| 200-400 | 20 |
| 400-700 | 30 |
| 700-800 | 24 |
| 800-1000 | 20 |

The data arranged in the above manner is not useful for statistical application as the class intervals are unequal. To make the data useful for statistical analysis it needs re-arrangement.

**Check your Progress -1**

Explain the term open-end classes and give a numerical example.

### 7.5.2 INCLUSIVE METHOD

In this method, the upper limit of the class interval will be ' included' in the same class interval. Further, ambiguity about the items identical to the limits of the class interval is removed. The following example will help to understand the inclusive method.

**Example -10**

**DISTRIBUTION OF MARKS OF 100 STUDENTS**

**IN AN EXAMINATION**

| marks Obtained | No.of students (Frequency) |
|---|---|
| 10 - 19.9 | 4 |
| 20 - 29.9 | 10 |
| 30 - 39.9 | 16 |
| 40 - 49.9 | 22 |
| 50 - 59.9 | 18 |
| 60 - 69.9 | 20 |
| 70 - 79.9 | 6 |
| 80 - 89.9 | 4 |

In the above example, the class 10-19.9 will include those students whose marks are in between 10-19.9. If marks of a student is exactly 20, he will be included in the next class. This type of class interval avoids confusion and makes the classification clearer.

If the data comprise whole numbers, the inclusive method of classification may be explained as below:

**Example -11**

**DISTRIBUTION OF MARKS OF 100 STUDENTS**

**IN AN EXAMINATION**

| Marks Obtained | No.of students (Frequency) |
|---|---|
| 10 - 19 | 4 |
| 20 - 29 | 10 |
| 30 - 39 | 16 |
| 40 - 49 | 22 |
| 50 - 59 | 18 |
| 60 - 69 | 20 |
| 70 - 79 | 6 |
| 80 - 89 | 4 |



This method facilitates the inclusion of absolute numbers in the class intervals without any ambigiuty. For example, the value 19 can be included in the first class interval and 20 can be included in the second class interval. But, when the values are given in fractions, their inclusion in the class intervals which are arranged in the above form, becomes difficult. For example, a doubt arises about the inclusion of the value 19.25 or 19.65, which may be interpreted to be included either in the first lass interval or in the second class interval. In such a case, the class

intervals are to be arranged in a modified form. In the modified form, in order to have continuity among the classes, the upper and lower limits of the class intervals are to be adjusted. For this purpose, the difference between the lower limit of the second class interval and the upper limit of the first class interval is to be taken. This value must be divided by two, which gives the value of correction factor, as explained by the following formula.

$$\text{Correction Factor} = \frac{\text{Lower limit of 2nd class interval} - \text{Upper limit of 1st class interval}}{2}$$

Thus, applying this formula for the above example, correction factor can be found by substituting the values.

Lower Limit of 2nd class = 20

Upper Limit of 1st class = 19

Substituting the values in the forumula,

$$\text{Correction Factor} = \frac{20-19}{2} = 0.5$$

While arranging the classes, ocrrection factor should be deducted from the lower limit of all the class intervals and be added to the upper limits of all the classes. Thus, the modified classes of the example - 11 are shown below:

**DISTRIBUTION OF MARKS OF 100 STUDENTS IN AN EXAMINATION**

| Marks Obtained | No. of Students (Frequency ) |
|---|---|
| 9.5 - 19.5 | 4 |
| 19.5 - 29.5 | 10 |
| 29.5 - 39.5 | 16 |
| 39.5 - 49.5 | 22 |
| 49.5 - 59.5 | 18 |
| 59.5 - 69.5 | 20 |
| 69.5 - 79.5 | 6 |
| 79.5 - 89.5 | 4 |

**Illustration - 4**

From the following data, prepare a continuous series by inclusive method.

13, 18, 16, 23, 25, 37, 40, 45, 60, 75,

74, 66, 71, 63, 61, 47, 43, 47, 36, 37,

39, 35, 31, 33, 36, 22, 25, 34, 38, 40,

11, 15, 16, 27, 45, 28, 47, 42, 24, 28,

51, 59, 56, 57, 55, 53, 60, 54, 52, 58.

**Solution**

The lowest value is 11 and the highest value is 75. Ther difference between these two extreme values is 64. If we take a class interval of 10, 7 classes would be formed. The first class will be 10 -19 .

| Size of items | Tally bars | No.of items |
|---|---|---|
| 10 - 19 | 𝍸 I | 6 |
| 20 - 29 | 𝍸 III | 8 |
| 30 - 39 | 𝍸 𝍸 | 10 |
| 40 - 49 | 𝍸 IIII | 9 |
| 50 - 59 | 𝍸 IIII | 9 |
| 60 - 69 | 𝍸 | 5 |
| 70 - 79 | III | 3 |

## 7.5.3 CUMULATIVE FREQUENCY DISTRIBUTION

A frequency distribution reveals as to how many number of times a particular value has been repeated. When the frequency distribution is shown by class intervals, the frequency indicates the value falling in the class intervals. It fails to furnish information relating to the total number of observations" Less than" particular value or " More than" a particular value. This type of information can be obtained from the cumulative frequency distribution. We start from one end and the frequencies of classes are added to the frequencies of the next class. The frequencies so obatined are called cumulative frequencies. Thus cumulative frequency distribution can be arranged in two ways, namely, "Less than" and " More than" frequency distribution. They are explained below:

### 7.5.3.1 LESS THAN CUMULATIVE FREQUENCY DISTRIBUTION

In this method, the frequencies are cumulated on downward basis by arranging the values in an ascending order. " Less than" cumulative frequency for any value of the variable is obtained by adding the frequencies which the cumulative total is arranged. The following illustration helps to understand the "Less than" cumulative frequency distribution.

**Illustration - 5**

Find out "Less than' cumulative frequency distribution from the following data:

**DISTRIBUTION OF MARKS OF 65 STUDENTS**

**IN STATISTICS**

| Marks Obtained | No. of students (Frequency) |
|---|---|
| 20 - 25 | 5 |
| 25 - 30 | 10 |
| 30 - 35 | 20 |
| 35 - 40 | 15 |
| 40 - 45 | 10 |
| 45 - 50 | 5 |

**Solution**

**DISTRIBUTION OF MARKS OBTAINED BY 65 STUDENTS**

**IN STATISTICS**

| Marks Obtained | No.of students (Cumulative frequency) |
|---|---|
| Less than 25 | 5 |
| Less than 30 | 15 |
| Less than 35 | 35 |
| Less than 40 | 50 |
| Less than 45 | 60 |
| Less than 50 | 65 |



## 7.5.3.2 MORE THAN CUMULATIVE FREQUENCY DISTRIBUTION

Arranging the ' more than ' Cumulative frequency distribution, is similar to 'less than' frequency distribution except that the data is arranged in descending order. Here, the distribution starts from the highest value of the variable and ends with the lowest value. The following illustration would further explain the 'more than' cumulative frequency distribution.

**Illustration - 6**

Prepare a cumulative frequency distribution by 'more than' method from the following data.

**DISTRIBUTION OF MARKS OF STUDENTS IN STATISTICS**

| Marks Obtained | No.of students (cumulative frequency) |
|---|---|
| 20 - 25 | 5 |
| 25 - 30 | 10 |
| 30 - 35 | 20 |
| 35 - 40 | 15 |
| 40 - 45 | 10 |
| 45 - 50 | 5 |

**Solution**

**DISTRIBUTION OF MARKS OF 65 STUDENTS IN STATISTICS**

| Marks Obtained | No. of student |
|---|---|
| More than 20 | 65 |
| More than 25 | 60 |
| More than 30 | 50 |
| More than 35 | 30 |
| More than 40 | 15 |
| More than 45 | 5 |

## 7.6 SUMMING UP

Seriation is the arrangement of data in a logical order. Series can be made on the basis of general characteristics and frequency distribution. Again, seriation on the basis of general characteristics may be further classified as time series, spatial series and condition series. While time series refers to the arrangement of data in a chronological order, the spatial series refers to the arrangement of data in relation to different geographical locations. Condition series refers to the arrangement of data by some attributes such as income, productivity, marks and weight. Seriation on the basis of frequency distribution refers to the arrangement of data on the basis of frequencies of occurrences of the items of a phenomenon. It may be in individual series or in grouped series. In case of individual series, every item is listed individually one after another. Grouped observaion refers to the arrangement of data according to their frequencies. It may be further arranged in two ways; namely; discrete series and continuous series. Discrete series is an arrangement of items in discrete variable, which differ from each other by a definite number. In continuous series, the values of the variables are stated in terms of class intervals and frequencies are shown against each class. The class intervals may be formulated by exclusive method, inclusive method and cumulative frequency method.

## 7.7 CHECK YOUR PROGRESS: MODEL ANSWERS

The open-end classes are those classes, where the lower limit of the first class and the upper limit of the last class will be missing. Now you should give an example.

## 7.8 MODEL EXAMINATION QUESTIONS

**A. Short Questions.**

1. What do you mean by seriation of data?
2. Explain the term 'time series'.
3. What is discrete series?
4. What is continuous series?
5. Distriguish between individual observation and grouped frequency distribution.
6. Explain the terms 'mid-value', 'class interval'and 'class frequency'.
7. Distinguish between cumulative frequency by 'more than method' and ' less than method'.
8. Define seriation of data. Explain in detail the various types of series.
9. Elucidate the basic principles of forming frequency distribution.
10. What do you mean by 'class interval'? What precautions are to be taken in selecting class intervals.

**EXERCISES**

11. From the following observations, prepare a discrete frequency distribution:

Income in Rs.

126, 110, 112, 126, 130, 134, 135, 114, 118,
123, 127, 126, 128, 132, 131, 130, 135, 110,
126, 111, 110, 134, 125, 135, 126, 114, 116,
128, 118, 119, 120, 129, 140, 145, 146, 141,
145, 111, 146, 140, 141, 135, 115, 116, 135.

12. Tabulate the following data by taking 10 as the class interval.

45, 31, 56, 64, 60, 94, 40, 36,126,135,
140, 65, 75, 80, 83, 46, 48, 35, 66, 72,
78, 83, 87, 90, 91, 95, 63, 54, 56, 63, 58,
115, 80,126,130, 35, 45, 61, 82, 86,110
71, 96, 86, 62,116,110,105,106, 96,109
120,123,129,130,140,123,127,134,108.

13. Following is a record of heights of 70 students in inches. Tabulate the data in a form of frequency distribution.

62.1, 60.4, 60.9, 63.2, 64.2, 65.6, 68.3,
70.1, 72.3, 70.2, 64.5, 66.3, 67.1, 68.4,
69.3, 70.2, 70.5, 70.9, 60.4, 67.3, 68.9,
66.3, 65.4, 60.0, 65.0, 67.0, 68.0, 70.0,
64.9, 69.3, 68.2, 68.0, 63.2, 64.4, 65.3,
68.2, 63.8, 61.4, 66.5, 63.7, 64.4, 67.6,
64.1, 59.6, 64.5, 66.8, 61.1, 65.7, 60.2,
60.0, 61.5, 66.5, 59.9, 60.0, 6[illegible] 63.4,
62.2, 67.0, 68.0, 63.4, 66.2, 65.[illegible] 62.1,
64.0, 62.5, 59.9, 61.8, 64.2, 67.8, 66.6.

14. Following data relate to marks secured by 50 students in statistics. Construct a frequency table taking 10 as class interval.

20, 24, 22, 36, 37, 41, 46, 48, 71,
75, 63, 62, 51, 56, 57, 48, 26, 27,
28, 29, 30, 31, 41, 42, 36, 38, 40,
50, 53, 55, 56, 48, 43, 49, 60, 61,
21, 23, 25, 35, 34, 38, 41, 53, 55,
60, 64, 65, 66, 68.

15. Present the following information in a frequency table with a class interval of 5.

Weights of students in Kgs.

30, 31, 34, 38, 25, 26, 27, 28, 29, 26, 25,
32, 38, 39, 40, 41, 48, 50, 52, 53, 54, 28,
29, 34, 48, 34, 26, 23, 25, 39, 43, 55, 44,
44, 26, 28, 29, 48, 58, 28, 31, 35, 41, 45,
59, 56, 28, 38, 39, 44.



16. You are given the following information. Prepare a frequency distribution by inclusive method.

20, 23, 61, 44, 63, 67, 40, 42, 75, 74,
41, 30, 31, 35, 38, 41, 45, 48, 50, 51,
55, 58, 60, 62, 64, 71, 73, 76, 74, 46,
42, 43, 36, 39, 36, 32, 31, 40, 41, 43,
22, 25, 28, 30, 31, 55, 58, 44, 50, 53.

17. Following are the marks secured by 60 students. Arrange the data in a frequency table by ' Exclusive method'.

09, 42, 18, 83, 64, 45, 76, 81, 83, 79, 11, 84,
34, 56, 67, 31, 30, 68, 00, 48, 64, 12, 36, 81,

12, 24, 61, 34, 41, 32, 43, 63, 74, 45, 56, 63,
71, 64, 75, 79, 56, 47, 60, 58, 61, 59, 54, 63,
58, 37, 38, 28, 39, 40, 38, 27, 61, 80, 49, 28.

18. Arrange the following frequency distribution by ' Less than' method.

| Class | Frequency |
|---|---|
| 0 - 10 | 10 |
| 10 - 20 | 18 |
| 20 - 30 | 25 |
| 30 - 40 | 36 |
| 40 - 50 | 38 |
| 50 - 60 | 43 |
| | 170 |

19. Arrange the following data in a frequency table by ' More than' method.

| Class | Frequency |
|---|---|
| 18 - 26 | 18 |
| 26 - 34 | 26 |
| 34 - 42 | 29 |
| 42 - 50 | 35 |
| 50 - 58 | 37 |
| 58 - 66 | 30 |
| 66 - 74 | 25 |
| | 200 |



20. Present the following data of percentage marks of 50 students in the form of a frequency table with 10 classes of equal width, one class being 40 - 49 .

10, 35, 49, 54, 49, 50, 42, 63, 61, 48,
36, 50, 65, 60, 46, 42, 48, 62, 31, 47,
40, 71, 68, 70, 50, 50, 64, 66, 11, 49,
60, 57, 71, 11, 70, 80, 42, 72, 11, 40,
82, 39, 79, 19, 39, 76, 40, 90, 60, 35,

21. You are given below the wages paid to some workers in a textile mill. Form a frequency distribution with class interval of Rs. 20.

150, 142, 234, 260, 300, 349, 370,
200, 211, 210, 265, 310, 360, 380,
350, 160, 206, 270, 315, 351, 380,
175, 165, 145, 290, 320, 311, 400,
160, 170, 239, 290, 340, 348, 420,
140, 175, 240, 280, 319, 350, 411.

22. From the following data prepare a discrete frequency distribution.

180, 170, 190, 200, 235, 175, 200, 190, 200,
180, 132, 180, 190, 140, 165, 170, 235, 165,
165, 170, 132, 165, 240, 170, 132, 140, 140,
132, 235, 175, 150, 150, 175, 150, 140, 132,
175, 132, 190, 180, 190, 140, 170, 165, 170,
150, 180, 175.

23. From the following data prepare a frequency distribution.

| Mid points | Frequency |
|---|---|
| 7.5 | 15 |
| 12.5 | 20 |
| 17.5 | 40 |
| 22.5 | 111 |
| 27.5 | 72 |
| 32.5 | 87 |
| 37.5 | 35 |
| 42.5 | 12 |
| 47.5 | 6 |

24. Arrange the following marks in a frequency table taking the lowest class interval as 10 - 20.

46, 39, 60, 65, 70, 41, 30, 81, 37, 70, 71, 80, 72, 18, 58, 11, 75, 30, 63, 65, 19,
15, 10, 62, 38, 71, 70, 27, 20, 30, 73, 47, 79, 39, 42, 16, 35, 61, 27, 80, 28, 30,
19, 12, 65, 75, 75, 42, 29, 32, 14, 48, 82, 80, 40, 72, 36, 15, 53, 73, 42, 39, 71,
42, 10, 55, 70, 39, 40, 63, 46, 35, 58, 69, 27, 41, 60, 45, 80, 60, 62, 28, 12, 43,



## 7.9 RECOMMENDED BOOKS


1. Gupta, S.P. : "Statistical Methods", Sultan chand & company, New Delhi.
2. Gupta, B.N. : "Statistics", sahitya Bhavan, Agra.
3. Gupta, S.C. : "Fundamentals of statistics", Himalaya Pub. House, Bombay.
4. Simpson and Kafka : "Basic Statistics ", Oxford and I.B.H. Publishing Company , Calcutta.


## 7.10 GLOSSARY

1. Class Interval : The difference between the upper and lower limit of a class is known as class interval of that class.
2. Class limits : The class limits are the lowest and the highest values that can be included in a particular class.

| | | | |
|---|---|---|---|
| 3. | Continuous series | : | In this series the values of a phenomenon are shown in a range i.e., between two values. |
| 4. | Cumulative Frequency | : | It is the total number of observations less than or more than a given number or class. |
| 5. | Discrete Series | : | The type of series where a definite gap is maintained between one group and another. |
| 6. | Frequency Distribution | : | A frequency distribution is a table in which the data are grouped into classes and the number of cases which fall in each class are recorded. |
| 7. | Individual series | : | A set of ungrouped data. |
| 8. | Mid value | : | The central point of the class interval is called its midvalue. It is found out by adding the upper and lower limits of a class and dividing the sum by 2. |
| 9. | Seriation of data | : | The systematic arrangement of data either on the basis of its general characteristics or on the basis of the methods of its construction, is referred to as seriation of data. |
| 10. | Open-end classes | : | The classes where either the lower limit of the first class or the upper limit of the last class or both will be unspecified. |

# UNIT - 8: TABULATION OF DATA

**Contents**



## 8.0 AIMS AND OBJECTIVES

The aim of this unit is to explain the meaning, significance and the process of tabulation.

After reading this unit you should be able to:

- list out the objectives of tabulation
- identify the elements of tabulation
- explain the rules of tabulation
- classify the methods of tabulation
- analyse the limitations of tabulation

## 8.1 INTRODUCTION

Having classified the data, such data have to be presented in a condensed or simplified form for easy comprehension. Tabulation is one of the devices available to present the data in a reduced form. This enables the investigator to ensure comparison. While tabulating the data one has to comprehend the elements of a table and rules of tabulation. Depending on the purpose he can use various types of tables. While understanding the tables one has to know the limitations of the tabulation also. Let us get into details.

## 8.2 MEANING AND DEFINITION

Data collected through a statistical investigation is classified according to some characteristics. The classified data should be presented to the readers in a concise, clear and definite form. Data presented in a descriptive form may not be clear, and it may require more time to read, understand and interpret the minute details. Tabulation is one of the simplest devices of presenting the data in an orderly and systematic manner.

According to D.Gregory and H.ward "Tabulation is the process of condensing classified data in the form of a table, so that it may be more easily understood and so that any comparisons involved may be more readily made." L.R. Conner observed that, "Tabulation involves the orderly and systematic presentation of numerical data in a form designed to elucidate the problem under consideration."

A close observation of the definitions reveals that tabulation is a process of condensing the data and presenting it in appropriate tables. Though the terms 'Tabulation' and 'Statistical table' are loosely used as synonyms, there is a subtle difference between the two. While tabulation refers to the process of presenting the statistical data, a table is a means of presenting the data.

## 8.3 OBJECTIVES OF TABULATION.

The basic objective of tabulation is to summarize the mass numerical data and to present it to the readers in the simplest possible form. According to A.L Bowley, the function of tabulation in the general scheme of statistical investigation is to arrange in easily accessible form the answer with which the investigation is concerned.

Some of the important objectives of tabulation are explained below:

**i) Simplification of Data**

The objective of tabulation is to present the statistical data in the simplest and most intelligible form. It avoids all unnecessary details and repetitions. Since the data is arranged systematically in rows and columns, the reader can understand the contents of a table without confusion. This saves considerable amount of time and space. Data presented through tables is more effective than that presented in a textual form.

**ii) Comparison of Data**

Tabulation facilitates easy and meaningful comparison as all the related aspects are shown side by side. The relationship among various aspects of the variable will become clear from the tables as they are arranged in a systematic and orderly manner.

**iii) Identification of Data**

Since data is arranged in a table in suitable columns and rows with appropriate title and number, it is easy to identify tabulated data at any future date. It can be used as a source of reference in the interpretation of a problem.

**iv) Facilitates Statistical Processing**

Statistical data which is in raw form is not convenient for statistical processing. Since data is arranged in a systematic and orderly manner, it is convenient to compute statistical measures such as averages, dispersion, skewness, correlation, regression, etc.

**v) It reveals Patterns**

Tabulation reveals the trends and behaviour of the variables which is not possible in the case of descriptive form of data presentation. Well prepared tables clearly reveal the true characteristics of data and highlight their significant features.

**vi) It economises Space**

Since tabulation requires arrangement of data in a systematic and condensed manner, it avoids unnecessary details and repetitions without losing the objectivity of the data. Thus we can achieve economy of space.

**vii) Provides more Information**

Usually tables contain numerical facts relating to different variables. They are more informative than any other kind of data presentation.

## 8.4 ELEMENTS OF A TABLE

It is not possible to specify the number of parts that a table should contain. It depends upon the nature of the data and may vary from case to case. However, the following are considered important parts of a table.

**a. Number**

Every table should be numbered. Though there are no hard and fast rules regarding the place where the number should be written, it is customary to mention the number either at the top of the table, in the centre, above the title or at the bottom of the table on the left hand side. When the data is arranged in many columns, it is also desirable to assign numbers to each column to facilitate easy reference.

**b. Title**

Every table should have a suitable title. The title of the table should describe the contents of the table in a clear, concise and self-explanatory manner. In general a table should state what, where and how the classified data occurred and also specify the period to which the data related. If the title is too lengthy, it is desirable to have a catchy-title above the main title. However, for the sake of brevity and clarity, objectivity should not be sacrified. Titles should not give scope for different types of interpretations. Both the title and the catchy-title should be lettered prominently to attract the attention of the readers.

**c. Caption**

Caption refers to the column headings. They explain what the contents of the columns represent. A table may contain more than one column and each column may also contain sub-

columns. If different columns are expressed in different units, the units should be mentioned within the captions. To distinguish between the caption and the main part of the table, the caption is shown in smaller letters.

**d. Stub**

Stub refers to the rows or row headings. They are shown at the extreme left side of the table. The stubs are usually wider than caption but they should be kept as narrow as possible. This should not be at the cost of precision and clarity.

**e. Body**

The body of a table is the most vital part of a table. It contains the numerical information. Information which is irrelevant must be avoided and should not be kept in the body of the table. Important items may be underlined or expressed in bold letters. The data should be arranged in the body of the table in appropriate captions and stubs. Such an arrangement of the data is generally shown from left to right in the rows and from top to bottom in the columns. Data can be arranged in the body of a table in any of the following manner.

i) Alphabetical order.

ii) Geographical order.

iii) Chronological order.

iv) Conventional order.

v) Progressive order.

vi) Ascending or descending order.

**f. Head note**

The head note of a table is intended to provide a brief explanation of the contents of the table. Usually it is placed below the title of the table and written within brackets. However, it can also be writen on the right hand upper corner of the table. The head note is a supplement to the title and explains certain points relating to the whole data of the table which is not included either in the table or in the stubs and captions. Units of measurement are usually given in head note, such as 'in thousands of rupees' or million tonnes' etc.

**g. Foot note**

Foot notes are intended to provide further classification about the data contained in the stubs, captions and main body of the table. Foot notes are placed below the body of a table. Generally, foot notes are used when:

i) the data is inconsistent

ii) the data contains ambiguity

iii) the data is affected by a special circumstance

iv) the source is acknowledged in the case of secondary data

Foot notes can be indentified by numbering them consecutively with small numbers such as 1, 2, 3, 4 and with samll letters like a, b,c,d. Alternatively foot notes can be indentified with stars. In such cases the frist foot note is given one star*, second two stars **, and so on.

**h. Source note**

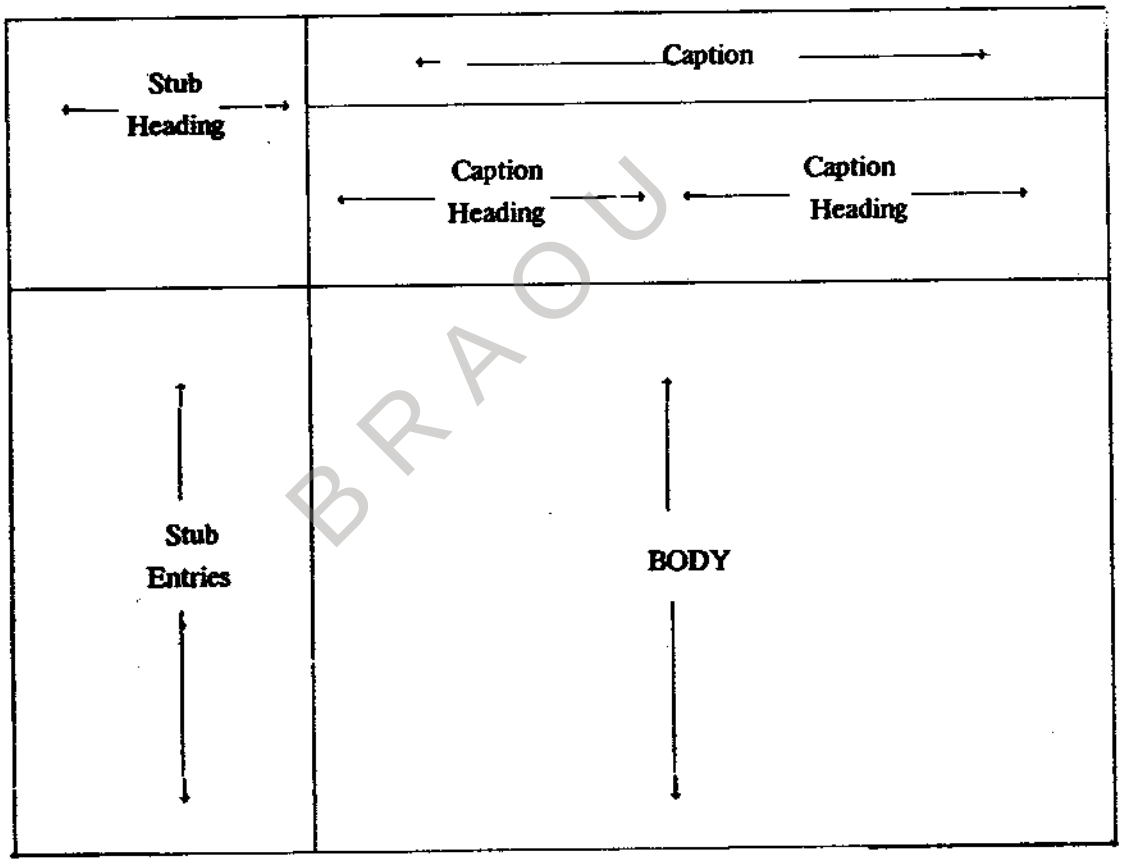
Source note refers to the source of information from where it is obtained. It contains the name of the author, the name of the book/Journal/Office, the date of publication, Volume number, page number etc. Source note is mentioned on the left hand side or right hand side below the table.

**STRUCTURE OF A TABLE**

**NUMBER**

**TITLE**

**HEAD NOTE**






**Foot Note :**

**Source Note :**

**Check your progress - 1**

List out the elements of a table

## 8.5. RULES OF TABULATION

Though it is difficult to prescribe exact rules, the following general rules must be followed in the construction of tables.

(i) Every table should be precise and easy to read. It should be simple to understand and interpret its contents. Data should be arranged in suitable stubs and columns. Ambiguity of data should be avoided as it creates confusion and leads to different interpretations. Hence, every table must be complete, concise, correct, clear and it should serve the purpose for which it is intended.

(ii) Every table should contain a proper title and head note, and they should represent the contents of the table.

(iii) Title, head note, main and sub-headings of the table should be self explanatory.

(iv) Tables should not be overloaded with large data. In case of large data, separate tables should be prepared for each important aspect of the table.

(v) the lay-out of the table should be designed keeping in view the availability of space. The size of columns and rows should be decided before hand in such a way as to incorporate all the details highlighting the chief characteristics and significant relations among various items.

(vi) Every table should be numbered serially in the order of their occurrence.

(vii) Thick lines should be used to indentify clearly between different subdivisions of the tables.

(viii) As far as possible, a table should contain very few main headings. At the same time it should contain as many number of sub-headings as possible. This helps the readers in understanding the important points of the table.

(ix) Data arranged in tables should facilitate comparisons. To get this objective fulfilled, related data should be arranged side by side along with their percentages, ratios or averages.

(x) Lengthy figures should be approximated and shown in the tables to avoid unnecessary minute details of the data. In such cases the method adopted to approximate the figures should be indicated in the foot notes which ensure better comprehension.

(xi) The units of measurement under each heading and sub-heading should be indicated precisely and clearly at their respective places.

(xii) Total of rows should be placed in the extreme right column. On the other hand totals of columns should be placed at the bottom of the table.

(xiii) Items should be arranged either in alphabetical or chronological or geographical order or in the order of their magnitude and significance. This facilitates easy and meaningful comparison between two or more sets of items.

(xiv) Significant figures should be written in bold letters or they should be kept in a 'box' or 'circle' or between 'thick lines'.

(xv) When both the original data and its percentages are given it is better to mention percentages in italics.

(xvi) Data which cannot be arranged in any class or sub-division should be shown under a separate class namely 'miscellaneous class'.

(xvii) Abbreviations should be avoided especially in titles and headings for example 'yr' should not be used for year.

(xviii) Wherever the figures are not available they should be indicated by dash (–) or by the letters N.A. The expanded form of such abbreviations should be clearly indicated in foot notes.

(xix) Expression like etc., viz., should be avoided to the extent possible as the readers may not readily understand their meaning.

(xx) Ditto marks should be avoided. There is the danger of reading ditto marks as 11. Hence repeat the figures or words as many times as they occur. Before the table is finalised the contents, validity and accuracy should be checked and reviewed.

All these guidelines are difficult to follow in any particular case. J.C. Capt has summarised these guidelines as follows:

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

## 8.6 TYPES OF TABLES

Broadly, tables may be classified into two categories:

8.6.1 Simple and Complex Tables.

8.6.2 General purpose and Special purpose Tables.

### 8.6.1 SIMPLE AND COMPLEX TABLES

The distinction between simple and complex table is based on the number of characteristics presented in the tables. While the data in respect of a single characteristics is presented in a simple table, two or more number of characteristics are presented in a complex table.

i) **Simple table:** As the name itself suggests, a simple table is easy to construct and simple to understand. A simple table is also known as one way table. The following is the example of a simple table.

Number of workers in factory 'X' according to income group

| Income (in Rs.) | No.of workers |
|---|---|
| 300-400 | 250 |
| 400-500 | 400 |
| 500-600 | 200 |
| 600-700 | 100 |
| 700-800 | 50 |
| Total | 1000 |

This type of table conveys only a limited message to the readers as more number of details of the characteristics cannot be shown in the table.

**ii) Two-way table:** If the information in rspect of two characteristics is shown in the table, it is called a two-way table. Two-way table can be prepared either by dividing the stubs or the captions into sub-divisions. The following is a specimen of two-way table:

Number of workers in factory 'X' in different income groups according to sex

| Income (in Rs.) | Workers | | Total |
|---|---|---|---|
| | Male | Female | |
| 300-400 | | | |
| 400-500 | | | |
| 500-600 | | | |
| 600-700 | | | |
| 700-800 | | | |
| Total | | | |

Two-way tables convey more infromation than one-way tables. But this type of tables fails to convey the detailed information which is required to understand the total background of the characteristics.

**iii) Higher order tables:** When three or more characteristics are shown simultaneously in a table, it is called a higher order table. In such cases it is necessary to establish an order of precedence among the characteristics to show them in the table. This can be done on the basis of the relative importance of each characteristic that is to be presented. The following is a specimen of a three-way table which presents three characteristics.

Number of workers in factory 'X' according to Income, Age and Sex

| Income (in Rs.) | AGE | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 20-30 | | 30-40 | | 40-50 | | 50-60 | | Total | | Total |
| | M | F | M | F | M | F | M | F | M | F | |
| 300-400 | | | | | | | | | | | |
| 400-500 | | | | | | | | | | | |
| 500-600 | | | | | | | | | | | |
| 600-700 | | | | | | | | | | | |
| 700-800 | | | | | | | | | | | |
| Total | | | | | | | | | | | |

NOTE : 'M' Indicates Male and 'F' Females.

Though higher order tables are more informative than one-way and two-way tables, it is difficult to construct such tables, especially when the characteristics are divided into a large number of sub-divisions.

## 8.6.2 GENERAL PURPOSE AND SPECIAL PURPOSE TABLES

Tables can also be classified as general or special purpose tables based on their utility. General purpose tables are also known as repository tables. They provide information in general and are useful for easy reference. Since general purpose tables contain more information they are highly useful to researchers. Tables published by various governmental agencies are best examples of this type of tables. For example, Statistical abstracts published by government of Andhra Pradesh. However, their use is limited, where specific description of the characteristics is needed. A special purpose table porvides the information relating to specific description of characteristics. It gives the information in a summary form. A special purpose table is also known as derivative table, as it is prepared with the help of figures and results derived from the original or primary data. It reveals the information in terms of ratios, percentages, aggregates or statistical measures like average, dispersion, skewness etc. For example, tables showing the trend values of time series data are also called derivative tables as they are derived form general purpose tables. A special purpose table is usually prepared in such a way that the reader can easily refer to the table, understand the contents, compare and analyse the facts shown in it. Though they describe the specific aspects of a characteristic in detail, they fail to convey complete information to the users.

## 8.7 LIMITATIONS OF TABLES

i) Tables contain only figures. It is not possible to describe the phenomena in detail. As such, they are not easy to read or simple to understand.

ii) The usefulness of tables is restricted to certain people only as they require specialised knowledge to read, understand and interpret the contents. They are not useful to illiterate

people or laymen.

iii) The construction of a table involves mathematical computations. As such, it cannot be prepared by those who lack a mathematical back ground.

iv) If too many details are shown in a single table, it creates confusion and difficulty of comprehension. People may feel fed up with numericals. This is possible especially in the case of complex tables.

(v) The Construction of table requires artistic talent. Tables prepared carelessly and haphazardly may defeat the very purpose of tabulation. Despite these limitations, tabulation is essential for any scientific statistical investigation.

To overcome these limitations, tables must be used as supplements to the textual reports.

## 8.8 SUMMING UP

Tabulation refers to the process of presenting the statistical data in an orderly and systematic manner. A table is a means of presenting the data in a condensed manner. The important objectives of tabulation include simplifcation, comparison, and identification of data. It makes the data suitable for futher statistical treatment in addition to providing more information and economising the space.

A table comprises number, title, captions and body. Some times a head note may be given to give brief explanation of the contents of the table. Foot notes may be used in case any clarifications are needed. Though a number of rules are prescribed for tabulation, the application of common sense and preparation of a table from the point of view of the user are very important.



## 8.9 CHECK YOUR PORGRESS:MODEL ANSWERS

1. The elements of a table are:

a) Number; b) Title; c) Caption;

d) Stub, e) Body f) Head note

g) Foot note h) Source note.

## 8.10 MODEL EXAMINATION QUESTIONS

**A. Short Questions**

1. What do you mean by tabulation of data?
2. State the objectives of tabulation.
3. Distinguish between classification and tabulation.
4. Distinguish between tabulation and statistical table.
5. Distinguish between simple table and complex table.

6. Distinguish between general purpose and a special purpose tables.

7. Explain the terms one-way and higher order tables.

8. Explain the significance of foot notes.

## B. Essay Questions

9. What are the elements of a table? Explain the precautions you would take in tabulating the data.

10. Explain the purpose of tabular presentation of statistical data. Draft a form of tabulation to show the distribution of population according to community by age, literacy and marital status.

11. Draw a blank table showing the distribution of employees in a University accourding to (i) Age; (ii) sex and (iii) Designation.

12. The following information relates to an Engineering College. Prepare a blank table showing the following details.

a) Class : Civil Engg., Mech. Engg., Engg.

b) Sex : Boys, Girls

c) Age : 20-25, 25-30, 30 and above

d) Years : 1980-81, 1981-82, 1982-83, 1983-84.

13. Prepare a specimen table covering the following information which relates to a commercial undertaking.

i) Employees : Male, Female

ii) Age : Below 30, 30-40, 40-50, 50-60

14. Prepare a table to show the distribution of employees of company during the years 1976-77 to 1983-84.

| Years | : 1976-77, | 77-78, | 78-79, | 79-80, | 80-81, | 81-82, | 82-83, | 83-84 |
|---|---|---|---|---|---|---|---|---|
| Employees : | 150 | 220 | 275 | 300 | 365 | 400 | 550 | 700 |

15. According to 1981 population census the populations of Nizamabad, Khammam, Warangal and Hyderabad are: 19,59,352; 24,07, 299; 23, 00, 295; and 22, 60, 702 respectively. The Female population of these districts are 9,90,767; 11,80,356; and 10,83,322 respectively. Prepare a suitable table.

16. In 1982, out of a total 2,000 workers of a factory, 1,500 workers were permanent. The number of women workers was 300 of which 200 were temporary. In 1983, the number of workers incresed to 2,500 of which 1,800 were men. On the otherhand, the number of temporary workers fell down to 150 of which 125 were women. Present the above data in the form of an appropriate table.

temporary workers fell down to 150 of which 125 were women. Present the above data in the form of an appropriate table.

17. The number of students admitted in a University were 7,500: 10,600; 14,300; and 20,500 during the years 1980-81; 1981-82; 1982-83; and 1983-84 respectively. Of the total number, B.A. Students were 4,500; 6,500; 8,000 and 12,000 during the years 1980-81, 1981-82, 1982-83, and 1983-84 respectively. The B.Com. students during the same period were 3,000; 4,100, 6,300, and 8,500 respectively. Present the above information in a suitable table.

## 8.11 RECOMMENDED BOOKS


1. Gupta, S.P : "Statistical Methods", Sultan Chand & Company, New Delhi.
2. Gupta, B.N. : "Statistics", Sahitya Bhavan,Agra.
3. Gupta, S.C. : "Fundamentals of statistics", Himalaya Pub. House, Bombay.
4. Simpson and Kafka : "Basic Statistics", Oxford and I.B.H. publishing company Calcutta.


## 8.12 GLOSSARY

1. Complex Table : A table wherein the data relating to two or more characteristics are presented.
2. Simple Table : A table wherein the data relating to a single characteristic is presented.

# UNIT-9 : DIAGRAMMATIC PRESENTATION OF DATA

**Contents**





## 9.0 . AIMS AND OBJECTIVES

The aim of this unit is to explain the various methods of diagrammatic presentation of data. After going through this unit you should be able to :

- recognise the importance of diagrams
- describe the rules for constructing diagrams
- identify the types of diagrams
- draw the appropriate diagram to the given data

## 9.1 INTRODUCTION

The purpose of classification and tabulation is to summarise and present the data in a definite and systematic manner. However, these techniques of presentation of data do not help the common man to interpret the data in a meaningful form as too many figures create confusion. Therefore statistical results are presented through diagrams and graphs are inportant media for presenting the statistical data and highlighting their basic facts and relationship. While

a graph is constructed on a graph paper a diagram is usually constructed on a plain paper. A graph represents mathematical relationship between two variables whereas a diagram does not represent such relationship . In this unit the diagrammatic presentation of data is explained. The graphic presentation of data will be explained in the next unit.

In the words of A.L. Bowley, "Any list of figures becomes less comprehensible as its length increases. A series of ten number can perhaps be easily grasped, of twenty with an effort, whereas a painted list of figures for one hundred successive years leaves hardly any impression on our mind ; we cannot see the wood for trees". Properly constructed diagrams help the readers in understanding the data as they exhibit the statistical results in a clear and appealing way.

## 9.2 SINGNIFICANCE OF DIAGRAMS

Diagrams are widely used in business, economics, social studies and other fields. They are extremely useful and occupy an important place in statistical methods for the following reasons:

i) **Diagrams are attractive and convincing :** Diagrams are visual form of presenting the data. They are impressive and eye catching. Since they highlight the facts, they convey the message convincingly to those who are fed up with numerical data. Further, they attract the attention of the readers.

ii) **Diagrams make the data simple and intelligible :** Even lengthy and complicated figures can be presented through diagrams in a simple and intelligible form. Thus the complex data can be understood through diagrams without any difficulty.

iii) **Diagrams facilitate comparisons :** The basic objective of diagrammatic presentation of data is to facilitate comparison between two or more sets of data. With the help of diagrams, data relating to different time periods of different geographical regions can be compared easily and effectively.

iv) **Diagrams have a memorising effect :** Unlike figures, diagrams act as effective aids of memory. They create an unforgetable image of the events in the minds of the readers.

v) **Diagrams save time and labour :** To interpret and understand the data presented in a diagram, we do not require much time and effort as in case of data presented in a tabular form.

vi) **Diagrams are more informative :** A diagram provides more information than a table. With the help of diagrams we can easily understand the trends in the data. Though tables show the trends in data, they may not clearly reveal the information as in case of diagrams.

However, diagrammatic presentation of data is not useful for in depth analysis as they reveal only approximate values of the data. They convey limited meaning to the readers. Further, they are not convenient to present all the minute details of the data. It is difficult to construct diagrams, especially when large number of items are to be presented.

**Check your progress- 1**

Out line the importance of diagrams briefly

## 9.3 GENERAL RULES FOR CONSTRUCTING DIAGRAMS

Diagrams must portray the information in a simple, attractive, clear and understandable manner. Hence construction of diagrams require both talent and artistic qualities. The following general rules should be followed while constructing the diagrams :

i) **Title :** Every diagram must be given an appropriate title. The title should be clear, precise and convey the central idea of the data. The title may be given either at the top or below the diagram.

ii) **Proportion between width and height :** A reasonable proportion between the width and height of a diagram should be maintained. If proper proportion between width and height is not maintained, the diagrams would display a shabby picture which would defeat the vey purpose of visual presentation of data. Though there are no hard and fast rules regarding the proportion between width and heigth, a ratio of 1 (short side) to 1.414 (Long side) is usually maintained. This proportion known as 'Root-two' was suggessed by Mr.Lutz in his book entitled "Graphic presentation".

iii) **Scale :** The scale should incorporate all the necessary details of the data. As far as possible, scale should be in even numbers or in the multiples of 5 or 10 e.g., 5,10,15,20,25, or 10,20, 30, 40 etc. Odd numbers such as 1,3,5,7,9,... should be avoided as far as possible. The scale should clearly specify the size of the unit and what it represents. For example, 'Rupees in lakhs', 'metric tonnes in millions', 'number of persons in thousands' etc.

iv) **Foot note :** Necessary foot notes should be given at the bottom of the diagram. This would help the readers in understanding the data without any ambiguity.

v) **Index :** Every diagram should contain an index describing the symbols used, such as dotted lines, colours, shades, etc.

vi) **Simplicity :** Diagrams should be simple and convey the message in clear and definite terms. However, simplicity should not be at the cost of objectivity.

vii) **Neatness and cleanliness :** Diagrams should be neat, clean and impressive and attractive to capture the attention of the readers. For this purpose, different colours should be used.

## 9.4 TYPES OF DIAGRAMMES

A large variety of diagrams are in use. However, the following are the important and widely used diagrams :

9.4.1 One-dimensional diagrams

9.4.2 Two-dimensional diagrams

9.4.3 Three-dimensional diagrams

9.4.4 Pictograms

9.4.5 Cartograms.

## 9.4.1 ONE DIMENSIONAL DIAGRAMS

One dimensional diagrams are also called bar diagrams. They are commonly in use due to their simplicity. In this type of diagrams, data are shown with the help of thick lines or bars. In bar charts, height of the bars indicate the size of the items. Width of the bar is shown only to make the diagrams more impressive. Thus width of the bars have no relation with the measurements. The width of all the bars should be uniform. The gap between one bar and the other should be uniform throughout the diagrams. To present a large number of items, simple lines can also be drawn instead of bars. It is customary to write the respective figures on the top of each bar which helps the readers to known the precise value of each item and enables him to have a comparative picture of various items.

Bar diagrams can be drawn either vertically or horizontally. However, vertical bars are preferred as they give a better look and facilitate easy comparison of various items of the data.

**Types of bar diagrams**

The following are the various types of bar diagrams :

i) Simple bar diagrams

ii) Sub-divided bar diagrams

iii) Multiple bar diagrams

iv) Percentage bar diagrams

v) Deviation bar diagrams.

**i) Simple bar diagrams**

Simple bar diagrams are used to present the data relating to only one variable. In this method, each and evey value of the variable is represented through a bar. The size of each item is identified with the height of the bars. If time series data are given, various values of the data must be shown in their chronological order. On the other hand, if the given data do not have any relation with time, values must be shown either in ascending order or in descending order.

As the name itself indicates, it is easy to construct and simple to under stand. However, its usefulness is restricted to present the data in respect of single variable only.

**Illustration 1.**

Construct a simple bar diagram from the following data relating to Firm 'A'.

| Years | 1977 | 1978 | 1979 | 1980 | 1981 | 1982 | 1983 | 1984 |
|---|---|---|---|---|---|---|---|---|
| Sales: (in'000 rupees) | 20 | 15 | 30 | 40 | 70 | 75 | 90 | 100 |

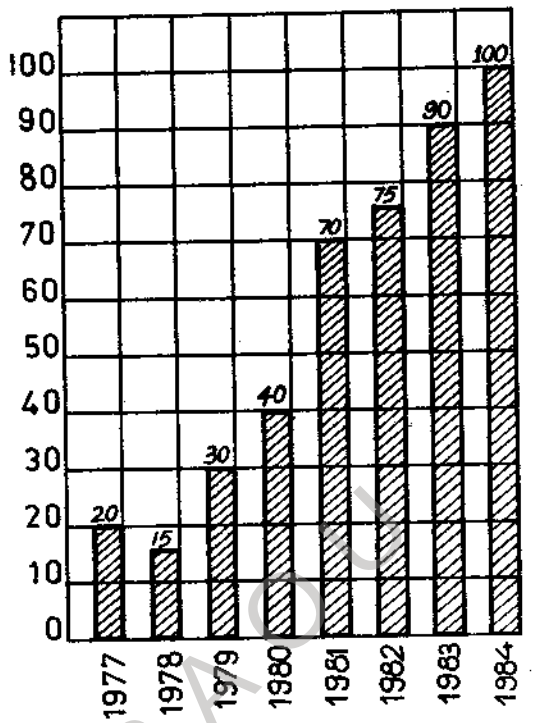



Fig. : 9.1 Sales of Firm 'A' (in '000 rupees)



**ii) Sub-divided bar diagrams**

Sub-divided bar diagrams are also called 'component bar' diagrams. According to this method, the bar is sub-divided into various parts. While bar represents the whole data, its sub-divisions represent different components of the data. To distinguish between various sub-divisions of the bars,different colours,shades, crossings or dottings are used. Usually, the largest component is shown at the base and the remaining components are shown in their order of magnitude.

Sub-divided bar diagrams are useful to present both absolute and relative values of the data. However, it is more convenient to show a set of distribution ratios such as percentage distributions. Though the sub-divided bar diagrams are useful visual aids, it is difficult to construct them, especially when the data are divided into large number of components,

**Illustration-2**

During 1980-81 to 1983-84, the number of students in a college are as follows :

| | Arts | Commerce | Science | Total |
|---|---|---|---|---|
| 1980-81 | 100 | 350 | 300 | 750 |
| 1981-82 | 200 | 300 | 400 | 900 |
| 1982-83 | 250 | 300 | 500 | 1050 |
| 1983-84 | 100 | 400 | 600 | 1100 |

Present the data in a suitable bar diagram.

**Solution :** The above data can best be represented by a sub-divided bar diagram

**iii) Multiple Bar Diagram**

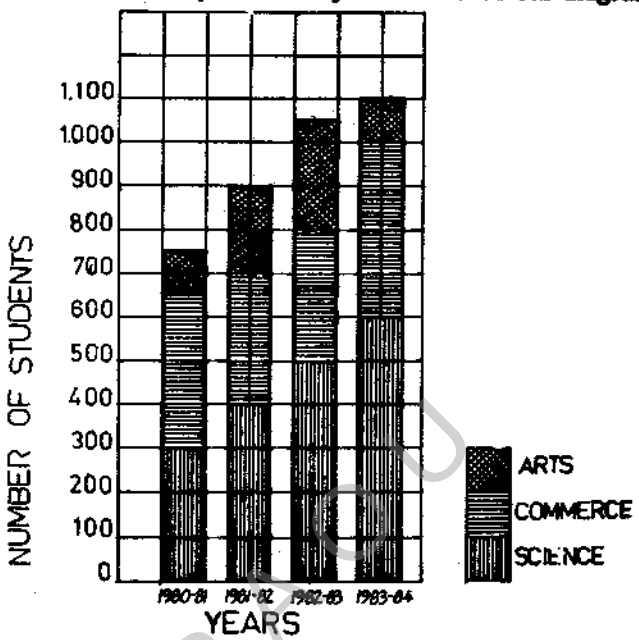



Fig. : 9.2 Number of Students in Science, Commerce and Arts in a College

Multiple Bar diagrams are constructed to present the data where comparison between two or more variables is needed. Construction of a multiple bar diagram resembles the simple bar diagram. The only difference between them is that the simple bar diagram is used to present the data of a single variable, whereas a multiple bar diagram is used to present the data of two or more variables.

**Illustration-3.**

Construct a multple bar diagram from the following data relating to the number of students in a university :

| Year | Arts | Commerce | Science |
|---|---|---|---|
| 1981 | 1000 | 2500 | 4000 |
| 1982 | 2000 | 2500 | 4500 |
| 1983 | 2000 | 4000 | 6000 |

**Solution :**

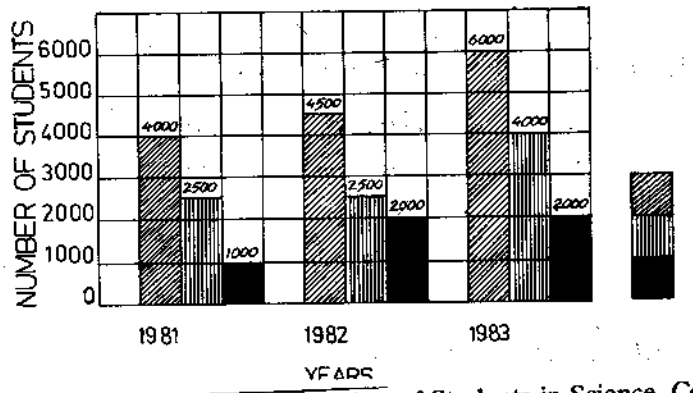



Fig. : 9.3 Number of Students in Science, Commerce and Arts in a University

**iv) Percentage bar diagrams**

Under this method, the total length of the bar is taken as 100. All the components of the variable are converted into percentages and they are cumulated. These cumulative percentages are shown on a diagram, dottings or crossings. Percentage bar diagrams are useful to present the relative changes in the data. But it is not easy to construct this diagram as it involves mathematical calculations.

The elements of cost, sales and profit per unit of article X for the years 1983 and 1984 are given below :

| | 1983 (Rs) | 1984 (Rs) |
|---|---|---|
| Material | 20 | 25 |
| Wages | 10 | 15 |
| Overheads | 5 | 10 |
| Total cost | 35 | 50 |
| Sale price | 40 | 60 |
| Profit | 5 | 10 |

Represent the data by a sub-divided bar drawn on a percentage basis.

**Solution :**

Take the sale price per article X as 100 and express the other figures in percentages. The percentages obtained are given in the following table :

| | 1983 % | 1984 % |
|---|---|---|
| Material | 50 | 41.7 |
| Wages | 25 | 25.0 |

| | | |
|---|---|---|
| Overheads | 12.5 | 16.7 |
| Total cost | 87.5 | 83.4 |
| Sale price | 100.0 | 100.0 |
| Profit | +12.5 | +16.6 |

The percentages would be represented diagrammatically as follows :

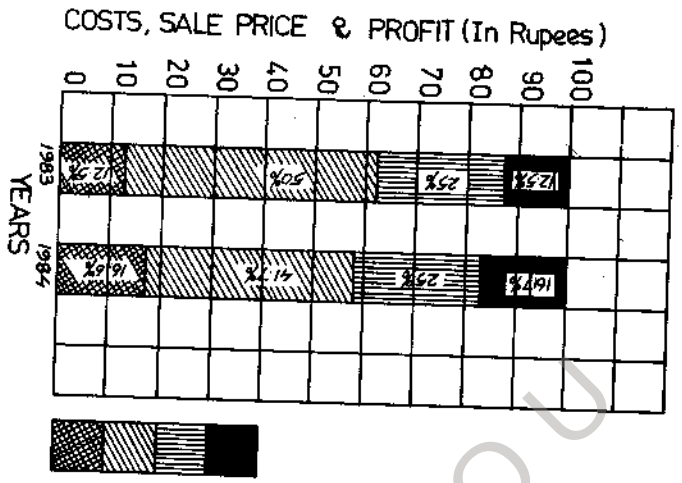



Fig. 9.4 Costs, Sales and Profit per Article.

**v) Deviation-Bar Diagrams**

Deviation-bar diagrams are used to represent the net deviations in the different values of the data such as net profit, net loss, net exports or imports (i.e., balance of trade). While positive values are shown above the base, negative values are shown below the base.

**Illustration-5.**

Present the following data by a suitable diagram showing the difference between sale proceeds and costs.

Sale proceeds and costs of Firm 'A'

| Year | Sale proceeds | Total costs |
|---|---|---|
| 1980 | 50 | 40 |
| 1981 | 100 | 50 |
| 1982 | 200 | 180 |

| | | |
|---|---|---|
| 1983 | 500 | 450 |
| 1984 | 700 | 725 |

**Solution :**

The data can be suitably represented by a deviation bar diagram :

| Year | Sale proceeds | Total cost | Profit |
|---|---|---|---|
| 1980 | 50 | 40 | 10 |
| 1981 | 100 | 50 | 50 |
| 1982 | 200 | 180 | 20 |
| 1983 | 500 | 450 | 50 |
| 1984 | 700 | 725 | (25) |

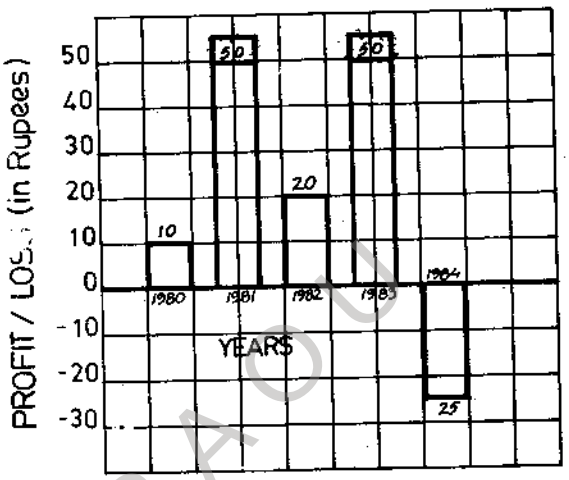



Fig.9.5. The Difference between sale proceeds and costs.

**Broken-Bar Diagrams**

Some times, there may be wide variations in the given values as some of the values may be very small and others may be very large. In such cases, to adjust the space and show all the components in an intelligible manner, the largest bar is broken.

**Illustration-6.**

Represent the following data by a suitable diagram :

| Colleges | No. of students in B.A.,Class |
|---|---|
| M | 50 |
| N | 100 |
| O | 300 |
| P | 25 |

**Solution :**

The number of students of College O is the maximum (12 times that of College 'P'). In order to gain space we have broken the bar for College 'O'. Otherwise the length of this bar would have been 12 times that of the bar of College 'P' and the diagram would have occupied a lot of space and give an ugly look

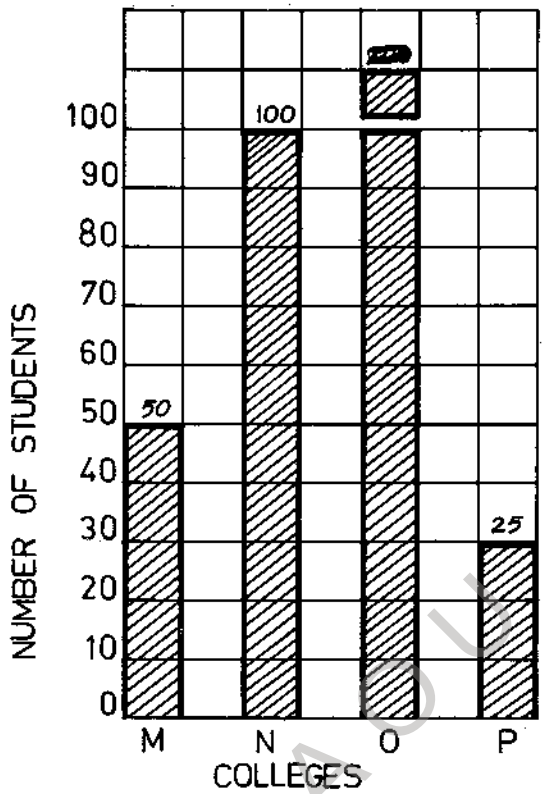



Fig. 9.6. Number of Students (College wise)

## 9.4.2 TWO DEMENSIONAL DIAGRAMS

Two dimensional bar diagrams are also called surface diagrams or area diagrams. While constructing them, both width and height of the diagram are to be taken into account. Thus following are some of the important types of two dimensional-bar diagrams :

i) Rectangles

ii) Squares

iii) Circles.

**i) Rectangles :** Rectangular diagrams are used to represent two or more sets of data. The area under a rectangle is equal to the product of its height and width. While drawing a rectangle, the original values of the data can be used. Alternatively, the percentage values of various components can be used to construct a rectangular diagram. The latter method is particularly useful to compare two or more sets of data.

However, it is very difficult to interpret a rectangular, diagram as it involves the problem of

judging the area. These diagrams are not useful to present the data, especially when the values of the data have a wide range of fluctuations.

**Illustration - 7**

Present the following data by a rectangular diagram :

| | Commodities | |
|---|---|---|
| | X | Y |
| Price per unit of commodity | 20 | 25 |
| Quantity sold | 50 | 60 |
| Cost of raw-materials | 500 | 540 |
| Other costs | 250 | 300 |
| Profit | 250 | 660 |

**Solution :**

Let us calculate the cost raw-materials. Other expenses and profit per unit.

| | Commodity X 50units | | Commodity Y 60units | |
|---|---|---|---|---|
| | Total Rs. | Per unit Rs. | Total Rs. | Per unit Rs. |
| Cost of raw-materials | 500 | 10 | 540 | 9 |
| Other costs | 250 | 5 | 300 | 5 |
| Profit | 250 | 5 | 660 | 11 |

The widths of the rectangles would be in the ratio of 50 : 60 or 5 : 6 or 1 : 1.2.

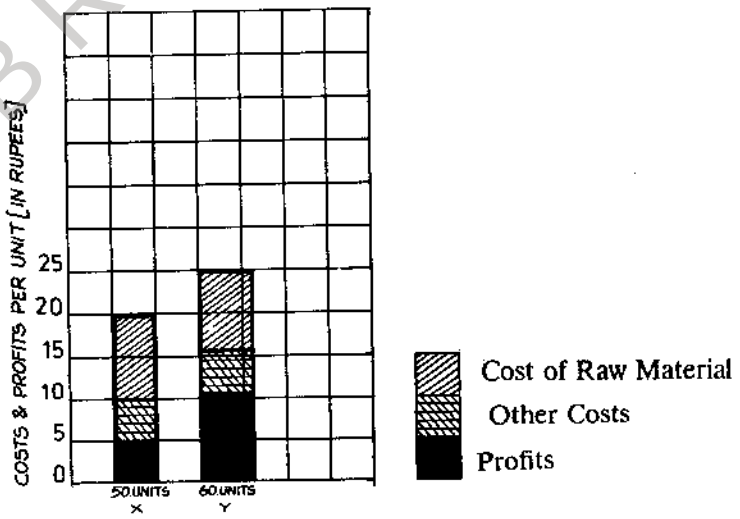



Fig. 9.7. Costs and Profits per unit of Commodities 'X' and 'Y'.

**ii) Squares :** Square diagrams are used to present the data, especially when the values of items have a wide range of variation. While constructing a rectangular diagram,width and height of the diagram are maintained in proportion to the given values. This rule, often results in an ugly shape of the diagram. To over-come this difficulty, square diagrams are preferred to rectangular diagrams. While constructing a square, the square root of the values of items are computed and presented with a suitable scale.

**Illustration-8**

Represent the following data of a firm by a square diagram:

| Year: | 1982 | 1983 | 1984 |
|---|---|---|---|
| Sales: (Rs) | 1600 | 2500 | 10000 |

**Solution :** Since there is a very big gap between the First year sales and fourth year sales; a square diagram may be quite suitable here. The side of a square will be determined by the square root of the value to be represented and the size of the side of various squares shall be proportional to the square roots of the various quantities to be presented.

Calculation for drawing square Diagram

| Years | Sales | Square root | Side of the Square in inches. |
|---|---|---|---|
| 1982 | 1,600 | 40 | 0.4 |
| 1983 | 2,500 | 50 | 0.5 |
| 1984 | 10,000 | 100 | 1.0 |

NOTE: Each figure of the square-root has been divided by 100 and the side of the square obtained.

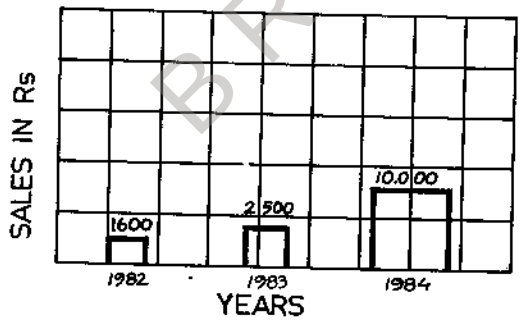



Fig. : 9.8 Square Dagram showing the sales of a firm

**iii) Circles or Pie-Diagrams :** These diagrams are used to present both the total and the component parts or sectors of the data. The area of a circle is proportional to the square of its radius. Seprate circles are drawn by taking the square roots of various figures. The area of each circle depicts the magnitude of different items that are presented.

Usually, a pie-chart is constructed to present the various components of the data. Con-

struction of a pie-chart requires the following steps :

a) Take the total value of the item as equal to 360°.

b) Convert all the components of the data into degrees.

c) Draw a circle. The size of the radius of the circle can be determined on the basis of the space available and other rules of diagrammatic presentation.

d) Measure the points on the circle representing the size of each sector with the help of a protractor.

The values of various components of data are shown in a pie-chart in order of their magnitude. The largest value is shown first at the top and others in sequence in a clock-wise pattern.

When more than one set of items which contain various components are to be presented, separate circles are drawn for each of such items and their components are depicted in the circles. Thus we obtain as many number of pie-charts as the number of items.

Though pie-diagrams are useful devices of presenting the data in an impressive manner, they are less effective when compared to bar diagrams. Accurate reading and interpretation are not possible, especially when the variables are divided into a large number of components. They are also not effective when the difference among various components is negligible.

**Illustration-9**

Represent the data of illustration-8 with the help of circles.

**Solution:**

| Years | 1984 | 1985 | 1986 |
|---|---|---|---|
| Sales(Rs) | 1600 | 2500 | 10000 |
| Dividing each figure by | 509.09 | 795.46 | 3181.82 |
| Square root | 22.56 | 28.20 | 56.41 |
| Side of squares | 0.23 | 0.28 | 0.56 |

**Note :** The side of squares has been obtained by dividing each square root figure by 100.

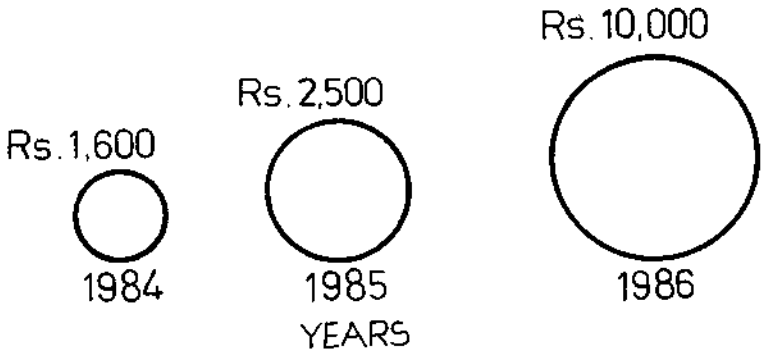



Fig.9.9 Sales of firm During 1984 to 1986.

**Illustration-10**

The following figures relate to sales, costs, and profit.

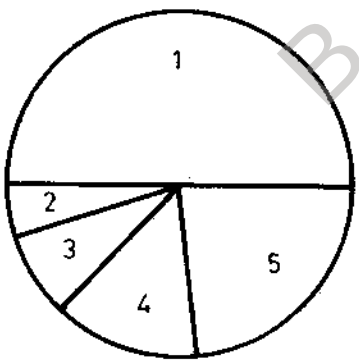
| | |
|---|---|
| Direct Materials | 50% |
| Direct Labour | 20% |
| Factroy overheads | 10% |
| Selling& Distribution Expenses | 5% |
| Profit | 15% |
| Sales | 100 |

Represent the data by a suitable diagram.

**Solution :**

CONVERSION OF COSTS, SALES AND PROFIT INTO DEGREES

| | |
|---|---|
| Direct Materials | = 50 x 3.6 = 180 |
| Direct Labour | = 20 x 3.6 = 72 |
| Factroy overheads | = 10 x 3.6 = 36 |
| Selling and Distribution Expenses | = 5 x 3.6 = 18 |
| Profit | = 15 x 3.6 = 54 |
| Total (sales)... | 360 |




1. Direct materials
2. Selling and distribution expenses
3. Factory overheads
4. Profit
5. Direct labour

Fig.9.10. Pie-diagram showing Sales, costs and Profit of a firm.

**Illustration-11**

The following data relates to the expenditure of Family 'X' and Family 'Y'. Represent this data by an angular (pie) diagram.

| Items of Expenditure | Family'X' | Family'Y' |
|---|---|---|
| Food | 45 | 100 |
| Clothing | 20 | 40 |
| Rent | 20 | 50 |
| Recreation | 10 | 20 |
| Micellaneous | 5 | 30 |

**Solution:** Calculations for constructing the pie-diagram.

| | Family 'X' | | Family 'Y' | |
|---|---|---|---|---|
| Items of Expenditure | Rs. | Degrees | Rs. | Degrees |
| Food | 45 | $\frac{45}{100} \times 360 = 162$ | 100 | $\frac{100}{240} \times 360 = 150$ |
| Clothing | 20 | $\frac{20}{100} \times 360 = 72$ | 40 | $\frac{40}{240} \times 360 = 6\ 0$ |
| Rent | 20 | $\frac{20}{100} \times 360 = 72$ | 50 | $\frac{50}{240} \times 360 = 7\ 5$ |
| Recreation | 10 | $\frac{10}{100} \times 360 = 36$ | 20 | $\frac{20}{240} \times 360 = 3\ 0$ |
| Miscellaneous | 5 | $\frac{5}{100} \times 360 = 18$ | 30 | $\frac{30}{240} \times 360 = 4\ 5$ |
| Total | 100 | 360 | 240 | 360 |
| Square root | 10 | 15.5 | | |
| Radius of Circle | 1.0 | 1.6 | | |

**Note:** The radius of the circles are in the square roots of their totals.

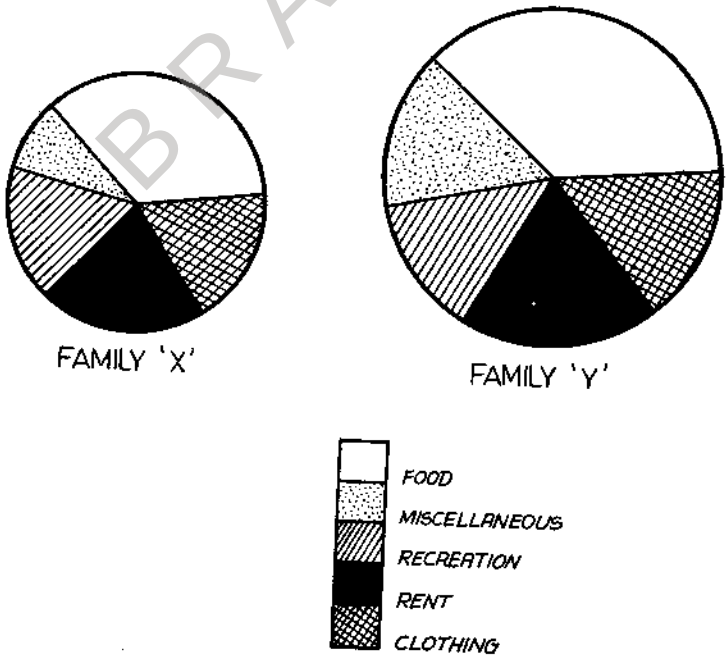



**Fig. 9.11. Pie-diagram showing monthly expenditure of family 'X' and family 'Y'**

## 9.4.3 THREE DIAMENSIONAL DIAGRAMS

Three dimensional diagrams are also called volume diagrams. While constructing these diagrams, length, width and height are taken into account. These diagrams are used to present the data where the difference between the largest value and smallest value is significantly large. Three dimensional diagrams comprise cubes, spheres, prisms, cylinders and blocks. Among them, the construction of a cube is relatively easy. Here construction of cubes alone is explained.

The side of the cube is shown in proportion to the cube-root of the magnitude of data. To ascertain the values of cubes, logarithms can be used. In this case, the logarithms of the figure is divided by 3 and its antilogarithm is taken as the cube-root. A cube is drawn as shown below:

**Illustration-12**

The number of students in University 'K' for two consecutive years was as follows:

| Year | No. of students |
|---|---|
| 1981-82 | 1,600 |
| 1982-83 | 2,500 |

Represent the data by a suitable diagram.

**Solution:** Since the gap between the smallest and the largest value is too large, the appropriate diagram would be the cube diagram. The sides of cubes would be determined as follows (Fig. 9.12) :

| Year | No. of students | Cube Root* | Side of Cubes@ |
|---|---|---|---|
| 1981-82 | 1,600 | 11.69 | 0.58 |
| 1982-83 | 2,500 | 13.57 | 0.69 |

* The cube roots are obtained with the help of logarithms as shown below:

Cube root for 1981-82 = AL 1/3 (Log 1600)

$$= \frac{AL(3.2041)}{3}$$

= AL 1.068

= 11.69

Cube root for 1982-83 = AL 1/3 (Log 2500)

$$= \frac{AL(3.3979)}{3}$$

= AL 1.1326

= 13 . 57

@The side of cubes is obtained by dividing the cube-roots by 20.

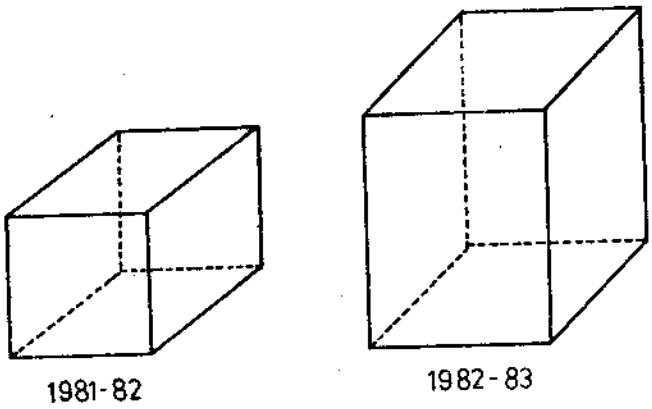



Fig. 9.12. Number of students in University 'K'

Construction of three-dimensional diagrams involves complicated mathematical calculations. It is not only hard to construct but also difficult to read and interprete. In view of these practical difficulties, three dimensional diagrams are not recommended for statistical presentation.

### 9.4.4 PICTOGRAMS

Pictograms attract the attention of the readers quickly as they really depict the kind of data that is presented. According to this method, data is represented with the help of a carefully selected pictorial symbol. Pictograms are highly useful statistical devices of presenting the data, especially to a layman.

The pictorial symbols selected should be self explanatory. For Example, if we wish to represent the data relating to telephones, the symbol should clearly indicate a telephone. As far as possible symbols should be clear, concise and interesting. Changes in numbers should be shown by more or fewer symbols but not by larger or smaller ones.

Pictograms are used to attract the attention of masses in public places like exhibitions, trade fairs, etc, as they create an immage which cannot be forgotten easily.

However, it is not easy to construct a pictogram as it requires an artistic talent. Further, it is not possible to show all the minute details of the data in a pictogram.

**Illustration 13:**

Represent the following data by a pictogram :

| | Production of cars |
|---|---|
| 1982 | 3,000 |
| 1983 | 4,000 |
| 1984 | 6,000 |
| 1985 | 8,000 |
| 1986 | 12,000 |

**Solution:** Let one symbol represent one thousand cars

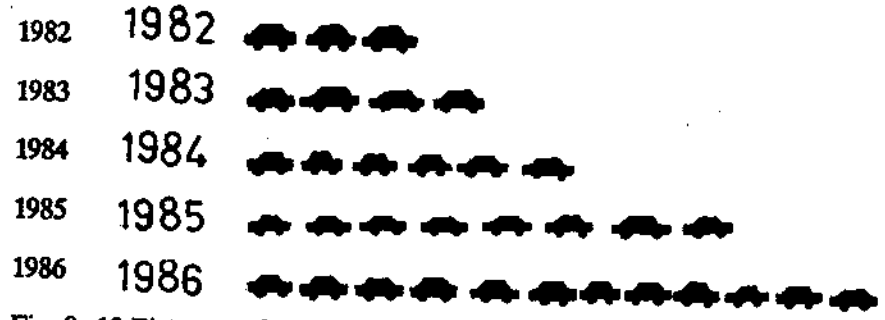



Fig. 9. 13 Pictogram Showing Production of cars during 1982 to 1986.

## 9.4.5 CARTOGRAMS

Cartograms or mapographs are used to present the data on geographical basis. Each geographical unit is clearly shown on the map and their respective values are identified with different colours, points, dots or crosses. Like pictograms, cartograms are also extremely useful visual aids of presenting the statistical data to a layman.

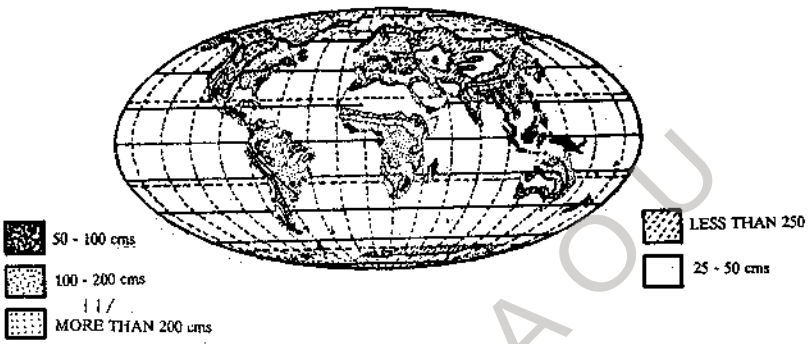



Fig. 9.14 Cartogram showing the rainfall in various Countries

## 9.5 CHOICE OF A SUITABLE DIAGRAM

The selection of a suitable diagram depends upon the nature of the data and also the type of people to whom the message is to be conveyed. If the message is to reach the illiterate people, it is better to use pictograms or cartograms. A pie-chart is useful when a large number of components is to be presented. Cubes are preferred where the difference between the smallest and the largest values to be represented is significantly large.

However, bar diagrams are widely in practice as they are easy to construct and simple to interpret. The selection of a suitable bar-diagram depends upon the following considerations:

i) Simple bar diagrams should be used to represent the magnitude of the values.

ii) Sub-divided bar diagrams should be used to present the relative changes and relative importance of various components of the data.

iii) Percentage bar-diagrams should be used to represent the relative changes and relative importance of various components of the data.

iv) Multiple bar diagrams should be used to represent two or more sets of data.

## 9.6 SUMMING UP

Diagrams and graphs are important media of presenting the statistical data and highlighting their basic facts and relationships. While a graph is constructed on a graph paper, a diagram is generally constructed on a plain paper. Properly constructed diagrams help the readers in understanding the data as they exhibit the statistical results in a clear and appealing way. Diagrams are widely used in business, economics, social studies and other fields. Diagrams are attractive and convincing. They make the data simple and intelligible. They facilitate comparisons and are easy to remember. Different types of diagrams, viz., one-dimensional, two-dimensional, three-dimensional diagrams, pictograms and cartograms are in use. However, the selection of a suitable diagram depends upon the nature of the data and also the type of people to whom the message is to be conveyed.

## 9.7 CHECK YOUR PROGRESS : MODEL ANSWERS

1.

- Attractive and convincing
- Make the mass data simple and intelligible
- Facilitate the comparison between two sets of data
- Have a memorising effect
- Save time and labour
- More informative in than table

You should give a brief description of each of these points.

## 9.8 MODEL EXAMNATION QUESTION

**A. Short questions**

1. State briefly the purpose served by the diagrammatic presentation of data.
2. What is a pictogram?
3. What is a cartogram?
4. What is a Pie-diagram?
5. Explain in brief the multiple Bar-diagram.
6. Distinguish between graphic and diagrammatic presentation of data
7. 'Diagrams help us to see the pattern and shape of any complex situation.' Comment.
8. What points should be taken into account in the construction of diagrams?

**B. Essay Questions**

9. Discuss the merits and limitations of diagrammatic presentation of data.
10. Explain any three important methods used for diagrammatic presentation of data.
11. Represent the following data by a simple bar diagram.

| Year | Production (m. Tonnes) |
|---|---|
| 1980-81 | 20.5 |
| 1981-82 | 35.0 |
| 1982-83 | 50.7 |
| 1983-84 | 72.5 |
| 1984-85 | 90.5 |
| 1985-86 | 120.6 |

12. Represent the following data by a sub-divided bar-diagram:

Price, Cost and Quantity sold of commodities, P and Q.

| | P | Q |
|---|---|---|
| | Rs. | Rs. |
| Price per Unit | 9 | 6 |
| Quantity Sold | 225 | 300 |
| Cost of raw-material | 525 | 450 |
| Other expensses | 90 | 75 |
| Profit | 60 | 75 |

13. Represent the following data by means of percentage sub-divided bar-diagram:

| Particulars | 1986 | 1987 | 1988 |
|---|---|---|---|
| | Rs. | Rs. | Rs. |
| **Cost per Cycle:** | | | |
| Raw material | 300 | 350 | 500 |
| Labour | 100 | 150 | 200 |
| Direct Expenses | 25 | 50 | 50 |
| Office Expenses | 30 | 40 | 50 |
| Total Cost | 455 | 590 | 800 |

14. Construct a sub-divided bar-diagram from the following data:

| Year | No.of students in University 'N' | | | |
|---|---|---|---|---|
| | Arts | Commerce | Science | Total |
| 1983-84 | 15,000 | 22,000 | 7,500 | 44,500 |
| 1984-85 | 16,000 | 23,000 | 10,500 | 49,500 |
| 1985-86 | 20,000 | 30,000 | 12,000 | 62,000 |
| 1986-87 | 22,000 | 35,000 | 10,000 | 67,000 |
| 1987-88 | 30,000 | 35,000 | 10,000 | 75,000 |

15. Given below are the number of students in a University in different faculties, over a period of 5 years. Present them in a multiple bar diagram:

| | No.of Students | | | | |
|---|---|---|---|---|---|
| Year | Arts | Commerce | Science | Law | Total |
| 1983-84 | 1,322 | 1,000 | 750 | 500 | 3,572 |
| 1984-85 | 1,430 | 1,200 | 925 | 600 | 4,155 |
| 1985-86 | 1,600 | 1,500 | 1,000 | 650 | 4,750 |
| 1986-87 | 1,650 | 1,500 | 1,500 | 700 | 5,400 |
| 1987-88 | 2,000 | 1,750 | 1,550 | 800 | 6,100 |

16. Draw a multiple bar-diagram to represent the following data:

| Year | Sales ('000 Rs.) | Gross Profit ('000 Rs.) | Net Profit ('000 Rs.) |
|---|---|---|---|
| 1984 | 300 | 90 | 30 |
| 1985 | 360 | 120 | 45 |
| 1986 | 390 | 135 | 75 |
| 1987 | 450 | 150 | 75 |
| 1988 | 500 | 200 | 90 |

17. Present the following data by a suitable diagram showing the difference between sale proceeds and costs:

Sale Proceeds of a firm

| Year | Sale Proceeds (Rs) | Total Costs (Rs) |
|---|---|---|
| 1983 | 330 | 285 |
| 1984 | 405 | 315 |
| 1985 | 420 | 450 |
| 1986 | 450 | 375 |
| 1987 | 480 | 390 |
| 1988 | 495 | 510 |

18. Represent the following data by a suitable diagram:

| Colleges | No. of students in the B.Com. Class |
|---|---|
| P | 320 |
| Q | 150 |
| R | 600 |
| S | 225 |
| T | 150 |

19. Present the follwoing data by a rectangular diagram :

| | K | L |
|---|---|---|
| | Rs. | Rs. |
| Price per unit of Commondity | 30 | 36 |
| Quantity Sold | 60 | 72 |
| Cost of Raw materials | 300 | 360 |
| Other costs | 180 | 288 |
| Profit | 120 | 210 |

20. Draw a Pie-diagram of the following data relating to areas under different crops :

| Food Crops : | Rice | Wheat | Maize | Jowar | Millets |
|---|---|---|---|---|---|
| Area (in '000 acres) | 16 | 24 | 10 | 8 | 5 |

21. Draw a pie-diagram for the following data:

| Items of Expenditure : | Family A | Family B | Family C |
|---|---|---|---|
| | Rs. | Rs. | Rs. |
| Food | 500 | 600 | 700 |
| Clothing | 200 | 125 | 250 |
| House Rent | 400 | 500 | 450 |
| Education | 100 | 200 | 150 |
| Recreation | 75 | 150 | 200 |
| Miscellaneous | 150 | 250 | 100 |

22. Represent the following data by means of volume diagram :

| | |
|---|---|
| Income of A | Rs. 32,000 |
| „ B | Rs. 14,000 |
| „ C | Rs. 8,000 |
| „ D | Rs. 2,000 |

23. The following data gives the number of lottery tickects (in thousands) sold in the month of April, May and June 1984, Warangal, Hyderabad, Khammam and Nalagonda.

| | Cities | | | |
|---|---|---|---|---|
| | Warangal | Hyderabad | Khammam | Nalagonda |
| April | 225 | 250 | 460 | 150 |
| May | 280 | 308 | 520 | 175 |
| June | 242 | 230 | 330 | 225 |

24. Three years' results of B.Com. students are given in the following table. Represent this by a bar-diagram :

| Year | No. of students | | | |
|---|---|---|---|---|
| | Frist Division | Second Division | Third Division | Failure |
| 1985-86 | 100 | 300 | 500 | 300 |
| 1986-87 | 120 | 400 | 600 | 280 |
| 1987-88 | 100 | 500 | 700 | 300 |

25. The following table gives the details of the cost of construction in town 'M'-

| | Rs. | | Rs. |
|---|---|---|---|
| Land | 9,000 | Cement | 2,500 |
| Labour | 5,000 | Stone | 1,500 |
| Bricks | 4,000 | Sand | 600 |
| Iron | 3,600 | Other- | |
| Timber | 3,000 | things | 2,600 |

Represent the above figures, by a suitable diagram.

## 9.9 RECOMMENDED BOOKS


1. Gupta, S.P. : "Statistical Methods" Sulthan chand & Company, New Delhi.
2. Gupta, B.N. : "Statistics", Sahitya Bhavan, Agra.
3. Gupta, S.C. : " Fundamentals of statistics", Himalay publishing House, Bombay.
4. Simpson and Kafka : "Basic Statisstics", oxford and IBH. publishing Company, Calcutta.


## 9.10 GLOSSARY

1. Bar Diagram : It refers to the presentation of data in rectangulars.
2. Cartogram : It is used to present the data on a geographical basis.
3. Diagram : a diagram is a visual form for presentation of Statistical data.
4. Graph : A graph is a visual form of presentation of statistical data on a graph paper.
5. Pictograms : It refers to the presentation of data through a carefully selected pictorial symbol.
6. Pie diagram : It consists of a circle or a pie divided into a number of sectors representing different components of the variable.

# UNIT-10 GRAPHIC PRESENTATION OF DATA

**Contents**





## 10.0 AIMS AND OBJECTIVES

The aim of this unit is to describe the various methods of graphic presentation of statistical data. After going through this unit, you should be able to :

- distinguish between diagrams and graphs
- describe the utility of graphic presentation
- list the techniques of constructing the graphs
- explain the rules for graphic presentation
- identify the various time series graphs
- identify the various graphs of frequency distribution
- draw an appropriate graph to the given data

## 10.1 INTRODUCTION

Graphs are more effective (in attracting the attention of people towards data), than any other methods of preseting statistical data. The graphic method of presentation of data implies the presentation of various economic and business variables by applying different types of geometrical

devices. The presentation of statistical series by two or more dimensions on a graph is known as graphic presentation. Graphs are useful in studying cause and effect relationship or in studying the extent of change in one variable if the other variable changes.

## 10.2 DISTINCTION BETWEEN DIAGRAMS AND GRAPHS

There are no set rules to distinguish between diagrammatic and graphic presentations, but the following may be mentioned as the differences between them.

i) Generally to construct a graph, a graph paper is used and it is helpful to study the mathematical relationship between two or more variables, whereas diagrams can be constructed on plain paper and they are not useful to study the mathematical relationship between variables.

ii) Graphs are clearer and more helpful in research studies for the analysis and study of slopes of curves, rate of change, and estimation of future values, whereas diagrams attract the attention of the readers on account of their colour and shape and are better suited for publicity and propaganda.

iii) Graphs are used to present data pertaining to time series and frequencies; on the other hand diagrams can be used to present categorical and geographical data.

iv) Construction of graphs is easier than that of diagrams.

## 10.3 UTILITY OF GRAPHIC PRESENTATION

The following are the utilities of graphic presentation of data:

i) Graphic presentation enables us to make the complex and mass data simple by way of drawing a chart or curve on the graph. Such types of presentation enable us to understand the pattern of the distribution and fluctuations of the data very easily.

ii) On account of their simplicity, the time and labour involved for drawing the graphs is minimum. To draw a graph, it is not compulsory that one should have mathematical knowledge. An ordinary man can also draw a graph relating to a variable and can try to analyse and examine the pattern and draw conclusions.

iii) The comparative analysis realting to two or more variables can be easily made by drawing them on a graph paper simultaneously in the form of curves. By close observation of these curves we can identify the significant fluctuations and the factors that influence such variations.

iv) The curves drawn on graphs can be used to estimate the missing values and future values by appling techniques of interpolation and extrapolation. Further, graphs can also be used to determine the values of median, mode, quartiles, deciles etc.

**Check your progress-1**

Distinguish between diagrammatic and graphic presentation of data.

## 10.4 TECHNIQUES OF CONSTRUCION OF GRAPHS

Generally graphs are drawn on a special type of paper which is known as graph paper. Graph paper contains fine net work of horizontal and vertical lines throughout its width and length. The thick lines of a graph paper measure a centimetre or an inch and the thin lines of it measure the small parts of a centimetre or an inch. The area of the entire graph is divided into four parts by drawing two simple lines at right angles to cut each other and intersecting at a point 'O'. The point of intersection i,e., 'O' is said to be the point of origin or the 'Zero point'. The four parts of the graph are known as 'quadrant'; 'XOX' is the axis of 'X' or the 'abscissa' and 'YOY' is the axis of 'Y' or the ordinate. In a graph both positive and negative values can be shown. In quadrant-I, both the values of X and Y are positive. Whereas in quadrant-II the value of Y is positive and that of X is negative. In quadrant-III, both the values of X as well Y are negative, whereas in quadrant-IV, the values of X is positive and that of Y is negative.

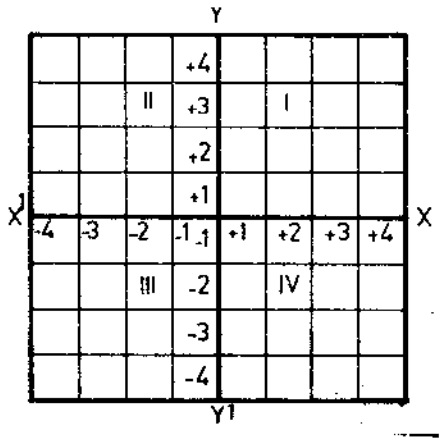
Quadrant-I
X-Positive
Y-Positive

Quadrant-II
X-Negative
Y-Positive

Quadrant-III
X-Negative
Y-Negative

Quadrant-IV
X-Positive
Y-Negative

Fig.10.1 Divisions of a graph



## 10.5 GENERAL RULES FOR GRAPHIC PRESENTATION

While presenting statistical data on a graph paper, the following rules are to be kept in mind.

i) **Title:** Every graph should be given a clear, comprehensive, unambiguous title that is self-explanatory. The title should indicate the subject matter and the contents of the data depicted in the graph.

ii) **Structural frame work:** While drawing a graph, the portion of axes should be selected in such a way that the graph drawn gives an attractive and proportionate get up. While drawing a graph, an independent value should always be accompanied by its corresponding dependent vaiable. It is general practice to draw the independent variable on the 'X' axis and the dependent variable on the 'Y' axis. For instance, if we wish to plot 'rainfall' and 'yield' on graph paper, 'rainfall' is the independent variable and is shown on 'X' axis, whereas the 'yield' is a dependent variable and is therefore shown on 'Y' axis.

iii) **Scale:** The selection of a scale of the graph is made in such a way that the entire values can be accommodated in the available space. In this connection, the words of A.L. Bowley are worth considering. According to him"It is difficult to lay down rules for the proper choice of the scale by which the figures should be plotted out. It is only the ratio between the horizontal and vertical scales that needs to be considered. The figures must be sufficiently small for the whole of it to be visible at one glance; if the figure is complicated, related to long series of years and varying numbers, minute accuracy must be sacrified to this consideration. Supposing the horizontal scale is decided, the vertical scale must be chosen so that the part of the line which shows the greatest rate of increase is well inclined to the vertical which can be managed by making the scale sufficiently small: and on the other hand, all important fluctuations must be clearly visible for which the scale may need to be decreased. Any scale which satisfies both these conditions will fulfill its purpose".

iv) **False Base Line:** The basic principle of graph is that the vertical scale must start with zero. We cannot depict the given variable effectively, when the fluctuations are maximum and the starting values are very distant from zero i.e., the point of origin. In order to magnify the scale on the vertical scale and for effective depiction the 'false base line' is used. In a false base lilne, the vertical scale is broken into two parts and the values of the dependent variable from zero to lowest value are omitted by drawing two zig-zag horizontal above the base line i.e., X-axis. For instance, in drawing a graph in respect of sales for five years, if the values given are Rs.15,000, Rs.17,000, Rs.20,000, Rs.23,000 and Rs.25,000, they can be effectively shown on a graph paper by using a false base line by omitting the values from 'O' to 15,000, on the vertical scale. This can be depicted as below:

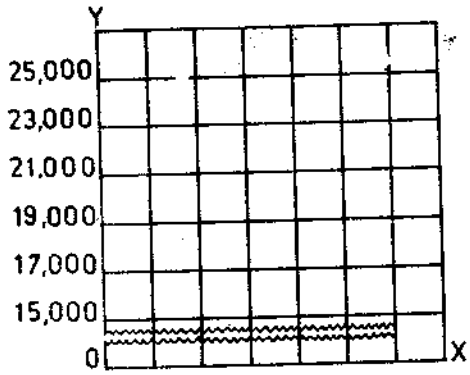



**Fig:10.2 False base line**

**Objectives of the false base line**

The following are the main objectives of the false base line:

i) The technique of the "false base line" is extensively used to magnify the minor fluctuations of time series data.

ii) The space in the graph paper can be economically used.

iii) The graph prepared by using a false base line can provide a better visual communication.

v) **Line designs:** When more than one variable is drawn on the same graph paper for the purpose of making a comparative study between them, different lines such as dotted lines, broken lines, dash - dot lines, thin or thick lines etc., are used to distinguish one line from the other to have a clear understanding about the curves. Further, an index is also prepared to identify them.

vi) **Ratio or Logarithmic scale:** Natural scale with absolute values is not used to study the relative changes in the magnitudes of the variables. In such cases, relative values such as percentages, logarithmic values etc., are used to make comparative studies and to draw valid inferences. The ratio scale is presented on "Y" axis i.e., vertical axis.

vii) Index and source of data: In every graph, an index is given to show the scale of 'X' and 'Y' axes as well as the meaning of different lines used. The source of the data i.e., where from the data is obtained is mentioned at the bottom of the graph.

## 10.6 TYPES OF GRAPHS

Generally, in practice, several types of graphs are used to draw various kinds of economic and business variables. Broadly, graphs are classified into two main categories, namely, Graphs of Time Series or Historigrams and Graphs of Frequency Distribution.

## 10.6.1 GRAPHS OF TIME SERIES OR HISTORIGRAMS

A time series is a sequence of values corresponding to successive points, or periods of time. When data such as sales, production. employment, bank deposits is arrranged chronologically, it constitutes economic time series. The time series data is presented geometrically by means of time series graph, also known as 'Historigram'. In this case the independent variable 'time' is taken on the 'X' axis and the dependent variable i.e., the values of the variable under study on the 'Y' axis. Different types of charts are employed in plotting the time series data. The following are the important graphs of time series data:

A. Graph of one variable or Horizontal line graph

B. Graph of two or more variables

C. Graph of two or more variables measured in different units or Graphs of two scales.

D. Range chart

E. Band graph

F. Semi-Logarithmic line graphs or Ratio charts

## A. GRAPH OF ONE VARIABLE OR HORIZONTAL LINE GRAPH

The horizontal line graphs are used to depict one variable graphically. In this Graph the time units are shown along 'X' axis and the dependent variables such as sales, production, consumption etc. shown on 'Y'axis. The suitable points are plotted on the graph by taking time unit and its corresponding value of variable. By joining all these points by a straight line we will get a horizontal curve.



**Illustration-I**

Draw a graph for the data given below.

| Year | 1977 | 1978 | 1979 | 1980 | 1981 | 1982 | 1983 |
|---|---|---|---|---|---|---|---|
| Sales in (000 of RS) | 120 | 135 | 130 | 138 | 128 | 120 | 150 |

**Solution :**

Taking the scale along X asix as 1 cm = 1 year and along Y-axis as 1 cm = 5,000 Rupees, the required graph can be drawn as below :

**Procedure**

In order to draw the band graph the following procedure is adopted:

i) Show years on the X axis and variables on the Y axis

ii) Draw a curve for the first component of the given data

iii) Draw another curve for the second component of the data over the first curve by calculating the cumulative totals for the first two components.

iv) Draw another curve for the third component over the second curve, on the basis of cumulative frequencies of the first three components. This process is continued till the curves are drawn for all the components.

v) The space between different curves can be shaded by different lines or colours.

**Illustration - V**

The data given in the following table relates to imports in India from various countries Present the data by using the band graph.

Imports (Rs. in crores)

| | 1978 | 1979 | 1980 | 1981 | 1982 | 1983 |
|---|---|---|---|---|---|---|
| Canada | 30 | 20 | 25 | 20 | 35 | 40 |
| U.S.A | 35 | 25 | 20 | 30 | 40 | 35 |
| U.K | 30 | 20 | 30 | 40 | 45 | 30 |
| U.S.S.R | 25 | 30 | 35 | 45 | 50 | 45 |
| Total | 120 | 95 | 110 | 135 | 170 | 150 |



**Solution:**

Drawing of Band graph.

On X-axis 1 cm = 1 year and on Y-axis 1 cm = Rs.5 crores are taken to draw the imports in India.

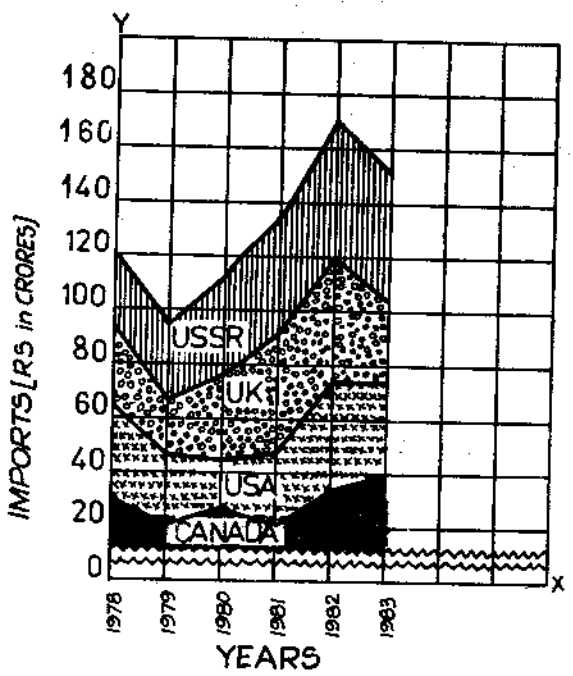



Fig. : 10.7 Imports of India from Different Countries

## F. SEMI LOGARITHMIC LINE GRAPHS OR RATIO CHARTS

The different types of graphs explained so far are drawn on natural scale or arithmetic scale. Such graphs indicate the absolute change in the values of variables which are not useful for comparison. To overcome this difficulty, ratio scale is used to study relative changes in the values. The ratio scale facilitates the comparison of rate of change in different variables on the same graph and helps to study the constant rate of increase or decrease in the variables.

The natural scale and ratio scales are given below

**Natural scale**

| |
|---|
| 6 |
| 5 |
| 4 |
| 3 |
| 2 |
| 1 |
| 0 |

**Ratio scale**

| | | |
|---|---|---|
| 64 | 640 | 6400 |
| 32 | 320 | 3200 |
| 16 | 160 | 1600 |
| 8 | 80 | 800 |
| 4 | 40 | 400 |
| 2 | 20 | 200 |
| 1 | 10 | 100 |

Through the ratio scale is a very significant technique of studying different variables in relative terms, they are seldom used on account of the following difficulties:

i) They are very difficult to understand.

**The formula of frequency density**

$$= \frac{\text{Frequency of the class}}{\text{Magnitude of the class (Class interval)}}$$

**Illustration - VIII**

Draw a histogram by using the following data:

| Wages (in Rs.) | 10-20 | 20-30 | 30-40 | 40-60 | 60-90 |
|---|---|---|---|---|---|
| No.of workers | 5 | 20 | 30 | 40 | 45 |

**Solution:** Construction of Histogram

The histogram is drawn by taking the scale 1 cm = 10 class intervals on X-axis and on Y-axis 1 cm = 5 workers.

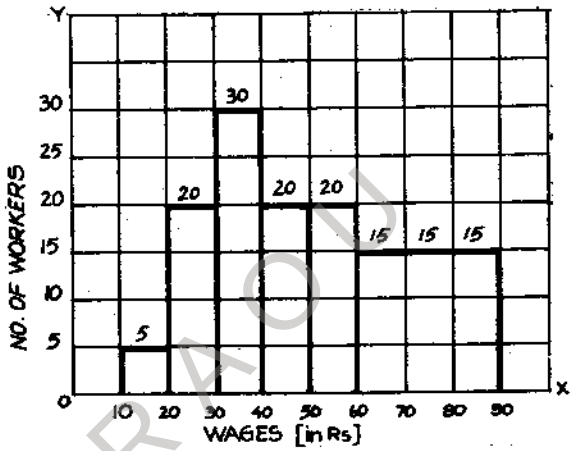



Fig. 10.10. Histogram

## B. FREQUENCY POLYGON

Frequency polygon is another way of presenting a frequency distribution graphically. A frequency polygon has more than four sides and is very useful to compare two or more frequency distributions on the same graph paper. The frequency polygon is an improved technique over the histogram, because it provides a continuous curve showing gradual decline or improvement in the frequency distribution.

**Methods of constructing frequency polygon**

Frequency polygons can be constructed either from histograms or from mid points.

**a) From histogram**

In order to draw frequency polygon from histogram the procedure mentioned hereunder is followed:

i) Draw a histogram for the frequency distribution.

ii) Identify the mid-points on the upper horizontal side of rectangles.

iii) Join the mid-poits by a line. The resulting curve is the frequency polygon. In order to estimate the area of the frequency polygon, the ends of the polygon are to be joined with the base line by assuming that both the frequencies of end-classes are zero.

**Illustration-IX**

The following frequency distribution pertains to profits of companies in India. Draw a histogram for the data and show the frequency polygon for it.

| Profits | 0-25 | 25-50 | 50-75 | 75-100 | 100-125 | 125-150 | 150-175 |
|---|---|---|---|---|---|---|---|
| No.of Companies | 10 | 20 | 30 | 18 | 12 | 10 | 8 |

**Solution:**

Construction of Histogram and frequency polygon.

The histogram and frequency polygon are drawn by taking scale 1 cm = Rs.2,500 on X-axis, and 1 cm = 5 companies on Y-axis.

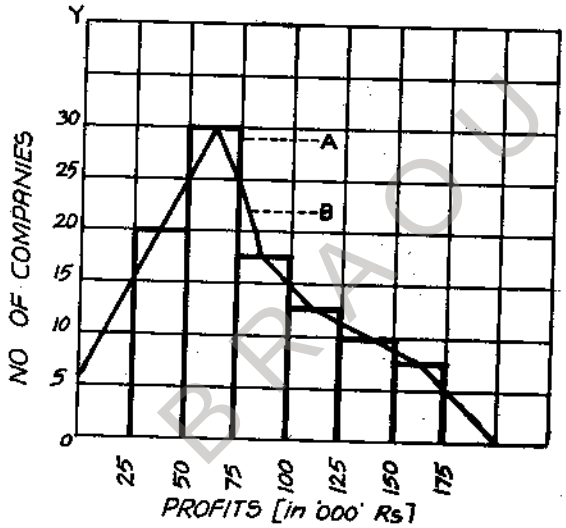



Fig. 10.11. Profits of Companies in India

**b) From Mid-points**

To construct a frequency polygon from mid-points of the classes the following procedure is followed.

i) Obtain mid-points for all the classes.

ii) Plot the mid-points and their respective frequencies on the graph paper.

iii) Join these points by a straight line. The resulting curve is the frequency polygon.

**Illustration-X**

Draw a frequency polygon with the help of the following data:

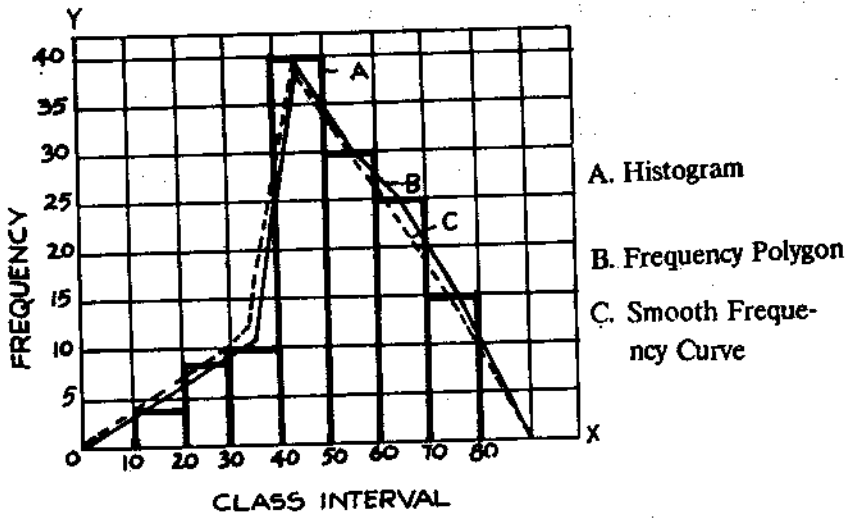



Fig. 10.13 Smooth frequency curve

## D. OGIVES OR CUMULATIVE FREQUENCY CURVES

A graph of a cumulative frequency distribution is referred as an 'ogive' curve (pronounced as 'Ojive'). An ogive curve facilitates the comparison of two or more variables. It can be used to determine the values of median, quartiles and deciles. The cumulative frequency curves enable us to answer the questions such as how many workers of a factory earn less than Rs.300 per month or how many workers earn more than Rs.400 per month?

**Construction of Ogive curves**

There are two methods by which ogive curves can be constructed. They are: a) Less Than Method and b) More Than Method.

**a) Less Than Method:** In Less Than Method, the frequencies are cumulated from the smallest value of the class to the greatest value of the class. Such cumulative frequencies are plotted on a graph paper which will give a rising curve.

**b) More Than Method:** In More Than Method, the frequencies are cumulated by using lower class limits. The cumulative frequencies so obtained are plotted on graph paper which gives a falling or declining curve.

**Illustration-XII**

Draw ogives (both, less than and more than types) from the following distribution of wages of 500 workers.

| Wages (in Rs.) | 50-60 | 60-70 | 70-80 | 80-90 | 90-100 | 100-110 | 110-120 | 120-130 |
|---|---|---|---|---|---|---|---|---|
| No.of workers | 20 | 60 | 100 | 150 | 75 | 50 | 25 | 20 |

**Find out:**

a) Number of workers earning wages less than Rs.95.

b) Number of workers earning wages more than Rs.75.

c) Number of workers earning wages between Rs.60 and Rs.100.

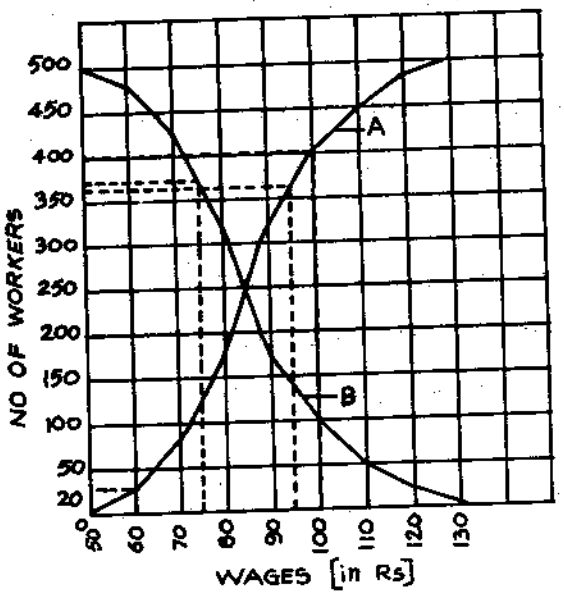
**Solution:**

**Construction of 'Ogive' curves**

Cumulative frequency

| Wages in Rs. | No.of workers | Wages less than | No.of workers | Wages more than | No.of workers |
|---|---|---|---|---|---|
| 50-60 | 20 | 60 | 20 | 50 | 500 |
| 60-70 | 60 | 70 | 80 | 60 | 480 |
| 70-80 | 100 | 80 | 180 | 70 | 420 |
| 80-90 | 150 | 90 | 330 | 80 | 323 |
| 90-100 | 75 | 100 | 405 | 90 | 170 |
| 100-110 | 50 | 110 | 455 | 100 | 95 |
| 110-120 | 25 | 120 | 480 | 110 | 45 |
| 120-130 | 20 | 130 | 500 | 120 | 20 |
| | | | | 130 | 0 |

The ogive curves are drawn by taking the scale 1 cm = Rs.10 on X-axis and on Y-axis 1 cm = 50 workers.




A. Less than Ogive

B.More than Ogive

Fig. 10.14 Less than and More than ogive curves

i) Number of workers earning wages less than Rs.95 are 365.

ii) Number of workers earning wages more than Rs.75 are 130.

iii) Number of workers earning wages between Rs.60 and Rs.100 are 380.

## 10.7 SUMMING UP

Graphic methods facilitate the presentation of quantitative data in a simple, clear and effective manner. Graphic method of presentation of data implies the presentation of various economic and business variables by applying differnt types of geometrical devices. For the presentation of statistical data on a graph, certain rules like title, structural frame - work, scale, false base line, line designs, etc., must be kept in mind. The data presented on graphs may be of two types; viz., (i) graphs of time series, and (ii) graphs of frequency distribution.

The graphs of time series are drawn either on arithmetic scale or on semi-logarithm scale. The time series graphs are further classified as graphs of one variable, graphs of two or more variables, range chart, band chart and ratio chart.

Graphs of frequency distribution are designed to reveal clearly the characteristic features of the frequency distribution. They are used to make comparative strudies about the pattern of two or more frequency distributions on the same graph paper. The frequency graphs are classified into histogram, frequency polygon, smooth frequency curve and ogive curve.

## 10.8 CHECK YOUR PROGRESS:MODEL ANSWERS

| | Graphic Presentation | Diagramatic Presentation |
|---|---|---|
| i) | Constructed on a graph paper and useful for studying mathematical relationship between two or more variables. | constructed on plain paper and is not useful for studying mathematical relationship between the variables. |
| ii) | These are useful in research for studying the slope of curves, rate of change, estimation of future values etc. | These draw the attention of public and are useful for publicity and propaganda. |
| iii) | Useful for presenting the time series data. | useful for presenting the categorical and geographical data. |

## 10.9 MODEL EXAMINATION QUESTIONS

**A. Short Questions**

1. What do you understand by graphic presentation of data?
2. What is a false base line?
3. What are time series graphs?
4. What are the rules of graphic presentation?
5. What is a Band graph?
6. What do you mean by ratio scale?
7. What are the limitations of arithmetic scale?
8. Distinguish between arithmetic scale and ratio scale?
9. What are the rules of interpretation of logarithmic curves?
10. What are the limitations of Ratio Scale?
11. What do you mean by graphs of frequency distribution?
12. What is a histogram?
13. What is a frequency polygon?
14. How a frequency polygon is prepared?
15. What do you understand by "Ogive" curves?
16. What is Less Than method of ogive curve?
17. What is the utility of an ogive curve?
18. What is a histogram? Distinguish between Histogram and Historigram?

32. Draw Less Than and More Than ogive curves for the following data and find the number of families with More Than 3 children and Number of families with Less Than 5 children.

| No.of Children | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| No.of Families | 150 | 72 | 50 | 30 | 15 | 10 | 5 |

33. The following information relates to the net woth of a company. Plot the data in the form of a semilogarithmic graph.

| Year | 1978 | 1979 | 1980 | 1981 | 1982 | 1983 |
|---|---|---|---|---|---|---|
| Net worth ('000 Rs.) | 100 | 115 | 120 | 135 | 140 | 145 |

34. Represent the following data by a band chart

| Year | Rice | Wheat | Pulses | Others |
|---|---|---|---|---|
| 1978 | 30 | 10 | 8 | 15 |
| 1979 | 32 | 12 | 10 | 20 |
| 1980 | 35 | 8 | 12 | 22 |
| 1981 | 38 | 14 | 13 | 23 |
| 1982 | 36 | 10 | 10 | 24 |
| 1983 | 40 | 15 | 8 | 25 |

35. Draw a histogram of the following frequency distribution and use it to find the total number of wage earners in the age group of 19-32 years.

| Age group (in years) | 14-15 | 16-17 | 18-20 | 21-24 | 25-29 | 30-34 | 35-39 |
|---|---|---|---|---|---|---|---|
| No. of Wage earners | 60 | 140 | 150 | 110 | 120 | 100 | 90 |

36. Prepare a graph of the following data by using a false base line :

Consumer price index numbers (1975=100)

| Year | 1977 | 1978 | 1979 | 1980 | 1981 | 1982 | 1983 |
|---|---|---|---|---|---|---|---|
| Centre: | | | | | | | |
| Delhi | 177 | 186 | 192 | 207 | 250 | 360 | 350 |
| Bombay | 185 | 199 | 211 | 222 | 265 | 337 | 340 |

37.Prepare the following data by means of a ratio scale graph:

| Year | Imports (Rs. in Crores) | Exports (Rs.in Crores) |
|---|---|---|
| 1978 | 440 | 335 |
| 1979 | 350 | 375 |
| 1980 | 370 | 365 |
| 1981 | 390 | 360 |
| 1982 | 420 | 365 |
| 1983 | 365 | 380 |

## 10.10 RECOMMENDED BOOKS


1.Gupta, S.P.: "Statistical Methods", Sultan chand & company, New Delhi.

2.Gupta, B.N.: "Statistics", Sahitya Bhavan, Agra.

3.Gupta,S.C.: "Fundamentals of Statistics", Himalaya Pub. House, Bombay.

4.Simpson and Kafka: "Basic Statistics", Oxford and I.B.H. Publishing Company, Calcutta.


## 10.11 GLOSSARY

1. False base line : It represents two zig-zag horizontal lines drawn above the line of X axis breaking the vertical axis. These lines are drawn where the starting value in the given data is very distant from zero.

2. Frequency polygon: Frequency polygon is a device of graphic presentation of frequency distribution of frequency. It is obtained by plotting frequencies on Y axis and the corresponding values of such frequencies on X axis in pairs. These points are joined by straight line. It can also be constructed with the help of a histogram.

3. Histogram: It is a graph constructed by taking class values on the "X"axis and the frequency on the "Y" axis. The widths of classes may be equal or may not be equal.

4. Historigram: It is a graph constructed with the help of time series data.

## BLOCK -II: MEASURES OF CENTRAL TENDENCY

## UNIT -11 : INTROUDCTION TO AVERAGES

**contents**



### 11.0 AIMS AND OBJECTIVES

The aims of this unit is to present the meaning, objectives, types of averages, requisites of a good average and limitations of averages.

After reading this unit, you should be able to :

- explain the meaning and importance of averages
- recognise the objectives of averages
- identify the various types of averages
- describe the requisites of a good average
- list the limitations of an average
- decide the choice of an average

### 11.1 INTRODUCTION

In the earlier units ' Tabulation', Diagrammatic and graphic presentation of Data have been discussed. Since these techniques do not tell the complete story of the phenomenon, there is a need to calculate single representative values which are also known as averages. Measures of

central tendency are the most popularly used tools to condense the mass data and represent it through single numbers. Thus averages are single figures which sum up all the characteristics of huge data.

## 11.2 MEANING AND IMPORTANCE OF AVERAGES

The object of statistical analysis is to get one single value which can represent the characteristic features of entire distribution. For this the tabulated data need to be processed and reduced to a single representative figure. Pointing out the need for such a figure, R.A.Fisher says " The inherent inability of the human mind to grasp in its entirety a large body of the numerical data compels us to seek relatively few constants that will adequately descirbes the data". K.A. Yeomans states, " A concise numerical discription is often preferable to a lengthy tabulation, and this form of discription also enables us to form a mental image of the data and interprete its significance ". Hence, there is a greater need to develope a single numerical figure which is representative of the data. One of the important and essential measures of summarsing the data in statistical analysis is average. Average is the method of reducing the mass and complex data into a single, representative numerical figure. Thus, in the words of A.E. Waugh, " An average value is a single value selected from a group of values to represent them in some way-a value which is supposed to stand for whole group of which it is a part, as typical of all the values in the group ". According to A.L.Bowley, " Averages are statistical constants which enable us to comprehend in a single effort the significance of the whole". For Croxden and Cowden . " An average is a single value with in the range of data that is used to represent all of the values in the series". Averages described as ' measures of Central Tendency' because the individual value of the variables usually cluster around such values. Accoding to Lawrence J.Kaplan " Statistical analysis, seeks to develop concise summary figures which describe large body of quantitative data. One of the most widely used set of summary figures is known as measures of location, which are often referred to as averages, measures of central tendency or central location. The purpose of computing an average value for a set of observations is to obtain a single value which is representative of all the items and which the mind can grasp easily and quickly. The single value is the point or location around which the individual items cluster".

Averages occcupy a prominent place in statistical analysis. Most of the statistical techniques are based on the averages. Averages are more widely used than any other measures because of their many applications. Many authors have realised the importance of averages in statistics. Bowley went to the extent of defining statistics as the ' Science of Averages'. While emphasing the importance of averages, L.H.C. Tippet observes,"The average has its limitations, but provided they have recognised, there is no single statistical quantity more valuable than an average".

## 11.3 OBJECTIVES OF AN AVERAGE

The objectives of averages are:

**i) To get single value Representing the characteristics of the data.**

An average, by summarising the complex data in to one single value, facilitates to get a

broad picture of the whole data. In this context M.J.Moreney says" The purpose of an average is to represent a group of individual values in a simple and concise manner, so that, the mind can get a quick understanding of the general size of the individual in the group". If the average value is computed, fruitful interpretations can be made. Thus, if the data relating to the wages paid to 1,000 workers in a factory are given, the characteristics cannot be interpreted. But if average wages paid are computed, it will be useful for interpretation.

**ii) To Facilitate Comparison**

As the average provides a common denominator, one set of data can be compared with another set of data and conclusions can be drawn on the characteristics of different sets of data. The comparisons can be made among different sets of data at a specific point of time, such as the average profits of company 'A' and company 'B' on 1st January, 1983. Comparisons can also be made for the data relating to different time periods, such as the average profits of a company in 1982 and 1983 etc.

**iii) To know about Population (Universe) from a Sample**

In statistical investigations, sample methods are frequently used. If the sample selected is representative, the average of the sample not only describes the characteristics of sample distribution but also the features of universe or population. Thus, averages of samples are often used to interpret the characteristics of universe.

**iv) To Trace Mathematical Relationship**

As the averages are expressed in numerical figures, they explain the mathematical relationship among various groups. For instance, if we know that the per capita income of a Russian is more than that of the per capita income of an Indian, it may not convey the correct meaning because it is not expressed quantitatively by a single figure. But if the average per capita income of Russia and India are given, they would convey a definite meaning.

**v) To Help in Decision Making**

The averages facilitate the decision making in different fields. Averages are quite helpful in setting standards, estimating, planning and other managerial decision making areas. Many examples such as average sales, average production, average cost, average expenditure etc., can be cited as bases for decision making.

**Check your progress - 1**

List out the objectives of averages.

## 11.4 TYPES OF AVERAGES

The following are the various measures of central tendencies

i) Arithmetic mean.

ii) Median

iii) Mode

iv) Geometric mean

v) Harmonic mean

vi) Moving average

vii) Progressive average

viii) Composite average

Among the above averages, median and mode are called positional averages, whereas arithmetic mean, geometric mean and harmonic mean are called mathematical averages. In addition the moving average, progressive average and the composite average are called commercial averages. Which are less used.

## 11.5 REQUISITES OF A GOOD AVERAGE

As the average is a single value which represents the group of items, it must possess certain properties. They are listed by J.F.Kenney and E.S.Keeping. According to them, an average should be rigidly defined on all the observed values and capable of mathematical treatment. Further, it should not be unduly influenced by one or two extreme values. The average should fluctuate relatively little from one random sample to another. Yule and Kendall also describe the requisites of a good average. They are as below:

**i) Average should be rigidly Defined**

An average should be rigidly defined so that only one interpretation can be made from it. If the definition is not specific and clear, it may give scope for personal prejudice and bias of the investigator. Further, the definition of an average should be expressed in the form of algebraic formula, so that, there can be no ambiguity in its computation.

**ii) Average should be based on all the items of a variable**

Every item of the distribution should be taken into account for computing the average. If the average is not so based it is not regarded as the representative value of the group.

**iii) Average should be easy to follow and simple to understand**

An average must be easily understood. If it is otherwise its utility will be restricted.

**iv) A rage should not be affected by any single item**

Th h every item of a variable is taken into account while computing the average, no single or group o ms should unduly influence the average. If one or two values in a distribution influence the ave age, it should not be considered as a typical value of the entire distribution.

**v) Average should be capable of further statistical treatment**

The average computed should lend itself to further statistical analysis. For example, if the mean and median of a distribution are known, mode can be computed easily. In the same manner, for computing Standard Deviation and Coeffient of Correlation, Arithmetic mean should be used.

**vi) Averages should have sampling stability.**

Averages should not be affected by sampling fluctuations. If two samples are drawn from universe the averages of samples should be very near to each other. The average which is least affected by fluctuations in the sample is considered as good one. Again, the average which shows little variation and more stability should be regarded as a representative one.

## 11.6 LIMITATIONS OF AN AVERAGE

Eventhough averages are widely applied in statistical analysis, they suffer from the following limitations.

i) As the average is a single value representing the phenomena of the given series, necessary care should be taken while interpreting its values. Otherwise, sometimes it may give scope for wrong conclusions. For instance, if the marks obtained by 10 students are 0,0,0,0,0,0, 100, 100, 100, 100 the arithmetic average works out to 40. If interpretation is not made carefully, this figure may give a wrong impression that the students are moderately intelligent.

ii) The average of a given distribution may give a value which is not found in the distribution. For example, the arithmetic mean of 10, 20, 30 and 40 is 25 which is not existing in the series.

iii) If a suitable average is not selected judiciously it may give wrong and fallacious conclusions.

iv) The measures of central tendency fails to give an idea about the formation of the series. Sometimes the average values of two or more series may be the same but they differ in distribution and composition. The following two series help in understanding the phenomena:

| X series | Y series |
|---|---|
| 100 | 800 |
| 200 | 600 |
| 300 | 50 |
| 500 | 40 |
| 400 | 10 |
| 1500 | 1500 |
| $\overline{X} = 300$ | $\overline{Y} = 300$ |

Here, the average values for the two distributions are the same but their composition is different. These differences are not revealed by the average.

## 11.7 CHOICE OF AN AVERAGE

The choice of a particular average depends upon the nature of data and definitions of representative value required for the statistical investigation. Different averages possess different characteristics and there is no average which suits all the requirements of all situtations. However, while selecting an average, the following, points should be taken into consideration.

i) **Objective :** The average should be selected in accordance with the purpose of statistical investigation. For example, if all the items of the distribution are to be given equal importance arithmetic mean is more suitable. To find out most frequently occuring items in the series, mode is an appropriate average. Where the averages are to be used to indicate the position, computation of median can be a useful measure. If the data requires that more importance be given to small items than big items, geometric mean is used. Further, if the data requires that more weight be given to small items, computation of harmonic mean is an appropriate measure.

ii) **Representative :** The selected average of a distribution should represent the basic characteristics of the data.

iii) **Nature of the Data :** The choice of the average to be used depends upon the nature of data.

a) When the data is symmetrically distributed, mean, mode or median may be used interchangeably.

b) If the frequency distribution is made by open-end classes, mean is not an appropriate average.

c) If the data given is in unequal class intervals, it is not possible to determine the mode accurately.

According to Ya-Lun-Chou, the following questions are to be answered in determining the choice of an average.

a) What is the purpose that the average is designed to serve?

b) Should we permit extreme values in the series to influence the averages?

c) should the average be used for further computation?

d) what is the model of distribution? Is it symmetrical or skewed?

e) How to record the observations to be averaged ? Natural numbers? ratios ? rates? or averages?

f) What would be the class of units used for the average? Are the observations expressed inversely to what is required in the average?

g) In what sense do we expect the average to be typical? To balance the individual values? To balance the ratios? To balance the number of items?

h) what weights, implicit or explicit, should be used ?

While discussing the choice of average, Secrist observed, "The justification of employing them (averages) must be determined by an appeal to all facts and in the light of peculiar characteristics of different types". Yule felt that the arithmetic mean is the most ideal average as it is familiar and has very wide applications in statistical theory at large. The student is advised to study once again the limitations and choice of averages after going through all the measures of central tendency for better understanding.

## 11.8 SUMMING UP

Averages are useful means of condensing the data into a single numerical figure which represents the characteristics of a given distribution . The averages are also called measures of central tendency because all the items of the distribution cluster around the average value. The types of averages include; arithmetic mean, median, mode, geometric mean, harmonic mean, moving average, progressive average and composite average. A good average must be rigidly defined, based on all the observations of series and capable of further algebraic treatment. While using the averages in statistical analysis, its limitations should be kept in mind. Otherwise, the results may be misleading. The choice of a particular average depends upon the objective of investigation and the nature of data.

## 11.9 CHECK YOUR PROGRESS : MODEL ANSWERES

1. The Objectives are :

i) To obtain a single value representing the whole data.

ii) To facilitate comparison.

iii) To know about the universe from a sample

iv) To trace mathematical relationship between the groups

v) To help in decision making



## 11.10 MODEL EXAMINATION QUESTIONS

**A. Short Questions**

1. What is meant by 'average'?
2. What are the objectives of averages?
3. What are the various types of averages?
4. List out the requisites of a good average?

**B. Essay Questions**

5. What considerations do you take into account while selecting an average?
6. Discuss the merits and limitations of measures of Central Tendency?

## 11.11 RECOMMENDED BOOKS


1. Gupta, S.P. : " Statistical Methods". Sultan Chand & Company, New Delhi.
2. Gupta, B.N. : " Statistics", Sahitya Bhavan, Agra.
3. Gupta, S.C. : " Fundamentals of Statistics" Himalaya Publishing house House, Bombay.
4. Simpson and Kafka : " Basic statistics", Oxford and I.B.H. Publishing Company, Calcutta.


## 11.12 GLOSSARY

Average : Average is a statistical measure which summarises the whole data into a single numerical value.

# UNIT-12 : ARITHMETIC MEAN

**Contents**



## 12.0 AIMS AND OBJECTIVES

The aims of this unit are to explain the meaning, methods of computation, properties, merits and limitations of Arithmetic Mean.

After going through this unit, you should be able to :

- explain the meaning of simple and weighted Arithmetic Means
- Compute Simple and Weighted Arithmetic means
- Identify the properties of Arithmetic mean
- List the merits and limitations of Arithmetic mean.

## 12.1 INTRODUCTION

The purpose of averages is to present a heap of data into a single value. As we have stated in unit 11, there are eight types of averages. In this course, we discuss the first five averages viz., arithmetic mean, median, mode, geometric mean and harmonic mean only. Arithmetic mean is a mathematical average whose computation is based on all the observations of a given series. This unit is about the theoretical and computational aspects of arithmetic mean.

## 12.2 MEANING OF ARITHMETIC MEAN

Arithmetic mean is a statistical measure which is widely used to represent the whole data with the help of a single value. While the arithmetic mean calculated for a sample data is denoted by '$\overline{X}$' (pronounced as X-bar), the arithmetic mean computed for a universe is denoted by $\mu$ (pronounced as 'mu'). Usually arithmetic mean is simply known as 'the mean ' or the average. Arithmetic mean may be of two types, viz., simple arithmetic mean and weighted arithmetic mean. Thus appropriate adjectives are added to the word 'arithmetic mean' to suit the requirements of the statistical analysis.

**Simple Arithmetic Mean**

According to Simpson Kafka, 'Arithmetic mean is the quotient that results when the sum of all the items in the series is divided by number of items'. Ya-Lun- chou defined arithmetic mean as 'the sum of the observations in a sample divided by the number of observations in the sample'.

## 12.3 COMPUTATION OF SIMPLE ARITHMETIC MEAN

Arithmetic mean is computed for the data relating to individual, discrete and continuous series.

**Individual series**

Arithmetic mean for the data of an individual series can be computed either by the direct method or short-cut method.

**Direct Method:** In this method, arithmetic mean is obtained by dividing the sum of the values of all observations in a series by the number of items constituting the series.

Symbolically:

$$\overline{X} = \frac{X_1 + X_2 + X_3 + X_4 + .....X_n}{N}$$

$$\overline{X} = \frac{\Sigma X}{N}$$

Where, $\overline{X}$ = Arithmetic mean

$\Sigma X$ = Sum of the values of all the observations of a variable

N = Number of observations.

**Illustration-I**

The following are the marks obtained by 10 students in accountacy subject.

| S.No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Marks in Accountancy | 56 | 63 | 45 | 75 | 66 | 47 | 80 | 36 | 29 | 54 |

Calculate arithmetic mean of the marks obtained by the students in accountancy subject.

**Solution:**

**CALCULATION OF ARITHMETIC MEAN**

| S.No. | Marks in Accountancy (X) |
|---|---|
| 1. | 56 |
| 2. | 63 |
| 3. | 45 |
| 4. | 75 |
| 5. | 66 |
| 6. | 47 |
| 7. | 80 |
| 8. | 36 |
| 9. | 29 |
| 10. | 54 |
| N = 10 | ΣX = 551 |

Arithmetic Mean $(\overline{X}) = \frac{\Sigma X}{N}$

$\Sigma X = 551, \quad N = 10$

Substituting the values in to formula,

$$\overline{X} = \frac{551}{10}$$

$$= 55.1$$

Thus the arithmetic mean of marks obtained by the students in Accountancy subject is 55.1.

**Short-cut Method:** The direct method of calculating, arithmetic is convenient only when the observations are very few and their values are represented by small numbers. When the variable constitutes large number of observations, computation of arithmetic mean involves more calculations. To overcome this difficulty, short-cut method is used. This method is based on the property of arithmetic mean, 'the sum of deviations of the variable from the mean is equal to zero'. Thus, the arithmetic mean of 1,2,3,4 and 5 is equal to 3. If the deviation of each of these observations from the mean (3) is calculated, it would $-2, -1$, and $+1$, $+2$ and their sum is equal to zero.

Symbolically,

$$\Sigma (X - \overline{X}) = 0$$

Arithmetic mean is computed by taking an arbitrary origin which is known as assumed mean. The assumed mean need not be picked up from among the values of the variable. It can be any figure as per the convenience of the user. The following steps are required to calculate the arithmetic mean.

i) Assume a value as mean and denote it by 'A'

ii) Take the deviations of all the values of the variable from the assumed mean (i.e., X−A) and denote these deviations by dx.

iii) Add all these deviations and obtain $\Sigma$ dx.

iv) Apply the formula given below and obtain $\overline{X}$.

$$\overline{X} = A + \frac{\Sigma dx}{N}$$

Where, $\overline{X}$ = Arithmetic mean

A = Assumed Arithmetic mean

$\Sigma$ dx = Sum of deviations of all values of the variable taken from an assumed mean

N = Number of observations.

**Illustration-II**

Following are the monthly incomes of 8 workers of a factory. Find the arithmetic mean of incomes of workers per month.

Monthly Income (in Rs.) 150, 162, 173, 179, 180, 200, 210, 205.

**Solution**

Calculation of Arithmetic mean ;

Let us take Rs. 180 as 'assumed mean'

Hence A = 180

| S.No. | Monthly income (in Rs.) X | (X−A) dx |
|---|---|---|
| 1. | 150 | − 30 |
| 2. | 162 | − 18 |
| 3. | 173 | − 7 |
| 4. | 179 | − 1 |
| 5. | 180 | 0 |
| 6. | 200 | 20 |
| 7. | 210 | 30 |
| 8. | 205 | 25 |
| N = 8 | | $\Sigma$ dx = 19 |

$$\overline{X} = A + \frac{\Sigma dx}{N}$$

Here, A = 180, $\Sigma$ dx = 19, N = 8

Substituting the values in the formula,

$$\overline{X} = 180 + \frac{19}{8}$$

$$= 180 + 2.37$$

$$= 182.37$$

Thus, the average monthly income of workers is Rs.182.37.

**Discrete Series**

Arithmetic mean for the data of discrete series can be computed either by direct method or by short-cut method.

The following steps are involved in the computation of arithmetic mean of discrete series by direct method.

i) Multiply each value of the variable with its corresponding frequency and denote the column by fX.

ii) Add all the values of fX and obtain $\Sigma$ fX

iii) Obtain the total frequencies of the distribution and denote it by N or $\Sigma$ f.

iv) Substitute these values in the formula given below and obtain the values of $\overline{X}$

$$\overline{X} = \frac{\Sigma fX}{N}$$

Where, $\overline{X}$ = Arithmetic mean

f = Frequency

$\Sigma$ fX = Sum of the products of the values of variable and their respective frequencies.

N or $\Sigma$ f = Total number of observations or total frequencies.

**Illustration - III**

Given below is the information relating to wages of workers of a factory. Find out the mean wages of workers per day.

| Wages per day (in Rs) : | 10 | 12 | 13 | 14 | 16 | 17 | 20 | 21 | 22 | 25 |
|---|---|---|---|---|---|---|---|---|---|---|
| Number of workers : | 1 | 6 | 10 | 20 | 23 | 15 | 16 | 4 | 3 | 2 |

**Solution :**

**COMPUTATION OF MEAN WAGES OF WORKERS**

| Wages per day (in Rs.) X | Number of Workers f | fX |
|---|---|---|
| 10 | 1 | 10 |
| 12 | 6 | 72 |
| 13 | 10 | 130 |
| 14 | 20 | 280 |
| 16 | 23 | 368 |
| 17 | 15 | 255 |
| 20 | 16 | 320 |
| 21 | 4 | 84 |
| 22 | 3 | 66 |
| 25 | 2 | 50 |
| | N = 100 | Σ f X = 1635 |

Arithmetic Mean $(\overline{X}) = \frac{\Sigma fX}{N}$

Here, $\Sigma$ fX = 1635, N = 100

Substituting the values in the formula,

$$\overline{X} = \frac{1635}{100}$$

$$= 16.35$$

Thus, the mean wages of workers per day is Rs. 16.35.

**Short-cut method:** In the case of short-cut method the following steps are required to calculate the arithmetic mean.

i) Assume any value as the mean of the series and denote it by 'A'.

ii) Take the deviations of the variable (X), from the assumed mean (A) and denote the column by dx (dx = X − A).

iii) Multiply the deviations (dx) with their respective frequencies (f) and denote the column by fdx.

iv) Find out the total of (fdx) column and obtain $\Sigma$ fdx.

v) Add all frequencies of the distribution and obtain N.

vi) Substitute the values in the formula given below and obtain $\overline{X}$

$$\overline{X} = A + \frac{\Sigma f dx}{N}$$

Where,

$\overline{X}$ = Arithmetic mean

A = Assumed mean

$\Sigma fdx$ = Sum of the products of deviations from assumed mean and their respective frequencies

N = Total number of observations.

**Illustration - IV :**

The following is the information relating to weekly wages of workers in a factory. Find out the mean wages of workers per week.

| Weekly wages (in Rs) | 25 | 30 | 35 | 40 | 45 | 50 | 55 | 60 |
|---|---|---|---|---|---|---|---|---|
| Number of workers | 3 | 61 | 132 | 153 | 70 | 70 | 51 | 2 |

**Solution:**

**COMPUTATION OF WAGES OF WORKERS**

Let us assume '40' as the mean of the series i.e., A = 40

| Weekly wages (in Rs.) X | Number of workers f | (X - A) dx | fdx |
|---|---|---|---|
| 25 | 3 | - 15 | - 45 |
| 30 | 61 | - 10 | - 610 |
| 35 | 132 | - 5 | - 660 |
| 40 | 153 | 0 | 0 |
| 45 | 70 | 5 | 350 |
| 50 | 70 | 10 | 700 |
| 55 | 51 | 15 | 765 |
| 60 | 2 | 20 | 40 |
| | N = 542 | | $\Sigma$ fdx = 540 |

Arithmetic Mean $\overline{X} = A + \frac{\Sigma f dx}{N}$

Here, $\Sigma$ fdx = 540, N = 542, A = 40

Substituting the values in the formula,

$$\frac{N}{pfC} + V = \underline{X}$$

$$\overline{X} = 40 + \frac{540}{542}$$

$$= 40 + 0.99$$

$$= 40.99$$

Weekly mean wages of workers mean wages of workers is Rs. 40.99.

**Check your progress - 1**

Compute Arithmetic mean for the following data.

| Marks | 52 | 55 | 58 | 60 | 63 | 67 | 69 | 70 | 80 | 85 | 90 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| No of Students | 10 | 30 | 40 | 50 | 30 | 25 | 45 | 65 | 60 | 20 | 10 |

**CONTINUOUS SERIES**

In case of continous series, arithmetic mean is calculated by the following methods

i) Direct method

ii) Short-cut method

iii) Step-Deviation method

iv) Summation method.

**Direct Method :** As per the direct method arithmetic mean is calculated with the help of the following steps.

i) Find out the mid values of each class and denote the column by 'm' (this can be obtained by dividing the total of upper value and lower value of each class by two).

ii) Multiply the mid values of each class with their respective frequencies to obtain 'fm'.

iii) Add all values of 'fm' to obtain $\Sigma$ 'fm'.

iv) Add all frequencies of the classes and denote it by N.

v) Apply the following forumula and obtain $\overline{X}$.

$$\overline{X} = \frac{\Sigma fm}{N}$$

Where,

$\overline{X}$ = Arithmetic mean

m = Mid values of various classes

$\Sigma fm$ = Sum of products of mid values and their respective frequencies.

N = Number of observations or total frequencies.

**Illustration - V**

From the following frequency distribution, find out arithmetic mean.

| Weight (in Kgs.) | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 | 70-80 |
|---|---|---|---|---|---|---|---|
| No. of Persons : | 15 | 20 | 25 | 30 | 18 | 12 | 6 |

**Solution :**

**COMPUTATION OF AVERAGE WEIGHT OF PERSONS**

| Weight (in Kgs) | Mid values | No.of persons | fm |
|---|---|---|---|
| 10-20 | 15 | 15 | 255 |
| 20-30 | 25 | 20 | 500 |
| 30-40 | 35 | 25 | 875 |
| 40-50 | 45 | 30 | 1,350 |
| 50-60 | 55 | 18 | 990 |
| 60-70 | 65 | 12 | 780 |
| 70-80 | 75 | 6 | 450 |
| | | N = 126 | Σ fm = 5,170 |

$$\overline{X} = \frac{\Sigma fm}{N}$$

Here, Σfm = 5,170, N = 126

Substituting the values in the formula,

$$\overline{X} = \frac{5,170}{126}$$

$$= 41.03$$

Thus, the mean weight of persons is 41.03 Kgs.

**Short-Cut Method** According to the short-cut method arithmetic mean is calculated as per the following procedure.

i) Find out mid values of each class.

ii) Take a mid value as assumed mean and denote it by 'A'

iii) Take the deviations of the mid values from the assumed (m − A) and denote the column by d.

iv) Multiply the deviations (d) by their respective frequencies and denote the column by 'fd'

v) Obtain the total of 'fd' and denote it by Σfd.

vi) Add the frequencies and obtain N.

vii) Apply the following formula and obtain $\overline{X}$

$$\overline{X} = A + \frac{\Sigma fd}{N}$$

Where,

$\overline{X}$ = Arithmetic mean.

A = Assumed mean

d = Deviations taken from the assumed mean of mid values (m − A).

Σfd = Sum of products of deviations and their respective frequencies. N = Total number of observations or total frequencies

**Illustration - VI**

Calculate arithmetic mean of marks from the following data.

| Marks : | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 |
|---|---|---|---|---|---|---|---|
| No.of students: | 5 | 13 | 19 | 20 | 25 | 12 | 6 |

**Solution :**

**COMPUTATION OF ARITHMETIC MEAN**

Let us take as assumed mean i.e., A = 35

| Marks | Number of students | m | d(A−m) | fd |
|---|---|---|---|---|
| 0-10 | 5 | 5 | − 30 | − 150 |
| 10-20 | 13 | 15 | − 20 | − 260 |
| 20-30 | 19 | 25 | − 10 | − 190 |
| 30-40 | 20 | 35 | 0 | 0 |
| 40-50 | 25 | 45 | 10 | 250 |
| 50-60 | 12 | 55 | 20 | 240 |
| 60-70 | 6 | 65 | 30 | 180 |
| | N = 100 | | | Σfd = 70 |

**Here, $\Sigma fd = 70$, $N = 100$, $A = 35$**

**Substituting the values in the formula,**

$$\overline{X} = 35 + \frac{70}{100}$$

$$= 35 + 0.7$$

$$= 35.7$$

**Thus, the mean marks of the students is 35.7**

**Step-Deviation Method :** In case of step deviation method, the calculations involved in the computation of arithmetic mean are further reduced by taking out the common factor from the deviations of mid values derived from the assumed mean. Usually the value of common factor would be the size of the class interval.

Symbolically,

$$\overline{X} = A + \frac{\Sigma f d^1}{N} \times C$$

Here,

$\overline{X}$ = Arithmetic mean

$d^1$ = Step deviation of mid values taken from the assumed mean.

$\Sigma f d^1$ = Sum of products of step diviation and their respective frequencies.

N = Number of observations.

C = Common factor.

**Illustration - VII**

Calculate Arithmetic mean from the following data.

| Marks | Number of students |
|---|---|
| 25-30 | 25 |
| 30-35 | 40 |
| 35-40 | 60 |
| 40-45 | 75 |
| 45-50 | 20 |
| 50-55 | 10 |

**Solution :**

**COMPUTATION OF AVERAGE MARKS OF STUDENTS**

Let us take 37.8 assumed mean i.e., A = 37.5

| Marks x | No.of students f | m | (m−A) d | d/c C = 5 $d^1$ | $fd^1$ |
|---|---|---|---|---|---|
| 25-30 | 25 | 27.5 | −10 | −2 | −50 |
| 30-35 | 40 | 32.5 | −5 | −1 | −40 |
| 35-40 | 60 | 37.5 | 0 | 0 | 0 |
| 40-45 | 75 | 42.5 | 5 | 1 | 75 |
| 45-50 | 20 | 47.5 | 10 | 2 | 40 |
| 50-55 | 10 | 52.5 | 15 | 3 | 30 |
| | N = 230 | | | | $\Sigma fd^1$ = 55 |

Arithmetic Mean $(\overline{X}) = A + \frac{\Sigma fd^1}{N} \times C$

Here, A = 37.5, $\Sigma fd^1 = 55$, N = 230

Substituting the values in the formula,

$$\overline{X} = 37.5 + \frac{55}{230} \times 5$$

$$= 37.5 + 0.239 \times 5$$

$$= 37.5 + 1.2$$

$$= 38.7$$

Thus, the average marks of students is 38.7

**Summation Method:** If the class intervals of continuous series are of equal size, arithmetic mean can also be computed with the help of summation method. This method involves the following steps.

i) Find out mid values of the distribution and denote the highest mid values as m.

ii) Add all frequencies of the distribution to obtain N.

iii) Calculate cumulative frequencies of all classes and add them to obtain $\Sigma$c.f.

iv) Divide sum of cumulative frequencies by N and obtain F.

v) Apply the following formula and obtain $\overline{X}$

$\overline{X} = m - i\,(F - 1)$

Where,

$\overline{X}$ = Arithmetic mean.

m = Mid value of the last class interval (i.e., mid value of highest class in the distribution)

i = Class interval

F = Sum of cumulative frequencies divided by the total frequencies.

**Illustration - VIII**

Calculate arithmetic mean of the following data by summation method.

| Wages (in Rs.) | Number of workers |
|---|---|
| 300-310 | 6 |
| 310-320 | 20 |
| 320-330 | 44 |
| 330-340 | 26 |
| 340-350 | 3 |
| 350-360 | 1 |

**Solution:**

**COMPUTATION OF AVERAGE WAGES OF WORKERS**

| Wages (in Rs.) X | No. of Workers f | cf | m |
|---|---|---|---|
| 300-310 | 6 | 6 | |
| 310-320 | 20 | 26 | |
| 320-330 | 44 | 70 | |
| 330-340 | 26 | 96 | |
| 340-350 | 3 | 99 | |
| 350-360 | 1 | 100 | 355 |
| | N = 100 | Σcf = 397 | |

$$\overline{X} = m - i\,(F - 1)$$

Here,

$i = 10, \quad F = \frac{397}{100} = 3.97, \quad m = 355$

Substituting the values in the formula,

$$\overline{X} = 355 - 10\,(3.97 - 1)$$
$$= 355 - 10\,(2.97)$$
$$= 355 - 29.7$$
$$= 325.3$$

Thus, the average wages Rs. 325.3.

## COMPUTATION OF ARITHMETIC MEAN, OPEN-END CLASSES

The arithmetic mean in case of open-end classes cannot be calculated unless the values of lower limit of first class and upper limit of the last class are ascertained. To determine such values, the pattern of class intervals of other class should be observed carefully. If the size of all the class intervals is equal, then the size of the first and last class should be determined in accordance with the size of other classes.

**Illustration-IX**

Find out the average wages of the workers from the following information.

| Wages (in Rs.) | Number of workers |
|---|---|
| Less than 150 | 10 |
| 150-200 | 15 |
| 200-250 | 27 |
| 250-300 | 25 |
| 300-350 | 15 |
| 350 and above | 8 |

**Solution:**

In the above frequency distribution, the size of all the class intervals is 50. Hence, the size of the class interval of the first and last class also can be deemed as 50. Thus the value of lower limit of the first class is determined as 100 and that of upper limit of the last class as 400.

Arithmetic mean can now be computed after rearranging given data as below:

Let us take 225 as assumed mean i.e.,

A = 225

C = 50

| Wages (in Rs.) | No. of workers f | m | (A−m) d | d/c $d^1$ | $fd^1$ |
|---|---|---|---|---|---|
| 100-150 | 10 | 125 | −100 | −2 | −20 |
| 150-200 | 15 | 175 | − 50 | −1 | −15 |
| 200-250 | 27 | 225 | 0 | 0 | 0 |
| 250-300 | 25 | 275 | 50 | 1 | 25 |
| 300-350 | 15 | 325 | 100 | 2 | 30 |
| 350-400 | 8 | 375 | 150 | 3 | 24 |
| | N = 100 | | | | $\Sigma fd^1 = 44$ |

$$\overline{X} = A + \frac{\Sigma f d^1}{N} \times C$$

Here, A = 225, $\Sigma f d^1$ = 44, N = 100, c = 50

Substituting the values in the formula,

$$\overline{X} = 225 + \frac{44}{100} \times 50$$

$$= 225 + 0.44 \times 50$$

$$= 225 + 22$$

$$= 247$$

Thus, the average wage is Rs. 247.

When the class intervals are not uniform for all the classes in the frequency distribution, the lower and upper limits of the classes cannot be ascertained rationally. In such cases the given data are not amenable for computing arithmetic mean and as such it is better to use some other measures of average such as median and mode.

## 12.4 COMPUTATION OF COMBINED MEAN

In case the arithmetic means and sizes of two more related groups are known, the combined mean of all the groups can be obtained by combining the given series. The combined mean is denoted by $\bar{X}_{12}n$ and computed with the help of the following formula:

$$\overline{X}_{12n} = \frac{N_1\overline{X}_1 + N_2\overline{X}_2 + \ldots\ldots\ldots\ldots N_n\overline{X}_n}{N_1 + N_2 + \ldots\ldots\ldots\ldots N_n}$$

Where,

$\overline{X}_{12n}$= Combined mean of all the groups under study;

$\overline{X}_1, \overline{X}_2 \ldots\ldots \overline{X}_n$ = Arithmetic means of various groups of data;

$N_1, N_2 \ldots\ldots\ldots N_n$ = Number of observations of various groups

**Illustration- X**

There are two sections of a class in a school in which the students are 80 and 70 respectively. If the arithmetic means of marks secured by the students of two sections are 65 and 55 respectively, find out the combined arithmetic mean of the marks secured by the students of the class as a group.

**Solution:**

$$\overline{X}_{12} = \frac{N_1\overline{X}_1 + N_2\overline{X}_2}{N_1 + N_2}$$

we are given that,

$N_1 = 80, N_2 = 70, \overline{X}_1 = 65, \overline{X}_2 = 55$

Substituting the values in the formula,

$$\overline{X}_{12} = \frac{(80 \times 65) + (70 \times 55)}{80 + 70}$$

$$= \frac{5,200 + 3,850}{150}$$

$$= \frac{9,050}{150}$$

$$\overline{X}_{12} = 60.33$$

The combined average marks secured by the students of a class as a group = 60.33

## 12.5 COMPUTATION OF WEIGHTED ARITHMETIC MEAN

While computing simple airthmetic mean all items in the distribution are given equal importance despite their varying magnitudes. In practice the data may consist some of the values which are more important than the other values. In such cases, simple arithmetic mean cannot be true representative of the distribution. Hence, proper weightage has to be given to the various items of the data. Weights are to be assigned to various items of the series in accordance with their relative importance. When all the values of variable are assigned equal weights, the weighted arithmetic mean is equal to arithmetic mean i.e., $\overline{X}_w = \overline{X}$. On the other hand if wieghts are assigned on the basis of relative importance of the values (i.e., large weights for large items) weighted arithmetic mean will be greater than the simple arithmetic mean, i.e., $\overline{X}_w > \overline{X}$. In the words of Boddington, 'This principle of weighted average is very important in all cases where varying quantities are in evidence and it will be necessary, for instance to use in a factory, where average cost per unit of the commodities manufactured is desired, also where the average output per machine is required and the machines are of different patterns or are working under varying conditions'.

Weighted arithmetic mean is denoted by $\overline{X}_w$ and is computed either by direct method or by short-cut method.

**Direct Method:** In the case of direct method, weighted arithmetic mean is computed with the help of the following steps:

i) Assign weights to each value of the series with their relative importance, if they are not given in the problem.

ii) Multiply each value of the series with their respective weights and denote the column by WX.

iii) Find out the total of weights and obtain $\Sigma W$.

iv) Find out the total of column (ii) and obtain Σ WX.

v) Apply the following formula and obtain the value of $\overline{X}_w$.

$$\overline{X}_w = \frac{\Sigma WX}{\Sigma W}$$

Where,

$\overline{X}_w$ = Weighted arithmetic mean

W = Weights

ΣW = Sum of weights

ΣWX = Sum of the products of values of variable and their respective weights assigned.

**Illustration- XI**

Find out weighted arithmetic mean of index numbers from the following data.

| Items | Index numbers | Weghts |
|---|---|---|
| A | 152 | 35 |
| B | 125 | 23 |
| C | 141 | 18 |
| D | 133 | 15 |
| E | 100 | 9 |

**Solution:**

COMPUTATION OF WEIGHTED ARITHMETIC MEAN OF INDEX NUMBERS

| Items | Index numbers | Weights | XW |
|---|---|---|---|
| A | 152 | 35 | 5.320 |
| B | 125 | 23 | 2,875 |
| C | 141 | 18 | 2,538 |
| D | 133 | 15 | 1,995 |
| E | 100 | 9 | 900 |
| | | ΣW = 100 | ΣXW = 13,628 |

$$\overline{X}_w = \frac{\Sigma XW}{\Sigma W}$$

**Here,**

**Σ WX = 13,628, $\Sigma W$ = 100**

**Substituting the values in the formula,**

$$\overline{X}_w = \frac{13,628}{100}$$

$$= 136.28$$

The weighted arithmetic mean of index number = 136.28.

**Short-cut Method:** In case of short-cut method the procedure given below is followed to compute weighted arithmetic mean:

i) Assign weights to each value of the series on the basis of their relative importance, if they are not given in the problem., i.e., W.

ii) Take deviations of all the values from the assumed weighted mean and denote the column by dx.

iii) Multiply the deviations with their respective weights and obtain the total i.e., ΣWdx.

iv) Add the weights and obtain Σ W.

v) Apply the following formula and compute $\overline{X}_w$.

$$\overline{X}_W = A_W + \frac{\Sigma W dx.}{\Sigma W}$$

Where,

$A_W$ = Assumed weighted arithmetic mean

dx = Deviations taken from assumed weighted mean

ΣWdx = Sum of the products of weights and their respective deviations taken from the assumed weighted arithmetic mean.

ΣW = Sum of weights.

**Illustration - XII**

While constructing the cost of living index numbers, the following group index numbers are found. Find out the wieghted arithmetic mean of index numbers.

| Group | Index numbers | Weights |
|---|---|---|
| Food | 250 | 100 |
| Clothing | 150 | 50 |
| Rent and Rates | 200 | 60 |
| Fuel | 300 | 80 |

**Solution:**

**COMPUTATION OF WEIGHTED AVERAGE OF INDEX NUMBERS**

Let us assume 200 as the mean of the series i.e., $A_W$ = 200

| Group | Index numbers X | Weights W | $(X - A_W)$ dx | wdx |
|---|---|---|---|---|
| Food | 250 | 100 | 50 | 5,000 |
| Clothing | 150 | 50 | −50 | −2,500 |
| Rent and Rates | 200 | 60 | 0 | 0 |
| Fuel | 300 | 80 | 100 | 8,000 |
| | | $\Sigma W = 290$ | | $\Sigma W d_X = 10{,}500$ |

$$\overline{X}_W = A_W + \frac{\Sigma W d_X}{\Sigma W}$$

Here,

$$A_W = 200, \qquad \Sigma W dx = 10{,}500, \qquad \Sigma W = 290$$

Substituting the values in the formula,

$$\overline{X}_W = 200 + \frac{10{,}500}{290}$$

$$= 200 + 36.20$$

$$= 236.20$$

Hence the weighted mean of index numbers = 236.20

**Utility of weighted arithmetic mean**

Weighted arithmetic mean is useful in case where the items of a variable do not carry equal importance. Weighted arithmetic mean is extremely useful in respect of the following:

i) Constructing the index numbers;

ii) Comparing the items of two or more related groups, where the total number of observations differ ; and

iii) Computation of the standardised birth and death rates.

## 12.6 PROPERTIES OF ARITHMETIC MEAN

The following are some of the important properties of arithmetic mean:

i) The sum of deviations of all the values of variable taken from its arithmetic mean is always equal to zero (taking into consideration the plus and minus signs). This property makes the arithmetic mean a 'point of balance'. i.e., when the deviations are taken from the arithmetic mean, the sum of the positive deviations is equal to the sum of the negative deviations. This is clear from the following example.

| Wages (in Rs.) X | $(X - \overline{X})$ dx |
|---|---|
| 100 | −200 |
| 200 | −100 |
| 300 | 0 |
| 400 | 100 |
| 500 | 200 |
| ΣX 1500 | − 300 + 300 = 0 |
| $\overline{X} = 300$ | $\Sigma(X - \overline{X}) = 0$ |

Here,

$$\Sigma X = 1500, \quad N = 5,$$

$$\overline{X} = \frac{\Sigma X}{N} = \frac{1,500}{5} = 300$$

ii) The sum of squares of deviations of the observations taken from the arithmetic mean is minimum, and is less than the sum of the square of deviations of the items from any other value.

Symbolically,

$\Sigma(X - \overline{X})^2$ is minimum.

**Note the following example :**

| X | $(X - \overline{X})$ d | $(X - \overline{X})^2$ $d^2$ |
|---|---|---|
| 4 | −2 | 4 |
| 5 | −1 | 1 |
| 6 | 0 | 0 |
| 7 | 1 | 1 |
| 8 | 2 | 4 |
| $\Sigma X = 30$ | 0 | $\Sigma d^2 = 10$ |

Since $\Sigma X = 30$, N = 5; $\overline{X} = \frac{\Sigma X}{N} = \frac{30}{5} = 6$

Thus, $d^2$ is the sum of squares of deviations taken from the arithmetic mean (i.e., $\overline{X} = 6$).

In the above example, the sum of squared deviations are 10, it will be greater than 10 if the deviations are taken from any value other than the value of mean i.e.,6.

Let us assume that the deviations are taken from 5.

| | (X−A) d | $(X-A)^2$ $dx^2$ |
|---|---|---|
| 4 | −1 | 1 |
| 5 | 0 | 0 |
| 6 | 1 | 1 |
| 7 | 2 | 4 |
| 8 | 3 | 9 |
| ΣX = 30 | Σd = 5 | $\Sigma d^2$ = 15 |

Thus the sum of squares of deviations is greater than 10.

iii) Arithmetic mean is susceptible to further mathematical treatment. Since $\overline{X} = \frac{\Sigma X}{N}$ if any two values are given other values can be obtained by substituting the values in the formula. Thus the mathematical relations of this property in this connection can be:

$$\overline{X} = \frac{\Sigma X}{N}, \qquad \Sigma X = \overline{X}N, \qquad N = \frac{\Sigma X}{\overline{X}}$$

iv) If the arithmetic mean and number of observations of two or more related groups are known, the combined mean of all the groups can be computed by combining all the series.

v) Standard error of the arithmetic mean is always less than the standard error of any other measure of central tendency such as Median, Mode, Geometric and Harmonic means.

## 12.7 MERITS AND LIMITATIONS OF ARITHMETIC MEAN

**Merits**

i) Computation of arithmetic mean is easy as it does not involve any tedious mathematical claculations. Simple mathematical knowledge of addition, multiplication and division is enough for computing the arithmetic mean.

ii) The arithmetic mean is a computed mean. Hence, it is amenable for further algebraic treatment.

iii) Since, the methods of computation of arithmetic mean are defined scientifically, any one who computes the arithmetic mean by any method for a given series will get the same answer.

iv) The arithmetic mean is typical and representative in the sense that it is the centre of gravity and a point of balance of the values on either side of it.

v) The arithmetic mean is a stable statistic and does not vary too much when repeated samples are taken from the same population.

vi) Arithmetic mean is computed on the basis of all the values of a variable. As such, it is considered to be a true representative of all the values of the variable unlike positional averages like median and quartiles.

**Limitations**

i) Since the arithmetic mean is computed on the basis of each and every item of the series, the extreme values of the data will unduly influence the value of arithmetic mean. In such cases the value of arithmetic mean cannot be a true representative of the data. For example, if in a class the marks secured by 5 students are 80, 75, 90, 85 and 5 the average marks would be $80 + 75 + 90 + 85 + 5 = \frac{335}{5} = 67$. All the values of the variables are more or less consistent except one value i.e.5. Hence 5 is considered to be extreme value and it is because of this value the arithmetic mean is pulled down to 67.

ii) Arithmetic mean computed for the data which contain open-end classes cannot be a typical value of the series as the values of lower limit of the first class and upper limit of the last class are determined arbitrarily.

iii) Very often interpretations of arithmetic mean may mislead the readers in the absence of the other details of the data. Note the following example:

| Year | No. of passes | |
|---|---|---|
| | College A | College B |
| 1980 | 1,500 | 5,000 |
| 1981 | 2,000 | 4,000 |
| 1982 | 2,500 | 2,500 |
| 1983 | 4,000 | 2,000 |
| 1984 | 5,000 | 1,500 |
| Total | 15,000 | 15,000 |



In the above example, the arithmetic mean of passes in both the colleges is 3000. On the basis of the original data, while college A is showing an improving performance, College B is showing a declining performance. But this has not been revealed by the arithmetic mean. In the words of Horace Secrist, If an average is taken as a substitute for detail, then the arithmetic mean inspite of the simplicity and care of calculation has little to recommend where series are non-homogeneous.

iv) Arithmetic mean cannot be located by inspection or by observation.

v) The computed arithmetic mean is not a typical value because such value is quite often altogether different from that of the values given in the series. For example, the average of 4, 8 and 18 is 10. Which does not occur in the observations. Despite the above limitations arithmetic mean is popularly used as it is simple to calculate and easy to understand.

## 12.8 SUMMING UP

Arithmetic mean is the most widely used measure of central tendency. It is the quotient that results when the sum of all items in the series is divided by the number of items. It can be computed by direct, indirect and step-deviation methods. The chief merits of arithmetic mean are that, it is easy to calculate, amenable to further algebraic treatment and takes all the items of the distribution into account. However, the arithmetic mean is influenced by the extreme values in the series and therefore, some times, there is scope for misleading interpretations. Weighted arithmetic mean is the modified form of arithmetic mean and recognises the relative importance of the items. It is normally applied in the computation of index numbers, birth and death rates.

## 12.9 CHECK YOUR PROGRESS: MODEL ANSWERS

This problem can be done either by direct method or short-cut method. Whatever the method you employ the answer is 67.81. Either of the following formulae can be applied:

Direct method = $\overline{X} = \frac{\Sigma fX}{N}$

Short-cut method $= \overline{X} = A + \frac{\Sigma fdx}{N}$

## 12.10 MODEL EXAMINATION QUESTIONS

**A. Short Questions**

1. Define arithmetic mean?
2. What is meant by weighted arithmetic mean?
3. What are the properties of arithmetic mean?
4. Explain the utility of weighted arithmetic mean in statistical analysis?
5. Explain the merits and limitations of arithmetic mean
6. Explain the procedure to compute arithmetic mean in a continuous series.
7. What do you mean by combined mean? How do you compute it?

**B. Essay Questions**

8. Distinguish between simple arithmetic mean and weighted arithmetic mean. Is weighted mean better than simple Mean? If so, why?

**EXERCISES**

9. Following data relates to the monthly income of 12 families in a town. Calculate arithmetic mean.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Monthly Income | 200, | 180, | 160, | 240, | 190, | 170, | 220, |
| (in Rs.) | 205 | 100, | 180, | 210, | 105 | | |

(Ans: 180)

10. Following are the marks obtained by 10 students in statistics.

Marks; 40, 37, 50, 42, 26, 34, 38, 42, 35, 45

Calculate the mean marks of the class.

(Ans: 38.9)

11. The data given below relates to the marks obtained by 60 students of a class. Find out arithmetic mean.

| Marks | No. of students |
|---|---|
| 30 | 5 |
| 35 | 10 |
| 46 | 20 |
| 42 | 15 |
| 38 | 6 |
| 50 | 4 |

(Ans : 41.3)

12. From the following data, calculate average height of the students.

| Height: (in feet) | 5 | 5.2 | 5.4 | 5.6 | 5.8 | 6 |
|---|---|---|---|---|---|---|
| No. of: students : | 6 | 10 | 20 | 30 | 15 | 8 |

(Ans : 5.54)

13. The following data shows the marks secured by students in a class. Compute arithmetic mean.

| Marks : | 30-40 | 40-50 | 50-60 | 60-70 | 70-80 |
|---|---|---|---|---|---|
| No.of Students : | 20 | 32 | 18 | 12 | 8 |

(Ans : 50.11)

14. The data given below relates to the weekly wages of labourers in a factory.

| Weekly wages (Rs.) : | 10-12 | 12-14 | 14-16 | 16-18 | 18-20 |
|---|---|---|---|---|---|
| No. of Labourers : | 24 | 18 | 15 | 8 | 7 |

Ascertain mean weekly wage.

(Ans: 13.78)

15. The marks obtained by 60 students are given below.

| Marks : | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 |
|---|---|---|---|---|---|---|
| No. of Students: | 3 | 6 | 10 | 21 | 12 | 8 |

Calculate arithmetic mean :

(Ans : 34.5)

16. The following data, relates to the weekly wages of workers of a Factory. Find out arithmetic mean :

| Wages : (in Rs.) | 20-29 | 30-39 | 40-49 | 50-59 | 60-69 | 70-79 |
|---|---|---|---|---|---|---|
| No. of workers | 15 | 25 | 28 | 22 | 18 | 12 |

(Ans : 47.75)

17. From the following data, calculate weighted arithmetic mean of the price indices.

| Commodity | Weights | Index Number |
|---|---|---|
| A | 42 | 250 |
| B | 14 | 200 |
| C | 22 | 135 |
| D | 30 | 325 |
| E | 16 | 400 |
| F | 10 | 214 |

(Ans : 257.91)

## 12.11 RECOMMENDED BOOKS


1. Gupta, S.P : "Statistical Methods", Sultan Chand & Company, New Delhi.
2. Gupta, B.N : "Statistics", Sahitya Bhavan, Agra.
3. Gupta, S.C. : "Fundamentals of statistics, Himalaya Publishing House, Bomoay.
4. Simpson and Kafka : "Basic Statistics", Oxford and I.B.H. Publishing Company, Calcutta.


## 12.12 GLOSSARY

1. Arithmetic Mean : It is the value obtained by adding the values of all the items and dividing the total by number of items.
2. Combined Mean : It is ascertained by multiplying the means of two series (i.e., $\overline{X}_1$ and $\overline{X}_2$) with their respective sizes (i.e., $N_1$ and $N_2$) and dividing the total by the sum of sizes (i.e. $N_1 + N_2$).
3. Weighted Arithmetic Mean : It is computed by assigning weights to each item according to their importance in the distribution. These weights are multiplied with their values and the total is divided by the sum of weights.

## Unit-13: MEDIAN AND QUARTILES

**Contents**



## 13.0 AIMS AND OBJECTIVES

This Unit aims at explaining the meaning and computation of median, quartiles, deciles and percentiles, and also the merits and limitations of median.

After going through this unit, you should be able to:

- describe the meaning of median
- work out the problems of median
- identify the properties of the median
- list out the merits and limitations of median
- explain the meaning of quartiles, deciles, and percentiles
- compute the problems of quartiles, deciles, and percentiles.

## 13.1 INTRODUCTION

In Unit 12, we dealt with arithmetic mean. After arithmetic mean, median is the most important measure of central tendencies. Median is a positional average. It can be described as the middle item when the values are arranged in the ascending order or descending order of magnitude.

Quartiles are the statistical measures which divide the total frequency into four quarters. We shall see both the theoretical and the practical aspects of median and quartiles.

## 13.2 MEANING OF MEDIAN

Computation of arithmetic mean is based on each and every item of the series. Due to this, the value of mean is influenced by the extreme values of the distrtibution. Thus, the computed value of this mean may mislead the reader. In view of this an alternative average is needed to represent the distribution. Median is one such popular measure of central tendency.

Median is the middle value of the distribution that divides the series into two parts in such a way that the individual values of one-half of the distribution are either equal to or smaller to the size of median value. And the individual values of the other half of the distribution are either equal to or greater than the size of the median value. Some of the important definitions of median are given below : In the words of Taro Yamane median of a frequency distribution is "a value that divides the frequency distribution into two equal parts". Yule and Kendall defined median as "the middle most or central value of the variable when the values are arranged in order of magnitude, or as the value is such that greater and smaller values occur with equal frequency". According to L.R.Connor "median is that value of the variable which divides the group into two equal parts, one part comprising all values greater, and the other, all values less than median.

## 13.3 COMPUTATION OF MEDIAN

Median can be computed for individual, discrete and continuous series.

**Individual Series**

In the case of individual series the following steps are needed to compute the Median.

i) Arrange the values of the distribution either in ascending order or in descending order.

ii) Apply the formula and obtain the value of median.

$$\text{Med} = \text{Size of } \left(\frac{N+1}{2}\right)\text{th item}$$

Where, Med = Median

N = Number of observations.

**Illustration-I**

The heights of students (in inches) are given below:

59, 68, 54, 61, 58, 62, 57, 61, 64.

Find out the median height of students.

**Solution :**

**Computation of Median**

To calculate median, the data are arranged in ascending order and written as below :

54, 57, 58, 59, 61, 61, 62, 64, 68

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

Since, N= 9.

Med= Size of $\left(\frac{9+1}{2}\right)$th item

= Size of 5th item

Size of 5th item in the distribution is 61. Hence, the median of the students is 61 inches.

**Illustration-II**

Given below are the marks secured by 12 students in statistics. Find out median marks secured by the students in Statistics.

Marks in Statistics : 64, 48, 46, 51, 55, 54, 58, 60, 56, 47. 52, 49

**Solution :**

**Computation of Median**

In order to calculate the median marks, the data are arranged in ascending order.

Marks in Statistics. 46, 47, 48, 49, 51, 52, 54, 55, 56, 58, 60, 64

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

Since, N = 12,

Med = Size of $\left(\frac{12+1}{2}\right)$ th item

= Size of 6.5th item

Since median is represented by size of 6.5th item, the value of Median would be the average of the value of adjoining items i.e., 6th and 7th items. Thus the size of 6.5th item

$$= \frac{value\ of\ 6th\ item + value\ of\ 7th\ item}{2}$$

$$\text{Med} = \frac{52+54}{2}$$

$$= \frac{106}{2}$$

$$= 53$$

Hence, median Marks of students in Statistics = 53.

**Discrete Series**

The formula used for computing the median value of individual series is also used to compute the median value in case of discrete series. However, the frequencies of the items are cumulated, in the case of discrete series. The following steps are required to compute the median.

i) Arrange the data in ascending or descending order

ii) Find out cumulative frequencies by adding the given frequency with the preceding frequency.

iii) Apply the formula and obtain the size of the median

iv) Locate the size of the median in the cumulative frequency column and find the total which is either equal to size of $(\frac{N+1}{2})$th item or next higher than that.

v) Determine the value of size of $(\frac{N+1}{2})$th item and obtain the median.

**Illustration-III**

Compute the median marks at a test in English from the following information :

| | | |
|---|---|---|
| Marks | : | 10, 20, 30, 40, 50, 60, 70, 80 |
| No.of Students | : | 6, 9, 15, 20, 25, 10 , 8, 7 |

**Solution :**

**COMPUTATION OF MEDIAN MARKS OF STUDENTS IN ENGLISH**

| Marks X | No.of Students f | Cumulative frequency cf |
|---|---|---|
| 10 | 6 | 6 |
| 20 | 9 | 15 |
| 30 | 15 | 30 |
| 40 | 20 | 50 |
| 50 | 25 | 75 |
| 60 | 10 | 85 |
| 70 | 8 | 93 |
| 80 | 7 | 100 |

Med = Size of $(\frac{N+1}{2})$th item

Since N = 100,

Med = Size of $(\frac{100+1}{2})$th item

= Size of 50.5th item

Size of 50.5th item lies in the cumulative frequency 75 and its corresponding value is 50. Hence, median marks of the students are 50.

**CONTINUOUS SERIES**

In the continuous series, the median is determined by using size of $\frac{N}{2}$ th item and median value is obtained with the help of the following steps:

i) Find out cumulative frequencies.

ii) Determine the median class by using size of N/2th item

iii) Apply the following formula and obtain the median value.

$$\text{Med} = L + \frac{\frac{N}{2} - cf}{f} \times i$$

Where,

Med = Median

L = Lower limit of the median class

c.f = Cumulative frequency of the class preceding the median class

i = Class interval of the median class

f = Frequency of the median class

$\frac{N}{2}$ = Size of the median class

**Illustration - IV**

The following frequency distributioin relates to the age of workers of a factory. Find out the median age of workers in the factory.

| Age in Years | Numbers of workers |
|---|---|
| 20 - 25 | 10 |
| 25 - 30 | 15 |
| 30 - 35 | 20 |
| 35 - 40 | 60 |
| 40 - 45 | 42 |
| 45 - 50 | 24 |
| 50 - 55 | 18 |
| 55 - 60 | 11 |

**Solution :**

**COMPUTATION OF MEDIAN AGE OF WORKERS OF A FACTORY**

| Age (in years) X | Number of workers f | Cumulative frequency cf |
|---|---|---|
| 20 - 25 | 10 | 10 |
| 25 - 30 | 15 | 25 |
| 30 - 35 | 20 | 45 |
| 35 - 40 | 60 | 105 |
| 40 - 45 | 42 | 147 |
| 45 - 50 | 24 | 171 |
| 50 - 55 | 18 | 189 |
| 55 -60 | 11 | 200 |

$$\text{Median class} = \text{size of } \left(\frac{N}{2}\right)\text{th item}$$

Here,

$$N = 200$$

$$\text{median class} = \text{size of } \left(\frac{200}{2}\right)\text{th item}$$

$$\text{median class} = \text{size of 100th item}$$

size of 100th item lies in 35-40 class

$$\text{median} = L + \frac{\frac{N}{2} - cf}{f} \times i$$

Here, $L = 35$, $\frac{N}{2} = 100$, c f $= 45$, i $= 5$, f $= 60$

Substituting the values in the formula,

$$\text{Med} = 35 + \frac{100-45}{60} \times 5$$

$$= 35 + \frac{55}{60} \times 5$$

$$= 35 + 0.916 \times 5$$

$$= 35 + 4.58$$

$$\text{Med} = 39.58$$

Alternatively the same illsutration will be worked out by applying the following formula to find out Median.

$$\text{Med} = U - \frac{\frac{N}{2} - cf}{f} \times i$$

Where,

U = Upper limit of the median class

cf = Cumulative Frequency of the class preceding the median class starting from largest class values.

f = Frequency of the median class

i = class interval

$\frac{N}{2}$ = Size of the median class

Further, the data given in the illustration - IV have to be rearranged in the following manner.

**COMPUTATION OF MEDIAN AGE OF WORKERS OF A FACTORY**

| Age (in years) | Number of workers f | Cumulative frequency cf |
|---|---|---|
| 20 - 25 | 10 | 200 |
| 25 - 30 | 15 | 190 |
| 30 - 35 | 20 | 175 |
| 35 - 40 | 60 | 155 |
| 40 - 45 | 42 | 95 |
| 45 - 50 | 24 | 53 |
| 50 - 55 | 18 | 29 |
| 55 - 60 | 11 | 11 |

$$\text{Median class} = \text{Size of } \frac{N}{2}\text{th item}$$

$$= \text{Size of } \frac{200}{2} \text{ th item}$$

$$= \text{Size of 100th item}$$

Size of 100th item lies in 35 - 40 class

Hence,

$$\text{Median} = U - \frac{\frac{N}{2} - cf}{f} \times i$$

Here,

$$U = 40, \quad \frac{N}{2} = 100, \text{ cf} = 95, \text{ i} = 5, \text{ f} = 60$$

Substituting the value in the formula.

$$\text{Med} = 40 - \frac{100-95}{60} \times 5$$

$$= 40 - \frac{5}{60} \times 5$$

$$= 40 - \frac{25}{60}$$

$$= 40 - 0.416$$

$$\therefore \quad \text{Median} = 39.58$$

**Illustration:V**

Compute median from the following data:

| Mid value | Frequency |
|---|---|
| 125 | 5 |
| 175 | 8 |
| 225 | 10 |
| 275 | 18 |
| 325 | 25 |
| 375 | 35 |
| 425 | 20 |
| 475 | 15 |
| 525 | 11 |
| 575 | 3 |

**Solution :**

**Computation of Median**

Since mid-values are given, the lower and upper limits of each class have to be ascertained. The difference between the two mid-values is taken as class interval. Hence, the class interval in this illsutration would be 50.

Half of this i.e., 25 is deducted from each mid-value to find out the lower limit of the class and 25 is added to all mid values to obtain the upper limits of the classes. After determining the various class limits the given data are rearranged as below:

| Class Interval | Frequency f | Cumulative frequency cf |
|---|---|---|
| 100 - 150 | 5 | 5 |
| 150 - 200 | 8 | 13 |
| 200 - 250 | 10 | 23 |
| 250 - 300 | 18 | 41 |
| 300 - 350 | 25 | 66 |
| 350 - 400 | 35 | 101 |
| 400 - 450 | 20 | 121 |
| 450 - 500 | 15 | 136 |
| 500 - 550 | 11 | 147 |
| 550 - 600 | 3 | 150 |

Median class = Size of N/2th item; Here, N = 150

Median Class = Size of 150/2th item = Size of 75th item

Size of 75th item lies in 350 - 400 class:

Applying the formula,

$$\text{Median} = L + \frac{\frac{N}{2} - cf}{f} \times i$$

Here,

$$L = 350, \frac{N}{2} = 75, cf = 66, f = 35, i = 50$$

Substituting the values in the formula,

$$\text{median} = 350 + \frac{75-66}{35} \times 50$$

$$= 350 + \frac{9}{35} \times 50$$

$$= 350 + 0.258 \times 50$$

$$= 350 + 12.9$$

$$\therefore \quad \text{Median} = 362.9$$

**Illustration:VI**

Find out median from the following data:

| Sales (Rs.in lakhs) | 10-19, | 20-29, | 30-39, | 40-49, | 50-59, | 60-69, | 70-79, |
|---|---|---|---|---|---|---|---|
| No. of firms | 4 | 8 | 18 | 25 | 30 | 10 | 5 |

**Solution :**

**Computation of Median**

In order to calculate median value, the data given in inclusive series are covered into exclusive series and rearranged as below:

| Sales (in Rs.lakhs) | No. of firms f | Cumulative Frequency cf |
|---|---|---|
| 9.5 - 19.5 | 4 | 4 |
| 19.5 - 29.5 | 8 | 12 |
| 29.5 - 39.5 | 18 | 30 |
| 39.5 - 49.5 | 25 | 55 |
| 49.5 - 59.5 | 30 | 85 |
| 59.5 - 69.5 | 10 | 95 |
| 69.5 - 79.5 | 5 | 100 |

Median Class = Size of $\frac{N}{2}$th item; Here, N = 100

Median Class = Size of 100/2th item = Size of 50th item

Size of 50th item lies in 39.5 - 49.5 Class

Applying the formula,

$$\text{Med} = L + \frac{\frac{N}{2} - cf}{f} \times i$$

Here,

$L = 39.5, \quad \frac{N}{2} = 50, \quad cf = 30, \quad f = 25, \quad i = 10$

Substituting the values in the formula,

$$\text{Med} = 39.5 + \frac{50-30}{25} \times 10$$

$$= 39.5 + \frac{20}{25} \times 10$$

$$= 39.5 + 0.8 \times 10$$

$$= 39.5 + 8 = 47.5$$

$\therefore$ Median = 47.5

**Calculation of median - When class intervals are unequal**

When class intervals are unequal,the computation of median does not require any adjustment. The formula used earlier is used to find out the Median.

**Illustration VII:**

Calculate median from the following frequency distribution.

Class : 0-5, 5-10, 10-20, 20-25, 25-30, 30-40

Frequency: 5, 8, 20, 15, 16, 10

**Solution:**

**COMPUTATION OF MEDIAN**

| Class (x) | Frequency (f) | Cumulative frequency (cf) |
|---|---|---|
| 0-5 | 5 | 5 |
| 5-10 | 8 | 13 |
| 10-20 | 20 | 33 |
| 20-25 | 15 | 48 |
| 25-30 | 16 | 64 |
| 30-40 | 10 | 74 |

Median class = Size of $\left(\frac{N}{2}\right)$th item

Median Class = Size of $\frac{74}{2}$th item

Median class = Size of 37th item

Size of 37th item lies in 20 - 25 class.

$$\text{Med} = L + \frac{\frac{N}{2} - cf}{f} \times i$$

Here,

$$L = 20, \quad \frac{N}{2} = 37, \quad cf = 33, \quad i = 5, \quad f = 15$$

Substituting the values in the formula,

$$\text{Med} = 20 + \frac{37-33}{15} \times 5$$

$$= 20 + \frac{4}{15} \times 5$$

$$= 20 + 0.267 \times 5$$

$$= 20 + 1.33$$

Median = 21.33

**Check your progress - 1**

Calculate the median for the data given below

| Wages (Rs.) | 0-5 | 5-10 | 10-15 | 15-20 | 20-25 |
|---|---|---|---|---|---|
| No. of workers | 8 | 15 | 20 | 30 | 25 |

## GRAPHIC LOCATION OF MEDIAN

In order to locate the value of median graphically either of the following methods is used:

a) Less than ogive curve method.

b) Less than and more than ogive curve method

**a)Location of median by 'less than ogive curve' method:**

The following procedure is adopted to locate median graphically by 'Less than ogive curve' method :

i ) Arrange the values by less than method and locate them on 'X' axis.

ii) Find out the cumulative frequencies and locate them on 'Y' axis.

iii) Draw an ogive cruve by less than method.

iv) Locate the size of N/2th item on Y - axis and through this point draw a horizontal line to X - axis.

v) Draw a perpendicular line on X - axis through the point where the ogive curve and the horizontal line to X - axis intersect.

vi) Locate the value of median on X - axis. the value where the perpendicular line touches the X - axis is taken as the value of median.

**Illustration - VIII.**

Locate median graphically for the following frequency distribution.

| Weights (Kgs) | Number of persons |
|---|---|
| 30 - 35 | 8 |
| 35 - 40 | 10 |
| 40 - 45 | 25 |
| 45 - 50 | 30 |
| 50 - 55 | 20 |
| 55 - 60 | 14 |
| 60 - 65 | 8 |
| 65 - 70 | 5 |

**Solution :** While locating the median graphically, the frequencies are to be cumulated.

| Weights (Kgs) | No.of persons | Cumulative frequency |
|---|---|---|
| Less than 35 | 8 | 8 |
| " 40 | 10 | 18 |
| " 45 | 25 | 43 |
| " 50 | 30 | 73 |
| " 55 | 20 | 93 |
| " 60 | 14 | 107 |
| " 65 | 8 | 115 |
| " 70 | 5 | 120 |

**Location of Median Graphically:**

Median = Size of N/2th item

Here,

$N = 120$

Med = size of $\frac{120}{2}$th item

Med = Size of 60th item

For locating median on the graph (Shown in fig:13.1), draw horizontal line parallel to 'X axis from 60 on 'Y' axis till it intersect the ogive curve. From the intersecting point ($M_1$) draw a perpendicular line parallel to 'Y' axis till it touches 'X' axis. This intersecting point ($M_2$) represents the value of median i.e., Median weight = 47.83 Kg.

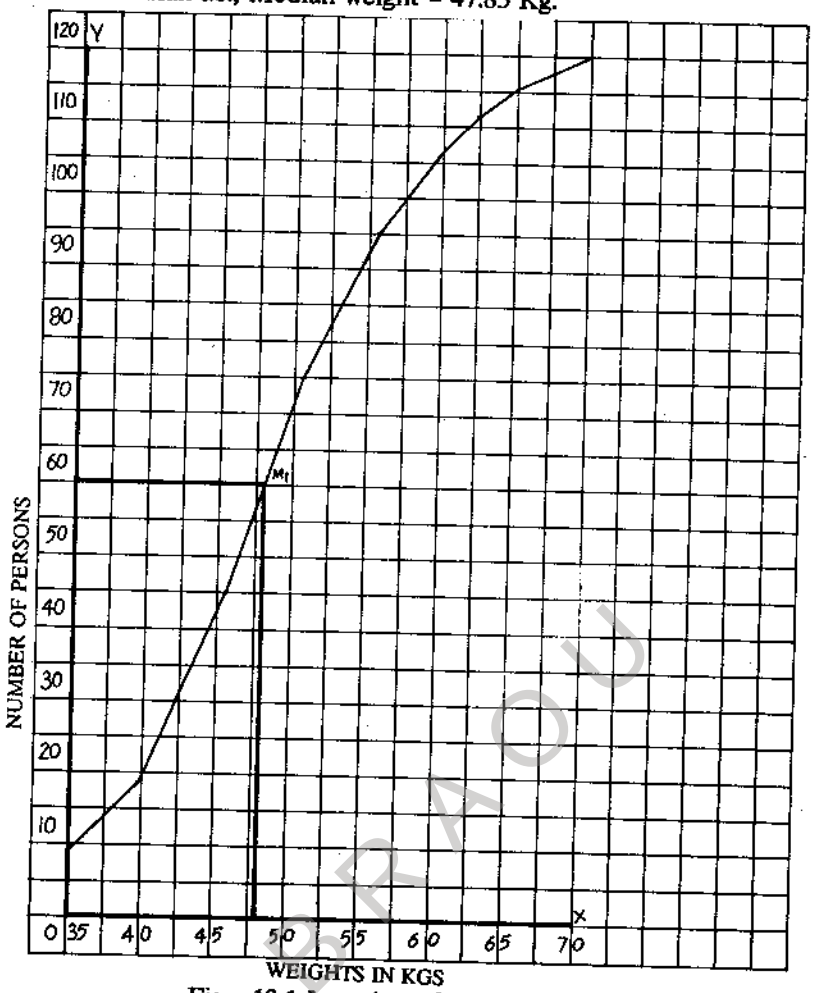



Fig : 13.1 Location of Median.

**Index :**

OY : 2 cms = 10 persons

OX : 2 cms = 5 Kgs.

Median Weight = $M_2$ = 47.83 Kgs.

**Location of Median by "less than" and more than "ogive curves":**

In order to locate median graphically lessd than and more than ogive curves, the following procedure is adopted.

i) Arrange the values as per less than method and locate them on X-axis.

ii) Arrange the cumulative frequencies in ascending order and obtain less than cumulative

frequencies and locate them on "Y" axis.

iii) Draw an ogive curve by less than method.

iv) Arrange the values by more than method and locate them on "X"axis.

v) Arrange the cumulative frequency in descending order and obtain more than cumulative frequency and locate them on "Y" axis.

vi) Draw an Ogive by more than method.

vii) Draw a perpendicular line on X-axis through the point where the two Ogive curves intersect.

viii) Locate the median at the point where the perpendicular line touches X - axis.

**Illustration - IX**

Find out median by using two Ogive curves for the following distribution.

| Monthly Savings(in Rs.) | Number of employees |
|---|---|
| 50 - 100 | 6 |
| 100 - 150 | 10 |
| 150 - 200 | 20 |
| 200 - 250 | 25 |
| 250 - 300 | 16 |
| 300 - 350 | 12 |
| 350 - 400 | 5 |

**Solution :** To locate median graphically, the frequencies are cumulated and arranged in ascending order and descending order.

| Monthly Savings (in Rs.) | No. of employees (Ascending Cumulative Frequency) | Monthly Savings (in Rs.) | No. of employees (Descending Cumulative Frequency) |
|---|---|---|---|
| Less than 100 | 6 | More than 50 | 94 |
| " 150 | 16 | " 100 | 88 |
| " 200 | 36 | " 150 | 78 |
| " 250 | 61 | " 200 | 58 |
| " 300 | 77 | " 250 | 33 |
| " 350 | 89 | " 300 | 17 |
| " 400 | 94 | " 350 | 5 |
| | | " 400 | 0 |

To locate median graphically, "more than" and "less than" ogive curves are plotted on the graph in fig 13.2. From the intersecting point ($M_1$) a perpendicular line is drawn which intersects "X" axis at ($M_2$). This intersecting point represents the value of median.

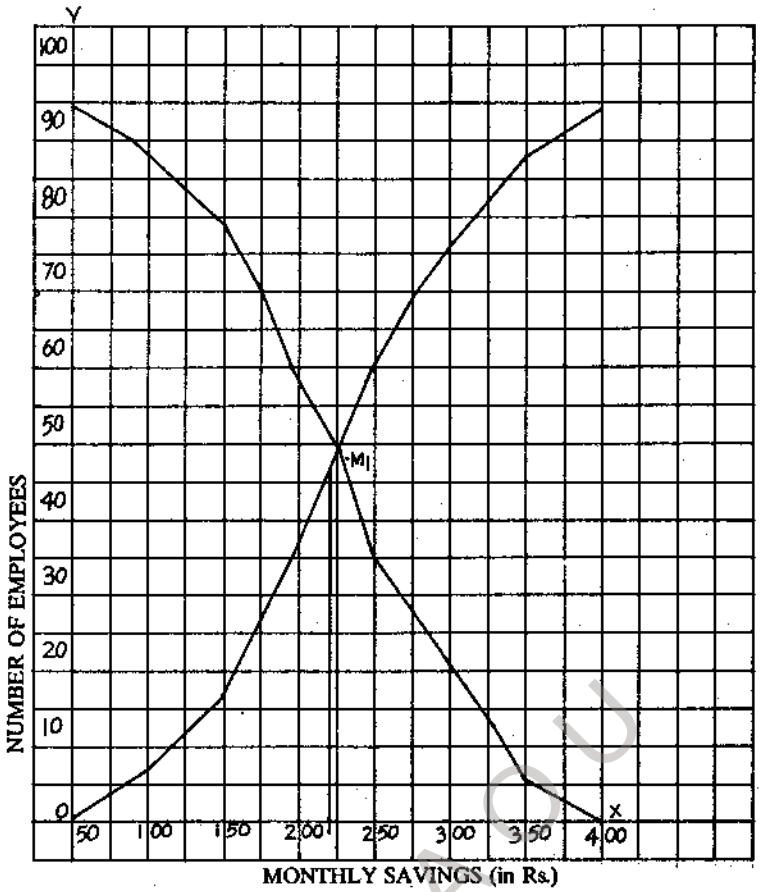



Fig : 13.2 Location of median

Index : OX : 2 cms = 10 persons  OY : 2 cms = Rs.25/-

Median Savings = $M_2$ = Rs.222/-

## 13.4 PROPERTIES OF MEDIAN

i) Median is a positional average.

ii) The value of the median is not influenced by the extreme values of the distribution.

iii) The sum of deviations of the values of a series taken from its median (ignoring the plus and minus signs) willl be less than the sum of deviations taken from any other value. Note the following example :

The median of 5, 6, 7, 8, 9, 10 and 11 is 8. The deviation from 8 ignoring the plus and minus signs are : 3, 2, 1, 0, 1, 2, 3, and their total is 12. If the deviations are taken from 9 the deviations are 4, 3, 2, 1, 0, 1, 2, and their total is 13.



221

## 13.5 MERITS AND LIMITATIONS OF MEDIAN

**Merits**

i) It is easy to compute median. It can also be located by observation.

ii) Median is not influenced by the extreme values of the distribution, where the data are unevenly distributed median is a better measure than mean as the calculation of mean is affected by the extreme items of the series.

iii) Median is useful when the data are classified in open-end classes as median refers to the middle value of the series.

iv) When the values of a series are not capable of direct quantitative measurement, median will be more useful measure of central tendency.

v) Unlike arithmetic mean, median can be calculated from the data which are incomplete and irregular.

vi) The value of median can be located graphically also, where the value of arithmetic mean cannot be determined graphically.

vii) Median is used for computing further statistical measures such as dispersion and skewness.

**Limitations**

1. Since median is a positional average, its computation is not based on each and every value of the series. As such median cannot be considered the true representative of the given data.
2. Unlike other averages, computation of the median needs the arrangement of data either in descending order or in ascending order. Thus its computation involves additional work.
3. Median is not amenable for further mathematical treatment.
4. When the data are given in even numbers, the exact value of median cannot be known, as median is based on the location of values.



## 13.6 MEANING OF QUARTILES, DECILES AND PERCENTILES

While median divides the given series into two equal parts, the first quartile divides the first part of the series into two equal parts and third quartile divides the other part of the series into two equal parts. Thus the total number of quartiles in a series is three. The first quartile also called lower quartile, is denoted by $Q_1$. Whereas third quartile, also called upper quartile, is denoted by $Q_3$ The second quartile is denoed by $Q_2$ which is equal to median.

Deciles divide the given distribution into ten equal parts consisting of nine diciles such $D_1$. $D_2$. $D_3$.....$D_9$ The value of $D_5$ is equal to the value of median.

Percentiles devide the given distribution into an equal parts consisiting of 99 percentiles such as $P_1$, $P_2$,$P_3$,.....$P_{99}$.

## 13.7 COMPUTATION OF QUARTILES, DECILES AND PERCENTILES

The values of quartiles, deciles and percentiles can be computed by adopting the same procedure which is used for computing median.

**Individual Series:** The value of quartiles, deciles and percentiles for an individual series is computed with the help of the following steps.

i) Arrange the given data either in ascending order or in descending order.

ii) Apply the formula to obtain Q, D and P.

Lower Quartile = $Q_1$ = Size of $\frac{N+1}{4}$ th item

Upper Quartile = $Q_3$ = Size of $3\frac{(N+1)}{4}$th item

First Decile = $D_1$ = Size of $\frac{N+1}{10}$ th item

First Percentile = $P_1$ = Size of $\frac{N+1}{100}$th item

**Illustration - X**

Compute the value of lower quartile, upper quartile, first Decile and 50th percentile from the following data.

40, 47, 45, 49, 48, 51, 54, 50, 53, 55, 57, 56, 60, 58, 65, 64, 63, 70, 69

**Solution:** In order to calculate quartiles and other measures, the data is arranged in ascending order.

40, 45, 47, 48, 49, 50, 51, 53, 54, 55, 56, 57, 58, 60, 63, 64, 65, 69, 70.

**Computation of Lower quartile**

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

Since, N = 19

$$Q_1 = \text{Size of } \frac{19+1}{4} \text{ th item}$$

$$Q_1 = \text{Size of 5th item}$$

Size of 5th item in the distribution is 49.

Hence,

$$Q_1 = 49$$

**Computation of Upper Quartile**

$$Q_3 = \text{Size of } 3(\frac{N+1}{4})\text{th item}$$

Since, N = 19

$$Q_3 = \text{Size of } 3\left(\frac{19+1}{4}\right) \text{ th item}$$

= Size of 15th item

Size of 15th item in the distribution is 63

Hence,

$$Q_3 = 63$$

**Computation of 1st Decile**

$$D_1 = \text{Size of } \frac{N+1}{10} \text{th item}$$

Here, N = 19

$$D_1 = \text{Size of } \frac{19+1}{10} \text{ th item}$$

= Size of 2nd item

Size of 2nd item is 45

Hence,

$$D_1 = 45$$

**Computation of 50th Percentile**

$$P_{50} = \text{Size of } 50\left(\frac{N+1}{100}\right) \text{ th item}$$

Here, N = 19

$$P_{50} = \text{Size of } 50\left(\frac{19+1}{100}\right) \text{ th item}$$

Size of 10th item is 55

$$P_{50} = 55$$

**Descrete Series.:** To compute the values of quartiles, deciles and percentiles, the formula used in the case of individual series is used. But the frequencies are to be cumulated.

Thus,

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

Where,

$Q_1$ = Lower quartile

N = Cumulative frequency

$$Q_3 = \text{Size of } 3\left(\frac{N+1}{4}\right)\text{th item}$$

Where,

$Q_3$ = Upper quartile

N = Cumulative frequency

$$D_1 = \text{Size of } \frac{N+1}{10}\text{th item}$$

Where,

$D_1$ = First decile

N = Cumulative frequency

$P_1$ = Size of $\frac{N+1}{100}$ th item

Where,

$P_1$ = First percentile

N = Cumulative frequency

**Illustration XI**

Find out Lower Quartile, Upper Quartile, 3rd decile and 20th percentile from the following data:

Earnings per day : (in.Rs) 8, 9, 10, 15, 12, 16, 13, 14, 11

No. of workers: 10, 15, 18, 8, 13, 4, 17, 9, 16

**Solution:**

**Computation of Quartiles, Decile and Percentile.**

| Earning per day (inRs.) | No.of Workers f | Cumulative frequency cf |
|---|---|---|
| 8 | 10 | 10 |
| 9 | 15 | 25 |
| 10 | 18 | 43 |
| 11 | 16 | 59 |
| 12 | 13 | 72 |
| 13 | 17 | 89 |
| 14 | 9 | 98 |
| 15 | 8 | 106 |
| 16 | 4 | 110 |

**Computation of Quartiles**

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

since,

N = 110

$Q_1$ = Size of $(\frac{110+1}{4})$th item

= Size of 27.75th item

Size of 27.75 item lies in the cumulative frequency of 43. Thus, the value of lower quartile is the value corresponding 43rd item i.e., = 10

Hence the lower quartile = 10

$$Q_3 = \text{Size of } 3\left(\frac{N+1}{4}\right) \text{ th item}$$

Since N = 110,

$$Q_3 = \text{size of } 3\left(\frac{110+1}{4}\right) \text{ th item}$$

$$= \text{Size of } 83.25\text{th item}$$

Since the size of 83.25th item lies in the cumulative frequency of 89, the upper Quartile will be the value corresponding to 89th item i.e., 13.

Hence, the upper Quartile = 13.

**Computation of 3rd Decile**

$$D_3 = \text{Size of } 3\left(\frac{N+1}{10}\right)\text{th item}$$

since, N = 110

$$D_3 = \text{size of } 3\left(\frac{110+1}{10}\right)\text{th item}$$

$$= \text{size of } 3(11)\text{th item}$$

$$= \text{size of } 33.3 \text{ rd item}$$

Size of 33.3rd item lies in the cumulative frequency. Hence, the value of third decile is the value corresponding to 43rd item i.e., 10.

**Computing of 20th percentile**

$$P_{20} = \text{Size of } 20\left(\frac{N+1}{100}\right)\text{th item}$$

Since, N = 110,

$$\text{Size of } 20\left(\frac{110+1}{100}\right)\text{th item}$$

$$= \text{size of } 20\ (1.11)\text{th item}$$

$$= \text{size of } 22.2\text{nd item.}$$

Size of 22.2nd item lies in the Cumulative frequency of 25. Hence, the value of 20th percentile is the value corresponding to 25th item i.e. 9.

**Continuous Series:** In continuous series, to compute quartiles, deciles and percentiles the following procedure is adopted.

i) Find out cumulative frequencies.

ii) Determine the quartiles, deciles and percentiles using the following formula.

$$Q_1 = \text{L} + \frac{\frac{N}{4} - cf}{f} \times \text{i}$$

$$Q_3 = \text{L} + \frac{\frac{3N}{4} - cf}{f} \times \text{i}$$

$$D_1 = L + \frac{\frac{N}{10} - cf}{f} \times i$$

$$P_1 = L + \frac{\frac{N}{100} - cf}{f} \times i$$

Where,

$Q_1$ = Lower Quartile
$Q_3$ = Upper Quartile
$D_1$ = First decile
$P_1$ = First percentile

**Illustration - XII**

calcualte quartile, 3rd decile and 50th percentile from the following distribution.

| Income (in.Rs) | Number of persons |
|---|---|
| 100-120 | 6 |
| 120-140 | 10 |
| 140-160 | 18 |
| 160-180 | 30 |
| 180-200 | 15 |
| 200-220 | 12 |
| 220-240 | 10 |
| 240-260 | 6 |
| 260-280 | 4 |
| 280-300 | 1 |

**Solution:**

**COMPUTATION OF QUARTILE, 3rd DECILE AND 50th PERCENTILE**

| Income (in Rs) | Number of persons (f) | Cumulative frequency (cf) |
|---|---|---|
| 100 - 120 | 6 | 6 |
| 120 - 140 | 10 | 16 |
| 140 - 160 | 18 | 34 |
| 160 - 180 | 30 | 64 |
| 180 - 200 | 15 | 79 |
| 200 - 220 | 12 | 91 |
| 220 - 240 | 10 | 101 |
| 240 - 260 | 6 | 107 |
| 260 - 280 | 4 | 111 |
| 280 - 300 | 1 | 112 |

**Computation of Lower Quartile**

$$Q_1 = L + \frac{\frac{N}{4} - cf}{f} \times i$$

Lower Quartile class = Size of $\frac{N}{4}$th item

= Size of $\frac{112}{4}$ th item

= size of 28th item

Size of 28th item lies in the class 140-160.

Here,

L = 140, cf = 16, f = 18, i = 20, N = 112

Substituting the values in the formula.

$$Q_1 = 140 + \frac{\frac{112}{4} - 16}{18} \times 20$$

$$= 140 + \frac{28-16}{18} \times 20$$

$$= 140 + \frac{12}{18} \times 20$$

$$= 140 + 0.67 \times 20$$

$$= 140 + 13.3$$

First Quartile = 153.3.

**Computation of upper Quartile**

$$Q_3 = L + \frac{\frac{3N}{4} - cf}{f} \times i$$

Upper Quartile class = Size of $3\frac{N}{4}$ th item

= size of $3\left(\frac{112}{4}\right)$ th item

= size of 84th item

= Size of 84 th items lies in the class 200-220.

Here,

L = 200, cf = 79, f = 12, i = 20, N = 112

Substituting the values in the formula,

$$Q_3 = 200 + \frac{3\frac{112}{4} - 79}{12} \times 20$$

$$= 200 + \frac{84-79}{12} \times 20$$

$$= 200 + \frac{5}{12} \times 20$$

$$= 200 + \frac{100}{12}$$

$$= 200 + 8.33$$

Upper Quartile = 208.33

**Computation of 3rd decile**

$$D_3 = L + \frac{\frac{3N}{10} - cf}{f} \times i$$

Third decile class = size of $3\ \frac{N}{10}$th item

= Size of $3\ \left(\frac{112}{10}\right)$ th item

= Size of 33.6th item

= size of 33.6th item in the class 140-160

Here,

$$L = 140,\ cf = 16,\ f = 18,\ i = 20,\ N = 112$$

Substituting the values in the formula

$$D_3 = 140 + \frac{3\left(\frac{112}{10}\right) - 16}{18} \times 20$$

$$= 140 + \frac{33.6 - 16}{18} \times 20$$

$$= 140 + \frac{17.6}{18} \times 20$$

$$= 140 + 0.978 \times 20$$

$$= 140 + 19.56$$

$$D_3 = 159.56$$

**Computation of 50th percentile**

$$P_{50} = L + \frac{50\left(\frac{N}{100}\right) - cf}{f} \times i$$

50th percentile class = size of $50\ \left(\frac{N}{100}\right)$ th item

= size of $50\ \left(\frac{112}{100}\right)$ th item

= size of 50 (1.12)th item

= size of 56th item

= size of 56th item lies in the class 160-18(

Here,

$$L = 160,\ cf = 34,\ f = 30,\ i = 20,\ N = 112$$

**Substituting the values in the formula,**

$$P_{50} = 160 + \frac{50\left(\frac{112}{100}\right) - 34}{30} \times 20$$

$$= 160 + \frac{56 - 34}{30} \times 20$$

$$= 160 + \frac{22}{30} \times 20$$

$$= 160 + \frac{440}{30}$$

$$= 160 + 14.66$$

50th percentile = 174.66

**Graphic location of Quartiles :** In order to locate the value of Quartiles graphically the following procedure is adopted.

1. Arrange the data by less than method and locate them on 'Y' axis
2. Cumulate the frequencies by less than method and locate them on 'Y' axis.
3. Find out the size of $\frac{N}{4}$th item and locate it on 'Y' axis
4. Draw an ogive curve by less than method.
5. Draw a horizontal line to 'X' axis through the point where the size of $\frac{N}{4}$th item is located.
6. Draw a perpendicular line on 'X' axis from the point where the Ogive curve and the horizontal line drawn from the point of the size of $\frac{N}{4}$th item intersects. Locate the value of lower quartile on 'X' axis.
   The value where the perpendicular line touches 'X' axis will be the value of first quartile.
7. Find out the size of $\frac{3N}{4}$th item and locate it on 'Y' axis
8. Draw a horizontal line to 'X' axis through the point where the size of $3\frac{N}{4}$th item is located.
9. Draw another perpendicular line on 'X' axis from the point where the ogive curve and the horizontal line through the point of $3\frac{N}{4}$th item intersect. Locate the value on 'X' axis at the point where the perpendicular line touches 'X' axis. The value at this point will be the value of Upper Quartile.

**Illustration - XIII:**

Locate the value of lower and upper quartile graphically from the following data :

| Marks | No.of Students |
|---|---|
| 0 - 10 | 5 |
| 10 - 20 | 6 |
| 20 - 30 | 8 |
| 30 - 40 | 18 |
| 40 - 50 | 20 |
| 50 - 60 | 15 |
| 60 - 70 | 12 |
| 70 - 80 | 4 |

**Solution :** To locate quartiles graphically, the frequency of the distribution needs cumulation.

| Marks (X) | No.of students (f) | Cumulative frequency (cf) |
|---|---|---|
| less than 10 | 5 | 5 |
| " 20 | 6 | 11 |
| " 30 | 8 | 19 |

| | | | |
|---|---|---|---|
| " | 40 | 18 | 37 |
| " | 50 | 20 | 57 |
| " | 60 | 15 | 72 |
| " | 70 | 12 | 84 |
| " | 80 | 4 | 88 |

**Location of quartile graphically** (Shown in fig:13.3)

Lower Quartile ($Q_1$) = Size of $\frac{N}{4}$th item

Lower Quartile = $\frac{88}{4}$ = Size of 22nd item

Upper Quartile ($Q_3$) = Size of 3 $\frac{N}{4}$th item

Upper Quartile = Size of 3 ( $\frac{88}{4}$ )th item

= size of 66th item

To locate lower quartile, draw a horizontal line parallel to 'X' axis from 22 to the ogvie curve which intersects at $R_1$. Draw a perpendicular line to the 'X' axis which intersects at $R_2$. The value $R_2$ represents lower quartile.

To locate upper quartile, draw a horizontal line parallel to 'X' axis from 66 to the ogive curve, which cuts at ($S_1$). The perpendicular line drawn from this point to the 'X' axis touching at $S_2$ represents the value of upper quartile.

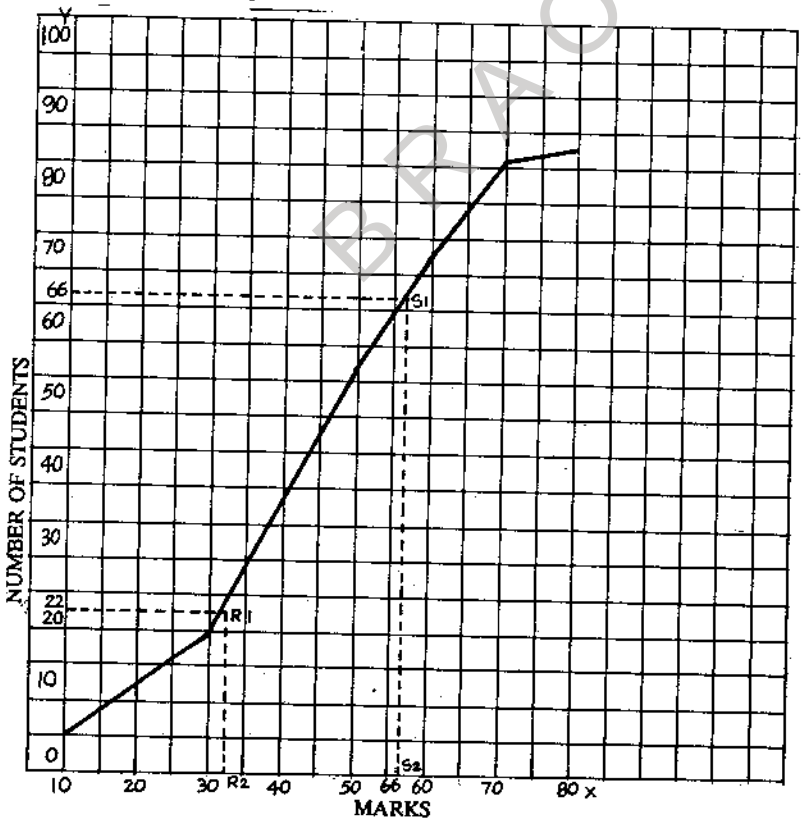



Fig : 13.3. Location of Lower and Upper quartiles

**Index :**

OY : 2 cms = 10 students

OX : 2 cms = 10 marks

Lower quartile: $R_2$ = 31.6 marks

Upper quartile : $S_2$ = 66 marks

## 13.8 SUMMING UP

Median is the middle value of the variable when the items are arranged in order of their magnitude. Median is not influenced by the extreme values of the distribution. In a distribution, where the data are unevenly distributed median is a better measure than mean. It is also useful in case of open-end classes. However, median cannot be considered the true representative of the given data as its computation is not based on each and every item of the series. It is also not amenable for further mathematical treatment.

Based on the median, certain other related positional measures like quartiles, deciles and percentiles have been developed. Median and other related positional averages can be graphically located by using ogive curves.

## 13.9 CHECK YOUR PROGRESS: MODEL ANSWERS

Obtain cumulative frequency and identify the median class. The median class is 15-20. Apply the following formula.

$$Md = L + \frac{\frac{N}{2} - cf}{f} \times i$$

The answer is 16.

## 13.10 MODEL EXAMINATION QUESTIONS

**A. short questions**

1. What is meant by 'median'?
2. Explain the terms:
   a) Quartiles
   b) Deciles
   c) Percentiles
3. "Median is a value which divides the frequency distribution into two equal parts." Elaborate this statement.
4. What are the properties of median ?
5. Discuss the merits and limitations of median.

6. Explain the procedure to locate median graphically.

**B. Essay Questions**

7. Find out median income from the following:

Monthly Income (Rs): 300, 220, 150, 180, 100, 50, 120, 260, 60, 80, 90

(Ans : 120)

8. The daily sales in a company are given below: Find out median.

| Name of the day: | Mon, | Tues, | Wed, | Thur, | Fri, | Sat. |
|---|---|---|---|---|---|---|
| Sales (Rs.'000): | 30 | 32 | 26 | 28 | 31 | 30 |

(Ans : 30)

9. Find out lower quartile and upper quartile from the following distribution:

| X | f |
|---|---|
| 0 - 5 | 10 |
| 5 - 10 | 20 |
| 10 - 15 | 25 |
| 15 - 20 | 30 |

(Ans : $Q_1$ = 7.81, $Q_3$= 16.46)

10. The data given below relates to the height of the students in a class:

| Height (in Inches) : | 57 | 60 | 62 | 65 | 70 | 55 | 58 |
|---|---|---|---|---|---|---|---|
| No.of students: | 10 | 12 | 14 | 8 | 6 | 20 | 15 |

Calculate median and quarties.

(Ans: Md = 58; $Q_1$ = 57; $Q_3$=62)

11. The following data relates to the income of workers in a factory:

| Income (Rs.): | 250 | 180 | 280 | 300 | 150 | 170 |
|---|---|---|---|---|---|---|
| No.of workers: | 10 | 15 | 18 | 12 | 22 | 7 |

Find out median and quartiles.

(Ans: Md = 180; $Q_1$ = 150; $Q_3$ = 280)

12. The following are the marks obtained by the students of B.com class in statistics.

| Marks : | 30 | 42 | 38 | 56 | 32 | 60 | 45 | 50 |
|---|---|---|---|---|---|---|---|---|
| No. of Students: | 5 | 20 | 8 | 6 | 7 | 4 | 10 | 5 |

Calculate median and quarties.

(Ans: Md = 42, $Q_1$ = 38; $Q_3$ = 45)

13. From the following data claculate the median

| Class interval : | 25-35 | 35-45 | 45-55 | 55-65 | 65-75 |
|---|---|---|---|---|---|
| Frequency: | 7 | 10 | 8 | 15 | 20 |

(Ans: 58.33)

14. The data given below relates to the income of persons in a village.

| Income (Rs.): | 105-115 | 115-125 | 125-135 | 135-145 | 145-155 | 155-16 |
|---|---|---|---|---|---|---|
| No. of persons : | 12 | 13 | 18 | 20 | 30 | 16 |

Calculate median, quartiles, 4th decile and 65th percentile.

(Ans: Md= 140.75; $Q_1$ = 126.25; $Q_3$=151.25; $D_4$ = 135.3; $P_{65}$=147.62)

15. Find out median, 7th decile and 45th percentile from the following:

| Class interval | Frequency |
|---|---|
| 8 - 12 | 10 |
| 12 - 16 | |
| 16 - 20 | |
| 20 - 24 | 18 |
| 24 - 28 | 17 |
| 28 - 32 | 12 |
| 32 - 36 | 8 |
| 36 - 40 | 6 |

(Ans: Md=22.44; $D_7$ = 27.06; $P_{45}$ = 21.33)

16. The marks obtained by 80 students in a class are shown below:

| Marks | No. of students |
|---|---|
| Less than 30 | 2 |
| " 40 | 10 |
| " 50 | 16 |
| " 60 | 18 |
| " | 20 |
| " 80 | 28 |
| " 90 | |

Calculate median, 6th decile and 27th percentile.

(Ans : Md=60; $D_6$ = 72; $P_{27}$ = 39.65)

17. Calculate the upper and lower quartiles from the data shown below:

| Class Interval : | 4-8 | 8-12 | 12-16 | 16-20 | 20-24 | 24-28 | 28-32 | 32-36 |
|---|---|---|---|---|---|---|---|---|
| Frequency : | 2 | 4 | 20 | 10 | 9 | 8 | 6 | 1 |

(Ans : $Q_1$ = 13.8 ; $Q_3$ = 24 )

18. Compute the values of quartiles and 2nd decile from the following data:

| Marks | Frequency |
|---|---|
| Below 30 | 2 |
| " 40 | 12 |
| " 50 | 14 |
| " 60 | 20 |
| " 70 | 24 |
| " 80 | 30 |

(Ans : $Q_1$ = 35.5; $Q_3$ = 66.25; $D_2$ = 34)

19. From the following data locate the value of median graphically.

| Size | Frequency |
|---|---|
| 5 - 10 | 5 |
| 10 - 15 | 15 |
| 15 - 20 | 25 |
| 20 - 25 | 20 |
| 25 - 30 | 20 |
| 30 - 35 | 18 |
| 35 - 40 | 12 |
| 40 - 45 | 10 |
| 45 - 50 | 5 |

20. From the following data, locate upper and lower quartiles graphically

| Marks : | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 |
|---|---|---|---|---|---|---|---|
| No. of Students: | 16 | 28 | 35 | 40 | 20 | 10 | 6 |

## 13.11 RECOMMENDED BOOKS


1. Gupta, S.P. : "Statistical Methods", Sultan Chand & company, New Delhi.
2. Gupta, B.N. : "Statistics," Sahitya Bhavan, Agra.
3. Gupta, S.C. : "Fundamentals of Statistics", Himalaya Publishing house, Bombay.
4. Simpson and Kafka : "Basic statistics," Oxford and I.B.H. Publishing Company , Calcutta.


## 13.12 GLOSSARY

1. Deciles : Decile divide the total frequency into ten equal parts.
2. Median : Median is the value of central item when the items arranged in the ascending or descending order of their magnitude.
3. Percetiles : Percentiles divide the total frequency into hundred equal parts.
4. Quartiles : Quartiles divide the total frequency into four equal quarters.

# UNIT - 14 : MODE

**Contents**



## 14.0 AIMS AND OBJECTIVES

The aims of this unit are to explain the meaning, computation, merits and limitations of mode.

After going through this unit, you should be able to :

- explain the meaning of Mode
- compute the model value
- identify mode by graphic method
- list the merits and limitations of Mode



## 14.1 INTRODUCTION

Like median, mode is a positional average. Mode as a statistical average is the observation that occurs with the greatest frequency and thus is the most fashionable value. In certain circumstances arithmetic mean and median fail to represent the mass data meaningfully and convincingly as their computed values tend to be illogical. To overcome this limitation, mode has been introduced.

## 14.2 MEANING OF MODE

The word mode is said to have been derived from the french work 'LAMODE' which means fashion. The Dictionary meaning of "mode" is the variate at which a relative or absolute maximum occurs more frequently or for a more number of times than any other value in the given frequency distribution. When the values of a variable are represented by a curve, the peak of the curve will be the value of mode as is clear from the fig.14.1

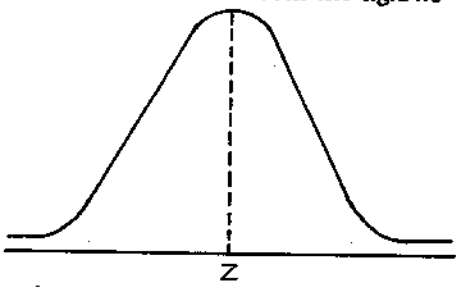



Fig.14.1 Showing the value of Mode.

For example, in a series comprising 25, 42, 28, 31, 30, 42, 29, 20, 42, 34, 42 the mode would be 42, as the value 42 occurred more number of times than any other value in the series. Since mode refers to the occurrence of an event more frequently, it is considered to be the modal value or typical value which represents the entire data. If all the values in a series occur in the same number of times, there is no modal value or mode.

Some of the important definitions of mode are given below:

According to Ya-Lun-Chou "The mode is that value of a series which appears more frequently than any other". Taro Yamane defined it as,. "The mode of distribution is the only value at which the frequency density is at a maximum or it is any value of the variable that occurs most frequently". A.M. Tuttle observed "Mode is that value which has the greatest frequency density in its immediate neighbourhood". In the words of Croxton and Cowden "The mode of the distri bution is the value around which the items tend to be heavily concentrated. It may be regarded as the most typical of a series of values."

## 14.3 COMPUTATION OF MODE

Mode can be computed from individual, discrete and continuous series.

### A. Individual Series

In the case of individual series, mode is determined by counting the number of times each value repeats. The value of that item which repeats maximum number of times is taken as modal value or mode. Practically this process of determining mode amounts to the conversion of individual series into discrete series.

**Illustration - I**

Find out modal value from the following data :

27, 28, 30, 33, 31, 35, 34, 33, 40, 41,

55, 46, 31, 33, 36, 33, 41, 33.

**Solution :**

**COMPUTATION OF MODE**

| Size of item | No.of times it occurred |
|---|---|
| 27 | 1 |
| 28 | 1 |
| 30 | 2 |
| 31 | 2 |
| 33 | 5 |
| 34 | 1 |
| 35 | 1 |
| 36 | 1 |
| 40 | 1 |
| 41 | 2 |
| 46 | 1 |
| 55 | 1 |

As seen from the above analysis, the item 33 occurred maximum number of times i.e., 5 times. Hence 33 is considered to be the modal value of the given data.

**B. Discrete Series**

In discrete series mode is computed either by inspection method or by grouping method.

**a) Inspection Method :** In this method, mode of a discrete series is determined by inspecting the values and their respective frequencies. The value which occurs maximum number of times in the series is regarded as mode. It implies that the value of the variable around which the items are most heavily concentrated is considered as mode.

**Illustration - II**

Following are the weights of 80 students. Find out the mode.

| Weights (in Kgs) | : 30 | 40 | 35 | 45 | 50 | 55 | 60 | 65 | 70 |
|---|---|---|---|---|---|---|---|---|---|
| Number of students | : 8 | 8 | 10 | 7 | 20 | 13 | 12 | 6 | 4 |

**Solution:**

By inspecting the data, it can be noticed that the item 50 Kgs has occurred as many number of times as 20. Since it has occurred maximum number of times, it is the modal value.

**b) Grouping method:** Some times the given data may be irregular and there may not be any significant difference between the occurrence of two or more values. In such cases the mode determined by inspection method may not be a typical value and this determination of mode necessitates the use of grouping method. Determination of mode by grouping method involves the preparation of a grouping table and an analysis table.

For the preparation of a grouping table, the following procedure is adopted.

i) Write the frequencies given in the distribution in column I and note the highest frequency.

ii) Group the frequencies in twos from the top and record them in column II and note the highest frequency.

iii) Group the frequencies in twos leaving the first frequency from the top and record them in column III and note the highest frequency.

iv) Group the frequencies in threes from the top and record them in column IV and note the highest frequency.

v) Group the frequencies in threes leaving first frequency from the top and record them in column V and note the highest frequency.

vi) Group the frequencies in threes leaving the first and the second frequencies from the top and record them in column VI and note the highest frequency.

Depending on the number of items, the above procedure can be continued further or terminated even at column IV and V as the case may be.

The analysis table is prepared from the grouping table as detailed below:

i) Write column numbers horizontally on the right side of the table.

ii) Write the items of the distribution vertically on left-hand side of the table.

iii) Record the highest frequencies of each column corresponding to their values with mark (I) in the table.

iv) Find out the value which has been repeated maximum number of times. The value so obtained will be the mode of a given distribution.

**Illustration - III**

Compute the modal value from the following data :

| Age (in Yrs) | : 20 | 25 | 30 | 35 | 40 | 45 | 50 | 55 | 60 |
|---|---|---|---|---|---|---|---|---|---|
| No. of persons | : 6 | 18 | 28 | 30 | 25 | 12 | 10 | 8 | 6 |

**Solution: Computation of Mode**

GROUPING TABLE

| Age in years | No.of persons Col.I | Col.II | Col.III | Col.IV | Col.V | Col.VI |
|---|---|---|---|---|---|---|
| 20 | 6 | | | | | |
| | | 24 | | | | |
| 25 | 18 | | | 52 | | |
| | | | 46 | | | |
| 30 | 28 | | | | 76 | |
| | | 58 | | | | |
| 35 | 30 | | | | | 83 |
| | | | 55 | | | |
| 40 | 25 | | | 67 | | |
| | | 37 | | | | |
| 45 | 12 | | | | 47 | |
| | | | 22 | | | |
| 50 | 10 | | | | | 30 |
| | | 18 | | | | |
| 55 | 8 | | | 24 | | |
| | | | 14 | | | |
| 60 | 6 | | | | | |

ANALYSIS TABLE

| Age in years | Col. I | Col. II | Col. III | Col. IV | Col. V | Col. VI | TOTAL |
|---|---|---|---|---|---|---|---|
| 20 | | | | | | | |
| 25 | | | | | 1 | | 1 |
| 30 | | 1 | | | 1 | 1 | 3 |
| 35 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 40 | | | 1 | 1 | | 1 | 3 |
| 45 | | | | 1 | | | 1 |
| 50 | | | | | | | |
| 55 | | | | | | | |
| 60 | | | | | | | |

Since the value 35 occurred maximum number of times i.e., 6 times. 35 is the modal value.

**C. Continuous Series**

The following procedure is adopted to determine the modal value in a continous series.

i) Prepare the grouping table

ii) Prepare the analysis table and obtain the modal class. The same procedure as adopted in the case of a discrete series, can be followed for preparing the grouping table and analysis table.

iii) Compute the value of mode by applying the formula given below.

$$Z = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

Where,

Z = Mode

L = Lower limit of the modal class

$f_0$ = Frequency of the class preceding the modal class

$f_1$ = Frequency of the modal class

$f_2$ = Frequency of the class succeeding the modal class

i = Class interval of the modal class.

The above formula can also be expressed as below.

$$Z = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

Where:

L = Lower limit of the modal class

$\Delta_1$ = (Pronounced it as 'Delta') = the difference between the frequency of the modal class and the frequency of the class preceding the modal class.

$\Delta_2$ = The difference between the frequency of the modal class and the frequency of the class succeeding the modal class.

i = Class interval of the modal class.

**Illustration - IV**

Find out the modal value from the following data.

| Profits earned (in Rs.'000) | No. of Companies |
|---|---|
| 10-20 | 5 |
| 20-30 | 7 |
| 30-40 | 9 |
| 40-50 | 12 |
| 50-60 | 15 |
| 60-70 | 10 |
| 70-80 | 4 |
| 80-90 | 3 |
| 90-100 | 2 |

**Solution : Computation of Mode**

**GROUPING TABLE**

| Profits earned in (Rs. '000) | No. of Companies Col. I | Col. II | Col. III | Col. IV | Col. V | Col. VI |
|---|---|---|---|---|---|---|
| 10-20 | 5 | | | | | |
| | | 12 | | | | |
| 20-30 | 7 | | | | | |
| | | | 16 | | | |
| 30-40 | 9 | | | | 28 | |
| | | 21 | | | | |
| 40-50 | 12 | | | | | 36 |
| | | | 27 | | | |
| 50-60 | 15 | | | 37 | | |
| | | 25 | | | | |
| 60-70 | 10 | | | | 29 | |
| | | | 14 | | | |
| 70-80 | 4 | | | | | 17 |
| | | 7 | | | | |
| 80-90 | 3 | | | 9 | | |
| | | | 5 | | | |
| 90-100 | 2 | | | | | |

**ANALYSIS TABLE**

| Profits in Rs. '000 | Col. I | Col. II | Col. III | Col. IV | Col. V | Col. VI | TOTAL |
|---|---|---|---|---|---|---|---|
| 10-20 | | | | | | | 0 |
| 20-30 | | | | | | | 0 |
| 30-40 | | | | | | 1 | 1 |
| 40-50 | | | 1 | 1 | | 1 | 3 |
| 50-60 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 60-70 | | 1 | | 1 | 1 | | 3 |
| 70-80 | | | | | 1 | | 1 |
| 80-90 | | | | | | | 0 |
| 90-100 | | | | | | | 0 |

Since the class 50-60 occurred for more number of times, it will be the modal class.

**NOTE :** Here, modal class can also be determined by inspection method. Inspection method avoids preparation of grouping table and analysis table:

Applying the formula,

$$Z = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

Here,

L = 50; $f_1$ = 15; $f_0$ = 12; $f_2$ = 10; i = 10

Substituting the values in the formula,

$$Z = 50 + \frac{15 - 12}{2 \times 15 - 12 - 10} \times 10$$

$$= 50 + \frac{3}{30 - 22} \times 10$$

$$= 50 + \frac{3}{8} \times 10$$

$$= 50 + 0.375 \times 10$$

$$= 50 + 3.75$$

$\therefore$ Mode = 53.75

**Illustration - V**

Compute modal value from the following data:

| Sales (in.Rs.lakhs) | 10-20 | 20-30 | 30-50 | 50-60 | 60-70 | 70-80 |
|---|---|---|---|---|---|---|
| No. of firms | 5 | 4 | 12 | 3 | 1 | 2 |

**Solution:**

Since the class intervals of all the classes of the given data are not equal, the data must be arranged on the assumption that the classes are equal and frequencies are equally distributed throughout the class. Thus, the given data can be adjusted and rearranged as below :

| Sales (in Rs.lakhs) | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 | 70-80 |
|---|---|---|---|---|---|---|---|
| No. of firms | 5 | 4 | 6 | 6 | 3 | 1 | 2 |

**GROUPING TABLE**

| Sales (Rs. lakhs) | No. of firms Col.I | Col.II | Col.III | Col.IV | Col.V |
|---|---|---|---|---|---|
| 10-20 | 5 | | | | |
| | | 9 | | | |
| 20-30 | 4 | | | 15 | |
| | | | 10 | | |
| 30-40 | 6 | | | | 16 |
| | | | | | |
| 40-50 | 6 | 12 | | | |
| | | | 9 | | |
| 50-60 | 3 | | | 10 | |
| | | | | | 6 |
| 60-70 | 1 | 4 | | | |
| | | | 3 | | |
| 70-80 | 2 | | | | |

**ANALYSIS TABLE**

| Sales (Rs.lakhs) | Col. I | Col. II | Col. III | Col. IV | Col. V | TOTAL |
|---|---|---|---|---|---|---|
| 10-20 | | | | 1 | | 1 |
| 20-30 | | | 1 | 1 | 1 | 3 |
| 30-40 | 1 | 1 | 1 | 1 | 1 | 5 |
| 40-50 | 1 | 1 | | | 1 | 3 |
| 50-60 | | | | | | - |
| 60-70 | | | | | | - |
| 70-80 | | | | | | - |



Since the class 30-40 occurred more number of times it will be modal class.

Applying the formula,

$$Z = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

Here,

L = 30, $f_1$ = 6; $f_0$ = 4; $f_2$ = 6; i = 10

Substituting the values in the formula,

$$Z = 30 + \frac{6-4}{2 \times 6 - 4 - 6} \times 10$$

$$= 30 + \frac{2}{12 - 4 - 6} \times 10$$

$$= 30 + \frac{2}{12 - 10} \times 10$$

$$= 30 + 1 \times 10$$

$$\therefore \quad \text{Mode} = 40$$

When the distribution possess one mode, it is called uni-modal distribution. If the distribution possesses two or more modes, it is called bi-modal and multi-modal distribution respectively. The fig:14.2 explains the bi-modal distribution of a variable

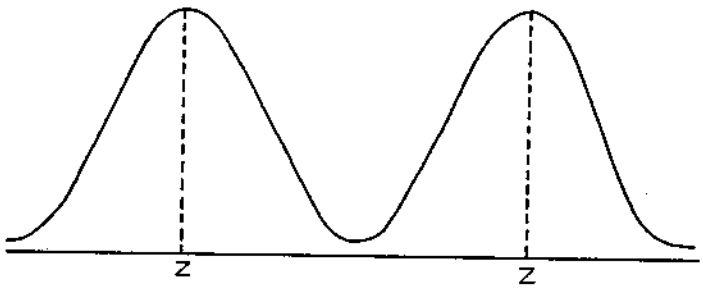



Fig: 14.2 Bi-modal distribution of a variable

When the distribution possess more than one mode, the value of mode cannot be determined by the formula used earlier. Hence, the mode of distribution is ill-defined.

The data may consist more than one modal value on account of inadequate size of the sample and heterogeneous composition of the variable.

When the mode of the distribution is ill-defined, its value is determined by using the following formula which is based on the realtionship between mean, median and mode.

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean.}$$

**Illustration - VI**

Compute the mode from the following data.

| Savings (in Rs.) | No. of families |
|---|---|
| 0-400 | 8 |
| 400-800 | 12 |
| 800-1200 | 40 |
| 1200-1600 | 64 |
| 1600-2000 | 66 |
| 2000-2400 | 34 |
| 2400-2800 | 16 |
| 2800-3200 | 4 |

**Solution : Computation of Mode**

**GROUPING TABLE**

| Savings (in Rs.) | No. of Families Col.I | Col.II | Col.III | Col.IV | Col.V | Col.VI |
|---|---|---|---|---|---|---|
| 0-400 | 8 | | | | | |
| | | 20 | | | | |
| 400-800 | 12 | | | 60 | | |
| | | | 52 | | | |
| 800-1200 | 40 | | | | 116 | |
| | | 104 | | | | 170 |
| 1200-1600 | 64 | | | | | |
| | | | 130 | | | |
| 1600-2000 | 66 | | | 164 | | |
| | | 100 | | | | |
| 2000-2400 | 34 | | | | 116 | |
| | | | 50 | | | |
| 2400-2800 | 16 | | | | | 54 |
| | | 20 | | | | |
| 2800-3200 | 4 | | | | | |

**ANALYSIS TABLE**

| Income (in Rs.) | Col. I | Col. II | Col. III | Col. IV | Col V | Col. VI | TOTAL |
|---|---|---|---|---|---|---|---|
| 0-400 | | | | | | | - |
| 400-800 | | | | | 1 | | 1 |
| 800-1200 | | 1 | | | 1 | 1 | 3 |
| 1200-1600 | | 1 | 1 | 1 | 1 | 1 | 5 |
| 1600-2000 | 1 | | 1 | 1 | 1 | 1 | 5 |
| 2000-2400 | | | | 1 | 1 | | 2 |
| 2400-2800 | | | | | 1 | | 1 |
| 2800-3200 | | | | | | | - |

Since the classes 1200 - 1600 and 1600 - 2000 have been repeated maximum number of times ie.,5 times each this is a case of bi-modal series.

Therefore, to obtain the mode, the following formula may be used

$$\text{Mode} = 3 \text{ median} - 2 \text{ mean.}$$

| Income (in Rs.) | No.of Families | c f | m | dx | $\frac{dx}{C=400}$ $d^1$ | $fd^1$ |
|---|---|---|---|---|---|---|
| 0-400 | 8 | 8 | 200 | −1600 | −4 | −32 |
| 400-800 | 12 | 20 | 600 | −1200 | −3 | −36 |
| 800-1200 | 40 | 60 | 1000 | −800 | −2 | −80 |
| 1200-1600 | 64 | 124 | 1400 | −400 | −1 | −64 |
| 1600-2000 | 66 | 190 | 1800 | 0 | 0 | 0 |
| 2000-2400 | 34 | 224 | 2200 | 400 | 1 | 34 |
| 2400-2800 | 16 | 240 | 2600 | 800 | 2 | 32 |
| 2800-3200 | 4 | 224 | 3000 | 1200 | 3 | 12 |
| | N= 224 | | | | | $fd^1 = -134$ |

**Computation of Median**

Median item = Size of $\frac{N}{2}$th item

Here,

N = 224

Median = Size of $\frac{244}{2}$th item = size of 122nd item

Size of 122nd item lies in the class 1200 - 1600

$$Median = L + \frac{\frac{N}{2} - cf}{f} \times i$$

Here,

N/2 = 122; L = 1200; cf = 60; i = 400; f = 64

Substituting the values in the formula,

$$Med = 1200 + \frac{122 - 60}{64} \times 400$$

$$= 1200 + \frac{62}{64} \times 400$$

$$= 1200 + 0.968 \times 400$$

$$= 1200 + 387.2$$

$\therefore$ Median = Rs. 1587.2

**Calculation of Arithmetic Mean**

$$\overline{X} = A + \frac{\Sigma f d^1}{N} \times i$$

Here,

A = 1800; N = 244; i = 400; $\Sigma f d^1 = -134$

Substituting the values in the formula,

$$\overline{X} = 1800 + \frac{-134}{244} \times 400$$

$$= 1800 + (-0.549) \times 400$$

$$= 1800 - 219.6$$

$$\therefore \text{ Mean} = \text{Rs. } 1580.4$$

Here,

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$$

Substituting the values in the formula,

$$Z = 3\ (1587.2) - 2\ (1580.4)$$

$$= 4761.6 - 3160.8$$

$$\therefore \quad \text{Mode} = \text{Rs. } 1600.8$$

## Check your progress - 1

In a distribution the arithmetic mean is 65 and the Median is 70. Find out the Mode.

_____

_____

_____

**Illustration - VII**

Compute modal value from the following frequency distribution :

| Age in years | Number of persons |
|---|---|
| more than 10 | 144 |
| " 20 | 138 |
| " 30 | 122 |
| " 40 | 98 |
| " 50 | 72 |
| " 60 | 42 |
| " 70 | 17 |
| " 80 | 5 |

**Solution :**

For computation of mode, the data given has to be rearranged in the following manner according to 'more than' method of cumulative frequencies.

| Age in Years | No. of persons |
|---|---|
| 10-20 | 6 |
| 20-30 | 16 |
| 30-40 | 24 |
| 40-50 | 26 |
| 50-60 | 30 |
| 60-70 | 25 |
| 70-80 | 12 |
| 80-90 | 5 |

**GROUPING TABLE**

| Age in years | No.o f Persons Col.I | Col.II | Col.III | Col.IV | Col.V | Col.VI |
|---|---|---|---|---|---|---|
| 10-20 | 6 | | | | | |
| | | 22 | | | | |
| 20-30 | 16 | | | | | |
| | | | 40 | | | |
| 30-40 | 24 | | | 46 | | |
| | | 50 | | | | |
| 40-50 | 26 | | | | 66 | |
| | | | 56 | | | 80 |
| 50-60 | 30 | | | 81 | | |
| | | 55 | | | | |
| 60-70 | 25 | | | | 67 | |
| | | | 37 | | | |
| 70-80 | 12 | | | | | 42 |
| | | 17 | | | | |
| 80-90 | 5 | | | | | |

**ANALYSIS TABLE**

| Age in years | Col. I | Col. II | Col. III | Col. IV | Col. V | Col. VI | TOTAL |
|---|---|---|---|---|---|---|---|
| 10-20 | | | | | | | - |
| 20-30 | | | | | | | - |
| 30-40 | | | | | | 1 | 1 |
| 40-50 | | | 1 | 1 | | 1 | 3 |
| 50-60 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| 60-70 | | 1 | | 1 | 1 | | 3 |
| 70-80 | | | | | 1 | | 1 |
| 80-90 | | | | | | | - |



Since the class 50-60 occurred more number of times i.e., 6 times, the modal class will be 50-60.

Applying the formula.

$$Z = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

Here,

L = 50; $f_1$ = 30; $f_0$ = 26; $f_2$ = 25; i = 10

Substituting the values in the formula,

$$Z = 50 + \frac{30 - 26}{2 \times 30 - 26 - 2} \times 10$$

$$= 50 + \frac{4}{9} \times 10$$

$$= 50 + 4.44$$

$$= 54.44$$

Hence,

$\therefore$ Mode is 54.44

**Illustration - VIII**

From the information given below compute the Mode.

| Marks | No. of students |
|---|---|
| 0-10 | 15 |
| 10-20 | 30 |
| 20-30 | 54 |
| 30-40 | 69 |
| 40-50 | 84 |
| 50-60 | 18 |
| 60-70 | 6 |

**Solution : Computation of Mode**

**GROUPING TABLE**

| Marks | No.of Students | | | | |
|---|---|---|---|---|---|
| | Col.I | Col.II | Col.III | Col.IV | Col.V |
| 0-10 | 15 | | | | |
| | | 45 | | | |
| 10-20 | 30 | | | 99 | |
| | | | 84 | | |
| 20-30 | 54 | | | | 153 |
| | | 123 | | | |
| 30-40 | 69 | | | | |
| | | | 153 | | |
| 40-50 | 84 | | | 171 | |
| | | 102 | | | |
| 50-60 | 18 | | | | 108 |
| | | | 24 | | |
| 60-70 | 6 | | | | |

**ANALYSIS TABLE**

| Marks | Col. I | Col. II | Col. III | Col. IV | Col. V | TOTAL |
|---|---|---|---|---|---|---|
| 0-10 | | | | | | - |
| 10-20 | | | | | 1 | 1 |
| 20-30 | | 1 | | | 1 | 2 |
| 30-40 | | 1 | 1 | 1 | 1 | 4 |
| 40-50 | 1 | | 1 | 1 | | 3 |
| 50-60 | | | | 1 | | 1 |
| 60-70 | | | | | | - |

Since the class 30 - 40 occurred more number of times, it is the modal class.

$$Z = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

Here,

L = 30, $f_1$ = 69; $f_0$ = 54; $f_2$ = 84; i = 10

Substituting the values in the formula,

$$Z = 30 + \frac{69 - 54}{2 \times 69 - 54 - 84} \times 10$$

$$= 30 + \frac{15}{138 - 138} \times 10$$

$$= 30 + \frac{15}{0} \times 10$$

Here, since $2f_1 = f_0 + f_2$

$$2f_1 - f_0 - f_2 = 0$$

Thus the denominator is equal to zero.

This indicates that some error is involved in the method adopted. In such a case an alternative formula which is given below is followed to find out the Mode.

$$Z = L + \frac{f_2}{f_0 + f_2} \times i$$

Where,

L = Lower limit of the modal class

$f_0$ = Frequency of the class preceding the modal class

$f_2$ = Frequency of the class succeeding the modal class

i = Class interval of the modal class.

Here, L = 30; $f_2$ = 84; $f_0$ = 54; i = 10.

Substituting the values in the formula,

$$Z = 30 + \frac{84}{54 + 84} \times 10$$

$$Z = 30 + \frac{84}{138} \times 10$$

$$= 30 + 0.608 \times 10$$

$$= 30 + 6.08$$

$$\therefore \quad \text{Mode} = 36.08$$

## 14.4 LOCATION OF MODE BY GRAPHIC METHOD

The value of mode of a frequency distribution can also be located graphically. The following steps are involved in the location of mode graphically.

i) Draw a histogram for the given data.

ii) Connect the inner corner points of the modal class bar with the upper corner points of the two adjacent bars. This is done by drawing two diagonal lines inside the modal class bar.

iii) Draw a perpendicular line to the X-axis from the point of inter section of the two diagonal lines.

iv) The value at which the perpendicular line touches X-axis is the mode of the given series.

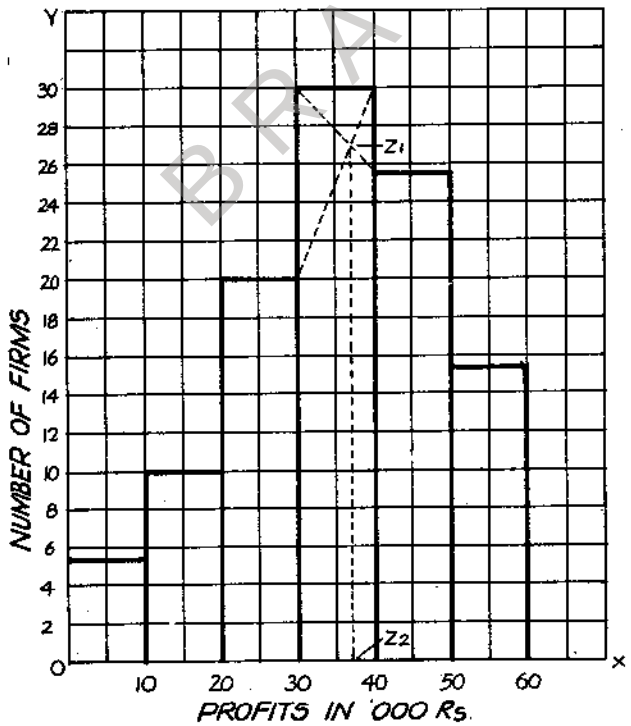
**Illustration- IX**

From the following data, locate the mode graphically and check up this value by direct calculation.

| Profits earned (Rs. in '000) | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 |
|---|---|---|---|---|---|---|
| No. of firms: | 5 | 10 | 20 | 30 | 25 | 15 |

**Solution :**

To locate mode, a histrogram is drawn (fig 14.3) with the diagonal values. By inspection modal class bar is found to be in 30-40. The diagonal lines joining the inner corner points of modal bar to the adjacent bars intersect at the point $Z_1$. From this point draw a perpendicular line to the 'X' axis which cuts 'OX' line at $Z_2$. This value represents mode i.e. 36.6




INDEX:

OX, 2cms. = Rs.10,000
OY, 1 Cm. = 2 Firms

fig. 14.3 Location of Mode

**Calculation of Mode :**

$$Z = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

By inspecting the given series, it can be said that the mode lies in the class 30-40.

Here,

L = 30; $f_1$ = 30; $f_0$ = 20; $f_2$ = 25; i = 10

Substituting the values in the formula,

$$Z = 30 + \frac{30 - 20}{2 \times 30 - 20 - 25} \times 10$$

$$= 30 + \frac{100}{15}$$

$$\therefore \quad \text{Mode} = 36.66$$

Hence the value of mode located graphically and computed through calculation is identical. However, mode is located graphically when the distribution contains one highest frequency only.

## 14.5 MERITS AND LIMITATIONS OF MODE

**Merits**

1) Mode refers to the value that occurs maximum number of times in the series. Hence mode is considered to be the typical value which represents the whole data. For example, if the modal wages in factory is Rs. 325 per month, it indicates that majority of the workers in the factory are getting wages of Rs. 325.

2) Unlike arithmetic mean, it is not affected by the extreme values of the variable. For example, if the values of a variable are 10, 80, 80, and 100, the mode is 80 or if the values of a variable are 2,80,10, 80,80 and 1000, then also the mode is 80.

3) Mode can be used to described qualitative phenomena such as comparison of consumer preferences for different types of articles. In this case, modal value of consumer preferences can be obtained before the production policy is decided. Here, the modal value reveals the preferences of majority of consumers towards a particular product.

4) Mode can be computed in open-end distribution even without ascertaining the class limits.

5) Unlike arithmetic mean, modal value of a distributioin can be determined graphically.

6) It is easy to compute and easy to understand the mode as in many a case mode can be determined simply by observing the data.

**Limitations**

1) Mode cannot be taken as the true representative of whole data as its computation is not based on all the values of a given variable.

2) Mode cannot be taken as the true algebraic treatment. Thus we can not compute the combined modal value for two or more related groups by combining the series unless the data are normally distributed. For example, given the modal wages of two factories as Rs.400 and 500 respectively, it is not possible to calculate the over all mode of combined data. Whereas combined mean can be calculated by combining the two series.

3) It is not easy to compute a modal value as it involves lengthy calculations both in grouping the items and analysing the values and their occurrences.

4) Mode determined by inspection method will not be the accurate and of precise value and it often will mislead the readers.

5) Mode located graphically may given an approximate value but not an accurate and precise value.

In the words of Ya-Lun-Chou, "Mode is the most unstable average and its true value is difficult to determine. Morever, the value of mode is affected significantly by the size of the class interval used in grouping data into any frequency distribution. A change in the size of the class interval will change the value of the mode".

## 14.6 SUMMING UP

Mode is the value of that item of a series which appears more frequently then any other item in a given distribution. Usually, mode can be located by inspection in the case of individual series, but in discrete and continuous series, grouping method is followed to locate modal value. Sometimes, more than one mode may exist in the series, in which case mode is said to be ill-defined. When mode is ill-defined, its value may be found by a method which is based on the relationship between mean, median and mode (mode = 3 median − 2 mean). Unlike arithmetic mean, mode is not affected by the extreme values of the distribution. It is a useful measure to describe the qualitative phenomenon. However, mode cannot be considered a true representative of the distribution as its computations are not based on all the items.

## 14.7 CHECK YOUR PROGRESS: MODEL ANSWER

The value of mode can be ascertained by applying the following formula.

Mode = 3 Median − 2 Mean

The answer is 80.

## 14.8 MODEL EXAMINATION QUESTIONS

**A. Short Questions**

1) What is meant by 'modal value' ?

2) What do you mean by 'uni-modal' distribution?

3) What is a 'bi-modal' distribution?

4) Discuss the merits and limitations of mode.

5) Explain the procedure to determine the mode. When do you say that mode is ill-defined in the distribution.

6) Explain the procedure for computing mode in continuous series.

**B. Essay Questions**

7) Explain the relationhsip between mean, median and mode in a symmetrical distribution.

**EXERCISES**

8) Given below are the marks obtained by 20 students in an examination. Obtain modal value.

Marks : 68 80 43 48 73 60 68 73 68 54 54
56 52 64 72 51 54 55 68 36

(Ans: 68)

9) The following data relates to the daily wages of workers

Wages (Rs.) 5, 6, 5.50, 4.50, 5, 6.50, 7.50, 4.75
6.50, 7.50, 8.50, 4.75, 5.00, 4.75

Calculate Mode

(Ans: 4.07)

10) The following are the marks obtained by students in an examination:

| Marks : | 55 | 60 | 66 | 70 | 72 | 75 |
|---|---|---|---|---|---|---|
| No. of students | 6 | 14 | 20 | 25 | 18 | 10 |

Calculate Modal marks

(Ans: 70)

11. find out mode from the following data:

| Age ( in years) : | 14 | 20 | 22 | 25 | 30 | 35 |
|---|---|---|---|---|---|---|
| No of persons : | 50 | 75 | 100 | 60 | 25 | 15 |

(Ans : 22)

12. From the following information compute mode.

| Size of items : | 6 | 8 | 9 | 12 | 13 | 15 | 16 | 20 | 21 | 22 | 25 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Frequency : | 3 | 9 | 7 | 11 | 10 | 12 | 8 | 12 | 14 | 6 | 50 |

(Ans : 22.83)

13. Calculate the mode from the daily wages of workers given below:

| Daily Wages (Rs.): | 10 | 11 | 13 | 15 | 17 | 19 | 18 | 20 | 22 |
|---|---|---|---|---|---|---|---|---|---|
| No of Workers: | 2 | 17 | 20 | 22 | 18 | 14 | 10 | 8 | 5 |

(Ans : 15)

14. Compute mode from the following data.

| Salary (Rs.) | No of persons |
|---|---|
| Below 60 | 15 |
| 60-80 | 20 |
| 80-100 | 32 |
| 100-120 | 28 |
| 120-140 | 15 |
| 140 and above | 10 |

(Ans : 95)

15. Find out mode from the data given below :

| Marks - | No.of students |
|---|---|
| Below 10 | 5 |
| " 20 | 10 |
| " 30 | 12 |
| " 40 | 20 |
| " 50 | 30 |
| " 60 | 35 |
| " 70 | 40 |
| " 80 | 45 |

(Ans : 42.86)

16. From the information given below calculate mode :

| Profits (Rs.) | No. of Companies |
|---|---|
| 5000 - 5500 | 8 |
| 5500 - 6000 | 10 |
| 6000 - 6500 | 20 |
| 6500 - 7000 | 25 |
| 7000 - 7500 | 30 |
| 7500 - 8000 | 18 |
| 8000 - 8500 | 14 |
| 8500 - 9000 | 12 |
| 9000 - 9500 | 8 |
| 9500 - 10000 | 5 |

(Ans : 7147.06)

17. Find out mode from the following data.

| Class Interval | Frequency |
|---|---|
| 0 - 5 | 6 |
| 5 - 10 | 10 |
| 10 - 15 | 12 |
| 15 - 20 | 14 |
| 20 - 25 | 18 |
| 25 - 30 | 20 |
| 30 - 35 | 26 |
| 35 - 40 | 15 |
| 40 - 45 | 12 |
| 45 - 50 | 8 |

(Ans : 31.76)

18. From the following information locate mode graphically locate and verify the results against calculated value:

| Class interval | Frequency |
|---|---|
| 0 - 10 | 15 |
| 10 - 20 | 25 |
| 20 - 30 | 40 |
| 30 - 40 | 50 |
| 40 - 50 | 65 |
| 50 - 60 | 70 |
| 60 - 70 | 60 |
| 70 - 80 | 30 |

19. Find arithmetic mean if the value of

Median = 37 & Mode = 42

(Ans : 34.5)

## 14.9 RECOMMENDED BOOKS


1. Gupta, S.P. : "Statistical Methods", Sultan Chand & Company, New Delhi.
2. Gupta, B.N : "Statistics", Sahitya Bhavan, Agra.
3. Gupta, S.C : "Fundamentals of Statistics", Himalaya pub. House, Bombay.
4. Simpson and kafka : "Basic Statistics", Oxford and IBH Publishing Company, Calcutta.


## 14.0 GLOSSARY

1. Bi-modal distribution : When a distribution consists of two modes, it is called bi-modal distribution.
2. Mode : The item which occurs more frequently in a distribution is called 'Mode'.
3. Multi-modal distribution : When a distribution contains more than two modes, it is called multi-modal distribution.
4. Uni-modal distribution : When a distribution possesses one mode, it is called uni-modal distribution.

## Unit-15: GEOMETRIC AND HARMONIC MEAN

**Contents**





## 15.0 AIMS AND OBJECTIVES

This Unit aims at explaining the meaning, computation, merits and limitations of geometric mean and harmonic mean and also the relationship among the averages.

After going through this unit, you should be able to :

— Explain the meaning of geometric mean and harmonic mean

— List the merits and limitations of both geometric and harmonic mean

— Compute geometric mean and harmonic mean

— Identify the relationship among the averages.

## 15.1 INTRODUCTION

Both geometric mean and harmonic mean are based on mathematical calculations. Geometric mean is the 'n'th root of the product of 'n' items in a given series. Harmonic mean is the reciprocal of arithmetic mean of the reciprocals of the given observations. Let us see the theoretical and practical aspects of these two.

## 15.2 MEANING OF GEOMETRIC MEAN

According to Freund and Williams, 'Given a set of numbers $X_1$ , $X_2$, $X_3$,.....$X_n$, the geometric mean is the 'n'th root of their products'.

Symbolically,

$$G.M. = n\sqrt{X_1.X_2.X_3.....X_n}$$

Where,

$X_1$, $X_2$, $X_3$, .....$X_n$ are the various values of a series. when the number of values is three or more for which geometric mean is to be calculated, computation of the product of various values, and finding out the value of nth root becomes a difficult task. Hence geometric mean is usually calculated with the help of logarithms.

## 15.3 COMPUTATION OF GEOMETRIC MEAN SIMPLE AND WEIGHTED

Geometric mean is computed for individual, discrete and continuous series.

**Individual Series**

In the case of individual series geometric mean is computed with the help of the following steps.

i) Find out logarithms of the values of X variable and obtain their total i.e., $\Sigma$ logX.

ii) Obtain the total number of observations i.e.,N.

iii) Apply the following formula and obtain G.M.

$$\text{G.M.} = \text{Antilog } \frac{\Sigma log X}{N}$$

**Where,**

G.M. = Geometric mean

$\Sigma$ log X = Sum of logarithms of values of X variable

N = Number of observations.

**Illustration - I**

Compute the geometric mean from the following data.

2000, 200, 20, 12, 10, 8, 4

**Solution :**

**Computation of Geometric Mean**

| X | Log X |
|---|---|
| 2000 | 3.3010 |
| 200 | 2.3010 |
| 20 | 1.3010 |
| 12 | 1.0792 |
| 10 | 1.0000 |
| 8 | 0.9031 |
| 4 | 0.6021 |
| N = 7 | Σ log X = 10.4874 |

$$\text{G.M.} = \text{Antilog } \frac{\Sigma log X}{N}$$

Here,

$$\Sigma \log X = 10.4874; \qquad N = 7$$

Substituting the values in the formula,

$$\text{G.M.} = \text{Antilog } \frac{10.4874}{7}$$

$$= \text{Antilog } 1.4982$$

$$\therefore \quad \text{Geometric mean} = 31.49$$

**Illustration - II**

Find out geometric mean from the following data :

52340, 2420, 340, 78, 9, 0, 8, 0.8, 0.04, 0.009

**Solution:**

Computation of Geometric Mean

| X | log X |
|---|---|
| 52340 | 4.7188 |
| 2420 | 3.3838 |
| 340 | 2.5315 |
| 78 | 1.8921 |
| 9 | 0.9542 |
| 0.8 | $\bar{1}.9031$ |
| 0.04 | $\bar{2}.6021$ |
| 0.009 | $\bar{3}.9542$ |
| N = 8 | Σ log X = 9.9398 |

$$\text{G.M.} = \text{Antilog } \frac{\Sigma log X}{N}$$

Here, $\Sigma \log X = 9.9398$; $N = 8$

Substituting the values in the formula,

$$G.M. = \text{Antilog } \frac{9.9398}{8}$$

$$= \text{Antilog } 1.2425$$

$$= 17.48$$

$\therefore$ Geometric mean = 17.48

**Illustration-III**

Find out geometric mean for the following distribution:

.0875, .8945, 0083, 0006, .03458

**Solution :**

COMPUTATION OF GEOMETRIC MEAN

| X | log X |
|---|---|
| 0.0875 | $\bar{2}.9420$ |
| 0.8945 | $\bar{1}.9515$ |
| 0.0083 | $\bar{3}.9191$ |
| 0.0006 | $\bar{4}.7782$ |
| 0.03458 | $\bar{2}.5388$ |
| N = 5 | $\Sigma \log X = \bar{8}.1296$ |

$$G.M. = \text{Antilog } \frac{\Sigma log X}{N}$$

Here,

$\Sigma \log X = \bar{8}.1296$; $N = 5$

Since, $\Sigma \log X = \bar{8}.1296$, the value of characteristic is negative. But the value of mantissa is positive. As such it is not possible to divide the values of characteristic and mantissa together with the denominator i.e., 5. Hence 2 is added to 8 to make it convenient to divide it with 5. At the same time, to neutralise the effect of adding 2 to characteristic, 2 is added to mantissa.

Substituting the values in the formula,

$$G.M. = \text{Antilog } \frac{\overline{10}+2.1296}{5}$$

$$= \text{Antilog } \bar{2} + .4259$$

$$= \text{Antilog } \bar{2}.4259$$

$\therefore$ Geometric mean = 0.02667

**Check your progress - 1** Find out the geometric mean of the following data

X 14 20 25 30

**Discrete Series**

The following is the procedure for calculating the geometric mean in discrete series.

i) Find out logarithms for the values of X variable and denote it by log X.

ii) Multiply the logarithms of the values of X variable with their frequencies and obtain their total.

iii) Obtain the total of frequency and denote it by N.

iv) Apply the formula given below to obtain the value of geometric mean.

$$G.M. = \text{Antilog } \frac{\Sigma f \log X}{N}$$

Where,

G.M. = Geometric mean

$\Sigma f \log X$ = Sum of the product of logarithm values of X variable and their respective frequencies.

N = Total number of observations.

**Illustration-IV**

Find out geometric mean from the following data.

| Profits per shop (Rs. in '000 s ) | : | 7 | 14 | 28 | 35 | 85 | 123 | 185 |
|---|---|---|---|---|---|---|---|---|
| No. of shops | : | 6 | 9 | 14 | 25 | 17 | 9 | 3 |

**Solution :**

**COMPUTATION OF GEOMETRIC MEAN**

| Profits per shop (Rs. in '000 s ) | No. of shops | log X | f log X |
|---|---|---|---|
| 7 | 6 | 0.8451 | 5.0706 |
| 14 | 9 | 1.1461 | 10.3149 |
| 28 | 14 | 1.4472 | 20.2608 |
| 35 | 25 | 1.5441 | 38.6025 |
| 85 | 17 | 1.9294 | 32.7998 |
| 123 | 9 | 2.0899 | 18.8091 |
| 185 | 3 | 2.2672 | 6.8016 |
| | N = 83 | | $\Sigma$ f log X = 132.6593 |

$$G.M. = \text{Antilog } \frac{\Sigma f \log X}{N}$$

Here,

$$N = 83;\ \Sigma f \log X = 132.6593$$

Subtituting the values in the formula,

$$G.M. = \text{Antilog } \frac{132.6593}{83}$$

$$= \text{Antilog } 1.5983$$

$$\therefore \quad G.M. = 39.70$$

**Continuous Series**

In continuous series, the following steps are required for computing the geometric mean.

i) Find out mid-values of the classes of the variable and denote it by m.

ii) Obtain logarithms for the mid-value of the variable and denote it by log m.

iii) Multiply the logarithms of mid-values with their respective frequencies and obtain the total i.e. $\Sigma$ f log m.

iv) Obtain the total of frequency i.e., N

v) Apply the following formula and obtain geometric mean.

$$G.M. = \text{Antilog } \frac{\Sigma f \log m}{N}$$

Where,

G.M. = Geometric mean

m = Mid-value of the class

log m = Logarithms of the mid-values of the variable.

$\Sigma$f log m = Sum of logarithms of the mid-values of the variable multiplied with their corresponding frequencies.

N = Number of observations.

**Illustration-V**

Calculate geometric mean from the following data:

| Class interval | : | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 |
|---|---|---|---|---|---|---|---|
| Frequency | : | 10 | 12 | 7 | 14 | 8 | 6 |

**Solution:**

COMPUTATION OF GEOMETRIC MEAN

| Class interval | Frequency | mid-value m | log m | f log m |
|---|---|---|---|---|
| 10-20 | 10 | 15 | 1.1761 | 11.7610 |
| 20-30 | 12 | 25 | 1.3979 | 16.7748 |
| 30-40 | 7 | 35 | 1.5441 | 10.8087 |
| 40-50 | 14 | 45 | 1.6532 | 23.1448 |
| 50-60 | 8 | 55 | 1.7404 | 13.9232 |
| 60-70 | 6 | 65 | 1.8129 | 10.8774 |
| | N = 57 | | | Σ f log m = 87.2899 |

$$\text{G.M.} = \text{Antilog } \frac{\Sigma f \log m}{N}$$

Here,

$$N = 57;\ \Sigma \text{ f log m} = 87.2899$$

Substituting the values in the formula,

$$\text{G.M.} = \text{Antilog } \frac{87.2899}{57}$$

$$= \text{Antilog } 1.5314$$

$$\therefore \quad \text{G.M.} = 33.99$$

**Computation of Weighted Geometric Mean**

Like weighted arithmetic mean we can also compute weighted geometric mean.

The following steps are required for computing the weighted geometric mean.

i) Find out logarithm values of X variable and denote it by log X.

ii) Multiply the logarithm values of X variable with their respective weights and obtain the total and denote it by ΣWlogX.

iii) Obtain the total of weights and denote it by ΣW

iv) Apply the formula given below to obtain weighted geometric mean.

$$\text{W.G.M.} = \text{Antilog } \frac{\log X_1 \times W_1 + \log X_2 \times W_2 + \ldots \log X_n \times W_n}{W_1 + W_2 + \ldots\ldots W_n}$$

Thus,

$$\text{W.G.M.} = \text{Antilog } \frac{\Sigma W \log X}{\Sigma W}$$

Where,

W.G.M. = Weighted Geometric mean

W = Weights

log X = Logarithm values of X variable

ΣWlogX = Sum of logarithm values of X variable multiplied with their corresponding weights.

ΣW = Total weights.

**Illustration-VI**

Compute geometric mean from the following information.

| Items | Index Numbers | Weights |
|---|---|---|
| M | 130 | 12 |
| N | 140 | 20 |
| O | 135 | 10 |
| P | 280 | 16 |
| Q | 193 | 8 |

**Solution:**

| Items | Index Numbers X | Weights W | logX | WlogX |
|---|---|---|---|---|
| M | 130 | 12 | 2.1139 | 25.3668 |
| N | 140 | 20 | 2.1461 | 42.9220 |
| O | 135 | 10 | 2.1303 | 21.3030 |
| P | 280 | 16 | 2.4472 | 39.1552 |
| Q | 193 | 8 | 2.2856 | 18.2848 |
| | | $\Sigma N = 66$ | | $\Sigma W log X = 147.0318$ |

$$G.M. = Antilog \frac{\Sigma W log X}{\Sigma W}$$

Here,

Σ W = 66; Σ W log X = 147.0318

Substituting the values in the formula.

$$G.M. = Antilog \frac{147.0318}{66}$$

$$= Antilog 2.2277$$

∴ G.M. = 169

## 15.4 MERITS AND LIMITATIONS OF GEOMETRIC MEAN

**Merits**

The following are the merits of geometric mean:

i) Geometric mean is considered to be the typical value of the series as its computation is based on each and every item of the series.

ii) It is a rigidly defined measure of central tendency.

iii) Unlike arithmetic mean, it is not affeced by the extreme items of the series as it gives less weight to large items and more weight to small items. This is the reason why geometric mean of a series is never larger than the arithmetic mean. But sometimes it may be equal to arithmetic mean.

iv) Geometric mean is amenable for algebraic treatment. For example, if the geometric mean of two or more series and their number of items are given a combined geometric mean of all the series can be computed.

v) Geometric mean provides a satisfactory measure of computing the average rate of change in the values of a frequency distribution.

**Limitations**

Geometric mean suffers from the following limitations:

i) It is neither easy to calculate nor simple to understand. It is rather difficult particularly to calculate when the items of a series are very large.

ii) It is not possible to compute geometric mean when the given series contains both positive and negative values.

iii) Quite often, the value of geometric mean is altogether a different value which may not be found in the series.

iv) Since it gives more weight to samll items of the series, it is not useful for the analysis of certain economic problems such as study of income levels, study of level of living standards etc. In such studies, disparities between small and large items will have to be brought out clearly.

Despite its limitations, geometric mean is widely used in finding out the average percent change in sales, production, population and other business and economic variables. It is widely used in the construction of index numbers. Since it gives equal weight to equal ratio of change it satisfies the time reversal test in index numbers.

## 15.5 PROPERTIES OF GEOMETRIC MEAN

i) If each individual value of the series is substituted with the value of geometric mean, their product is equal to the product of the values of the series. For example, if the geometric

mean of 4, 8, 16, is 8 and if the individual values are replaced with geometric mean i.e., 8, it will be

$$4 \times 8 \times 16 = 8 \times 8 \times 8$$

ii) The sum of positive and negative deviations of the logarithms of original values of a series taken from the logarithms of the geometric mean is equal to Zero.

## 15.6 MEANING OF HARMONIC MEAN

According to F.C.Mills, "the harmonic mean of series of members is the reciprocal of the arithmetic mean of the reciprocals of the individual numbers".

Symbolically,

$$H.M. = \frac{N}{\left(\frac{1}{X_1} + \frac{1}{X_2} + \ldots\ldots\frac{1}{X_n}\right)}$$

or

$$= \frac{N}{\Sigma\left(\frac{1}{X}\right)}$$

Where,

H.M. = Harmonic mean

N = Number of observations

$\frac{1}{X_1}, \frac{1}{X_2} \ldots\ldots \frac{1}{X_n}$ denote the reciprocals of the values of X variable.

## 15.7 COMPUTATION OF HARMONIC MEAN - SIMPLE AND WEIGHTED

Harmonic mean is computed from individual, discrete and continuous series.

**Individual Series**

The following steps are required to compute harmonic mean:

i) Find out the reciprocal values of X variable and denote it by $\frac{1}{X}$.

ii) Obtain the total of $\frac{1}{X}$ value i.e., $\Sigma\frac{1}{X}$.

iii) Obtain the number of observations i.e., N.

iv) Apply the formula and obtain the harmonic mean.

Harmonic mean is computed with the help of the following formula:

$$H.M. = \frac{N}{\Sigma\frac{1}{X}}$$

Where,

H.M. = Harmonic mean

N = Number of observations

$\Sigma\frac{1}{X}$ = Sum of the reciprocal values of variable X.

**Illustration-VII**

Following are the heights of 10 students in a class. Compute the harmonic mean of heights.

55", 56", 60", 61", 64", 67", 68", 69", 70", 72",

**Solution:**

**COMPUTATION OF HARMONIC MEAN**

| Heights in inches X | Reciprocal Values $\frac{1}{X}$ |
|---|---|
| 55 | 0.0182 |
| 56 | 0.0178 |
| 60 | 0.0167 |
| 61 | 0.0163 |
| 64 | 0.0156 |
| 67 | 0.0149 |
| 68 | 0.0147 |
| 69 | 0.0145 |
| 70 | 0.0143 |
| 72 | 0.0139 |
| N = 10 | $\Sigma\frac{1}{X}$ = 0.1569 |

$$H.M = \frac{N}{\Sigma\frac{1}{X}}$$

Here,

$N = 10$ ; $\Sigma\frac{1}{X} = 0.01569$

Substituting the values in the formula,

$$H.M. = \frac{10}{0.1569}$$

$$H.M. = 63.73$$

**Discrete Series**

In the case of descrete series, the following steps are required to compute harmonic mean.

i) Divide frequencies of X variable with their respective values and obtain the total i.e., $\Sigma\frac{f}{X}$

ii) Obtain the total frequencies i.e., N.

iii) Apply the following formula and obtain harmonic mean.

$$H.M. = \frac{N}{\Sigma(f \times \frac{1}{X})} = \frac{N}{\Sigma(\frac{f}{X})}$$

Where,

H.M. = Harmonic mean

N = Number of observations

$\Sigma(f \times \frac{1}{X})$ or $\Sigma(\frac{f}{X})$ = Sum of reciprocals of the values of X variable multiplied with their corresponding frequencies.

**Illustration - VIII**

Calculate the harmonic mean from the following:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Size | : 6 | 10 | 21 | 27 | 30 | 45 | 60 |
| Frequency | : 5 | 15 | 25 | 30 | 10 | 8 | 4 |

**Solution :**

**COMPUTATION OF HARMONIC MEAN**

| Size | Frequency (f) | $\frac{f}{X}$ |
|---|---|---|
| 6 | 5 | 0.8333 |
| 10 | 15 | 1.5000 |
| 21 | 25 | 1.1904 |
| 27 | 30 | 1.1111 |
| 30 | 10 | 0.3333 |
| 45 | 8 | 0.1777 |
| 60 | 4 | 0.0666 |
| | N = 97 | $\Sigma\frac{f}{X}$ = 5.2124 |

Here,

N = 97, and $\Sigma\frac{f}{X}$ = 5.2124

Substituting the values in the formula.

$$H.M. = \frac{97}{5.2124}$$

$$H.M. = 18.61$$

**Continuous Series**

In continuous series, the following steps are required to calculate harmonic Mean.

i) Obtain mid-values of all the classes and denote the column by 'm'

ii) Divide the frequencies with their respective mid-values and obtain the total i.e., $\Sigma\frac{f}{m}$.

iii) Obtain the total frequencies i.e., N.

iv) Apply the following formula and obtain the harmonic mean.

$$H.M. = \frac{N}{\Sigma(f \times \frac{1}{m})} = \frac{N}{\Sigma(\frac{f}{m})}$$

Where,

H.M. = Harmonic mean

$\Sigma f \times \frac{1}{m}$ = Sum of reciprocals of the mid-values of the variable and multiplied with their respective frequencies.

N = Number of observations.

**Illustration - IX**

Compute harmonic mean from the following frequency distribution

| Marks | No.of students |
|---|---|
| 0-10 | 7 |
| 10-20 | 10 |
| 20-30 | 12 |
| 30-40 | 21 |
| 40-50 | 25 |
| 50-60 | 20 |
| 60-70 | 17 |
| 70-80 | 6 |

**Solution :**

**COMPUTATION OF HARMONIC MEAN**

| Marks | No. of Students f | m | $\frac{f}{m}$ |
|---|---|---|---|
| 0-10 | 7 | 5 | 1.4000 |
| 10-20 | 10 | 15 | 0.6666 |
| 20-30 | 12 | 25 | 0.4800 |
| 30-40 | 21 | 35 | 0.6000 |
| 40-50 | 25 | 45 | 0.5555 |
| 50-60 | 20 | 55 | 0.3636 |
| 60-70 | 17 | 65 | 0.2615 |
| 70-80 | 6 | 75 | 0.0800 |
| | N = 118 | | $\Sigma(\frac{f}{m}) = 4.4072$ |

$$H.M. = \frac{N}{\Sigma\left(\frac{f}{m}\right)}$$

**Here,**

N = 118; $\Sigma\frac{f}{m} = 4.4072$

Substituting the values in the formula,

$$H.M. = \frac{118}{4.4072}$$

$$H.M. = 26.77$$

**Computation of Weighted Harmonic Mean**

Like weighted arithmetic mean, we can also compute weighted harmonic mean. The following steps are required to calculate the weighted harmonic mean.

i) Find out the total weights i.e., $\Sigma W$.

ii) Divide weights of X variable with their respective values and obtain the total i.e., $\Sigma\frac{W}{X}$

iii) Apply the formula and obtain weighted harmonic mean i.e., H.Mw

Weighted harmonic mean is computed with the help of the following formula :

$$\text{H.Mw} = \frac{W_1+W_2+\ldots W_n}{(1/X_1+W_1)+(1/X_2+W_2)+\ldots(1/X_n+W_n)}$$

Thus,

$$H.Mw = \frac{\Sigma W}{\Sigma(\frac{W}{X})}$$

Where,

H.Mw = Weighted harmonic mean

$\Sigma W$ = Sum of weights

$\Sigma\frac{W}{X}$ = Sum of reciprocals of values of variable X multiplied with their respective weights.

**Illustration - X**

A traveller covers his first 225 Kms at an average speed of 5 Kms per hour, next 75 Kms at an average speed of 4 Kms per hour and last 100 Kms at an average speed of 3 Kms per hour. Find out the average speed for the entire journey.

**Solution :**

Since the traveller covered the journey with varying speeds it is appropriate to compute weighted harmonic mean.

COMPUTATION OF WEIGHTED HARMONIC MEAN

| Speed X | Distance in Kms. W | $\frac{W}{X}$ |
|---|---|---|
| 5 | 225 | 45 |
| 4 | 75 | 18.75 |
| 3 | 100 | 33.33 |
| | $\Sigma W = 400$ | $\Sigma\frac{W}{X} = 97.08$ |

$$H.Mw = \frac{\Sigma W}{\Sigma(\frac{W}{X})}$$

Here,

$\Sigma W = 400$ ; $\Sigma\frac{W}{X} = 97.08$

$$H.Mw = \frac{400}{97.08}$$

$$= 4.12$$

The average speed of the entire journey is 4.12 Kmph.

## 15.8 MERITS AND LIMITATIONS OF HARMONIC MEAN

**Merits**

i) Harmonic mean is rigidly defined and its computations are based on every value of the variable.

ii) It is also amenable for algebraic treatment.

iii) Since the reciprocals are averaged, harmonic mean is more suitable when greater importance is to be given to small items.

iv) Computation of harmonic mean is not affected very much by fluctuations of sampling.

v) Harmonic mean measures relative changes and it is utilised to average the time rates and price movement.

**Limitations**

1. Computation of harmonic mean involves complicated process, hence it is not understood easily.
2. While computing the harmonic mean, larger weights are given to smaller values. Hence this average is not amenable to statistical analysis.
3. When the data consists of positive and negative values, or one or more values are zero, harmonic mean cannot be computed from such data.

## 15.9 RELATIONSHIP AMONG THE AVERAGES

1. When the series are normal, symmetrical and unimodal, Mean = Mode = Median.
2. When series is positively skewed (fig : 15.1), the mean is the highest value and mode is the lowest value, and the median is about one third of the distance from the mean towards the mode.

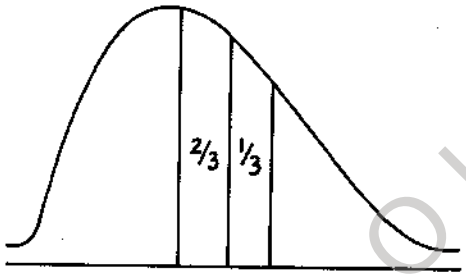



Fig : 15.1 Showing the Mean, Median and Mode in positively skewed series.

3. When the distribution is negatively skewed (fig : 15.2), the mean is the lower and the mode is the largest and the median is about one-third of the distance from the mean towards the mode.

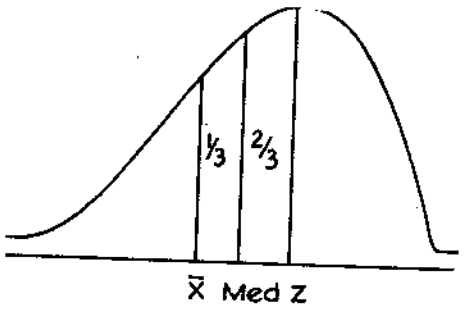



Fig : 15.2 Showing the Mean, Median and Mode in Negatively skewed distribution

4. When all the values of a variable are identical then the relationship among arithmetic mean, geometric mean and harmonic mean is Arithmetic mean = Geometric mean = Harmonic mean

5. When the values of a distribution are unequal the relationship among arithmetic mean, Geometric mean and Harmonic mean is :

   Arithmetic mean > Geometric mean > Harmonic mean

## 15.10 SUMMING UP

Geometric mean is the 'n'th root of the product of 'N' items of the series. Logarithms are used in calculating geometric mean. Geometric mean is considered to be the typical value as its computation is based on each and every item of the series. Unlike arithmetic mean, it is not affected by the extreme items of the series as it gives less weight to large items and more weight to small items. It is amenable for further algebraic treatment. However, it is very difficult to compute geometric mean where the number of items is large in the series.

Harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of individual numbers. It is generally used to average the time rates and price movements.

## 15.11 CHECK YOUR PROGRESS: MODEL ANSWERS

Obtain $\Sigma$ log X and apply the following formula.

GM = Antilog of $\frac{\Sigma log X}{N}$

The answer is 21.40

## 15.12 MODEL EXAMINATION QUESTIONS

**A . Short questions**

1. Define Geometric mean.
2. Explain the meaning of weighted geometric mean.
3. What are the properties of geometric mean ?
4. Distinguish between simple and weighted geometric mean.
5. What are the merits and limitations of geometric mean ?
6. What is meant by harmonic mean ?
7. Explain the term 'weighted harmonic mean'
8. What is the difference between simple and weighted harmonic mean ?
9. Explain the merits and limitations of harmonic mean.
10. Explain the relationship among various averages.

## EXERCISES

11. Compute geometric mean for the following data :

670 820 1020 1190 520 60 10

(Ans : 297.9)

12. Find out geometric mean for the following :

4000 400 40 0.4 0.004

(Ans : 10.05)

13. Calculate geometric mean

0.66 0.24 0.66 1.66 0.02

(Ans : 0.3222)

14. Following data gives the monthly wages of employees in a factory. Find out geometric mean.

| Wages (Rs.) : | 120 | 125 | 130 | 135 | 145 | 150 |
|---|---|---|---|---|---|---|
| Employees : | 3 | 13 | 24 | 26 | 15 | 10 |

(Ans : 134.79)

15. Data given below relates to the marks of students in a calss. Calculate geometric mean.

| Marks : | 30 | 35 | 45 | 55 | 65 | 75 |
|---|---|---|---|---|---|---|
| No of students : | 8 | 14 | 20 | 10 | 5 | 3 |

(Ans : 43.97)

16. Calculate the weighted geometric mean

| Commodity | Price (Rs.) | Weights |
|---|---|---|
| A | 120 | 200 |
| B | 105 | 300 |
| C | 130 | 500 |
| D | 100 | 150 |
| E | 160 | 140 |
| F | 80 | 200 |

(Ans : 114.63)

17. Find out geometric mean of the following series.

| Wages | No. of labourers |
|---|---|
| 50-60 | 6 |
| 60-70 | 8 |
| 70-80 | 12 |
| 80-90 | 25 |
| 90-100 | 15 |

(Ans : 79.29)

18. From the following data, compute geometric mean :

| Class interval | frequency |
|---|---|
| 0-2 | 3 |
| 2-4 | 6 |
| 4-6 | 10 |
| 6-8 | 14 |
| 8-10 | 18 |
| 10-12 | 16 |
| 12-14 | 12 |
| 14-16 | 8 |

(Ans : 7.941)

19. Calculate weighted geometric mean for the following data :

| Commodity | Price Index | Weights |
|---|---|---|
| P | 120 | 2 |
| Q | 140 | 5 |
| R | 105 | 7 |
| S | 80 | 9 |
| T | 70 | 6 |
| J | 100 | 2 |

(Ans : 94.48)

20. Find weighted geometric mean from the following table :

| Commodity | Price Index | Weignts |
|---|---|---|
| Rice | 360 | 50 |
| Wheat | 200 | 32 |
| Sugar | 160 | 20 |
| Oil Seeds | 220 | 10 |
| Barley | 150 | 8 |

(Ans : 243.4)

21. The following are the monthly earnings of 10 workers :

(Earnings in Rupees)

180, 200, 205, 150, 170, 190, 210, 185, 160, 156.

Calculate harmonic mean

(Ans : 177.62)

22. Find out the harmonic mean for the following data :

200, 110, 156, 182, 190, 210, 160, 172, 164.

(Ans : 165.75)

23. The data given below relates to the marks obtained by students in an examination.

| Marks | : | 30 | 35 | 40 | 50 | 60 | 70 | 80 |
|---|---|---|---|---|---|---|---|---|
| No. of Students | : | 8 | 12 | 30 | 25 | 18 | 6 | 4 |

Compute harmonic mean.

(Ans : 44.87)

24. The following are the daily wages of workers of a factory

| Wages (Rs.) | : | 10 | 15 | 20 | 25 | 30 | 35 | 40 |
|---|---|---|---|---|---|---|---|---|
| No. of Workers | : | 12 | 16 | 24 | 16 | 15 | 10 | 8 |

Compute harmonic mean

(Ans : 19.83)

25. From the following frequency distribution, calculate harmonic mean.

| Class Interval | Frequency |
|---|---|
| 10-20 | 6 |
| 20-30 | 4 |
| 30-40 | 10 |
| 40-50 | 12 |
| 50-60 | 18 |
| 60-70 | 15 |
| 70-80 | 5 |

(Ans : 40.29)

26. The following data relates to the marks of students in Accountancy:

| Marks | No. of students |
|---|---|
| 10-20 | 12 |
| 20-30 | 14 |
| 30-40 | 25 |
| 40-50 | 28 |
| 50-60 | 10 |
| 60-70 | 8 |
| 70-80 | 5 |

Compute harmonic mean.

(Ans : 33.25)

## 15.13 RECOMMENDED BOOKS


1. Gupta, S.P. : "Statistical Methods", Sultan Chand & Company, New Delhi.
2. Gupta, B.N. : "Statistics", Sahitya Bhavan, Agra.
3. Gupta, S.C. : "Fundamentals of Statistics", Himalaya Pub. House, Bombay.
4. Simpson and Kafka : "Basic Statistics", Oxford and I.B.H. Publishing Company, Calcutta.


## 15.14 GLOSSARY

Geometric Mean : It is the nth root of the product of 'n' items in the group.

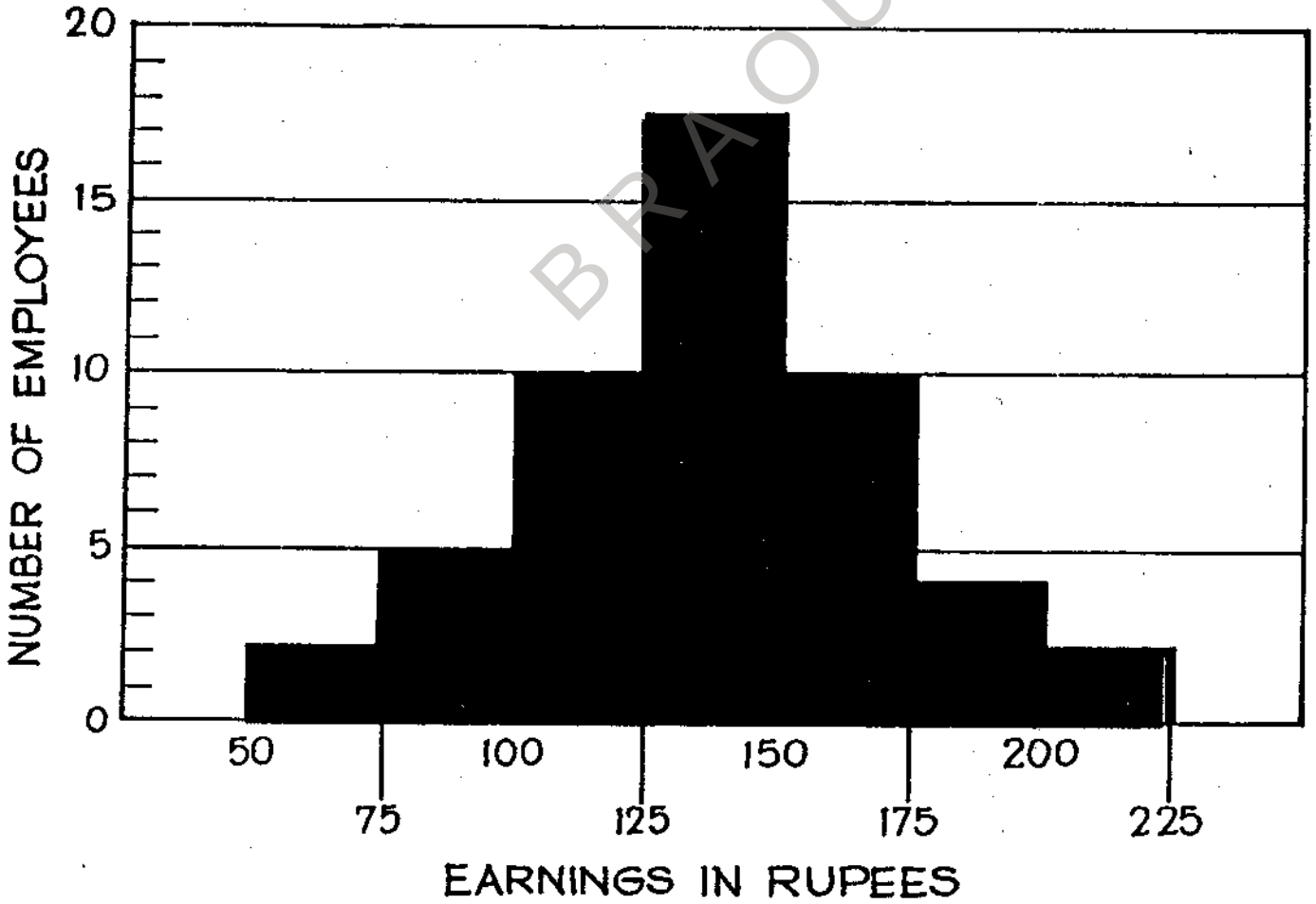
Harmonic Mean : It is the reciprocal of the arithmetic mean of the reciprocals of the given observations.

Dr. B.R. AMBEDKAR OPEN UNIVERSITY

2 - B1

# BUSINESS STATISTICS

## BLOCKS : III – V



# PREFACE

This book deals with the topics in Bussiness Statistics included in the syllabus for the second year of the B.Com. course offered by the Andhra Pradesh Open University. These topics generally cover the "core" area of the subject to be studied in the second Year of the Three Year Degree Course in Commerce (B.Com.) . The syllabus for the sake of convenience is divided into Blocks, each of which comprises a number of units. Each Block generally covers a specific area of the subject. The units are prepared by specialists in accordance with a format so designed as to enable the student to read and understand them without much difficulty. Each unit begins with contents followed by Aims and objecitives and has at its end Model Examination Questions intended to test the student's comprehension of its subject matter. Technical terms with which the student may not generally be familiar are given at the end of each unit under the head, "Glossary".

This book is concerned with the study of statistics as applicable to business in which decisions may have to be taken with regard to complex transactions and strategies in the interest of its furtherence and development. Statistics as a subject uses its own tools and techniques to analyse complex phenomena and presents its conclusions in the form of graphs and numerals. It is, therefore, of practical importance as much to businessmen as to students of commerce.

The University hoped that this meterial will help the student to get acquainted with the principal issues in Business Statistics which make for its distinctiveness and significance.

## BLOCK - III : MEASURES OF VARIATION OR DISPERSION

## UNIT-16 : DISPERSION

**Contents**





## 16.0 AIMS AND OBJECTIVES

The aim of this unit is to present the meaning, need, objectives, characteristics and types of dispersion.

After going through this unit, you should be able to :

- explain the meaning of dispersion
- identify the need for measuring the dispersion
- describe the objectives of measures of dispersion
- recognise the types and methods of dispersion
- distinguish between absolute and relative measures of dispersion.

## 16.1 INTRODUCTION

Though, the averages serve the purpose of describing the characteristics of the distribution, they cannot give a comprehensive idea and as such conceal many important facts about the distribution. One of the important aspects concealed by the measures of central tendency is regarding the variability in the values. It fails to explain the extent of deviation from the average value in the given distribution. In the absence of such information, the averages among the different distributions cannot be meaningfully compared. Comparisons can be made only in cases where all the values are the same as the average or when there are no significant deviations in

the values from the average. However, this situation is very rare especially, in case of the data pertaining to various socio-economic problems. Hence, any average by itself fails to give the complete description of the given distribution. It can be misleading, if it is not identified and accompanied by other information describing the range of things and their deviations from the central value. Some other values are required for describing the characteristics of the distribution and for making comparisons with the other distributions. As Simpson and Kafka rightly pointed out "An average does not tell the full story. It is hardly fully representative, of a mass unless we know the manner in which the individual items scatter around it. A further description of the series is necessary if we are to gauge how representative the average is". To support and supplement the measures of central tendency, other measures, like dispersion, and skewness are devised.

While the measures of dispersion explain the degree of variation in the individual items from the central value, the skewness measures the degree of symmetry or asymmetry of the distribution.

While, the measures of dispersion are discussed in this unit, skewness is dealt in the forthcoming units.

## 16.2 MEANING OF DISPERSION

'Dispersion' or 'variation' in statistics means the degree of spread of each individual item or value from the central value in the given distribution. According to Bowley it is "the measure of the variation of the items". In the words of Spiegel "the degree to which numerical data tend to spread about an average value is called the variation or dispersion of the data". The term dispersion indicates that within a given group, the items differ from one another in size, in other words there is a lack of uniformity in their sizes.

The techniques adopted to describe the variation or dispersion are called the measures of dispersion. According to John I. Griffin "A measure of variation or dispersion describes the degree of scatter shown by observations and is usually measured as an average deviation about some central value". The object of measuring dispersion is to arrive at a single summary figure which adequately demonstrates the extent of scatteredness of the variables in the distribution. The measures of dispersion are also called the average of second order, since they measure the average of the deviations taken from the central tendency of the distribution.

## 16.3 NEED FOR MEASUREMENT OF DISPERSION

The need for measurement of dispersion arises from the facts that, (a) the distribution may have similar averages, but the degree of scatteredness in the individual values is different, and (b) the distributions may have different averages, but the variation among the values in the distributions may be quite less. Hence, the difference between the averages is not important. It also arises from the fact that changes in the behaviour of a group of values are often reflected in a change in variation rather than the change in central value. The need for the measurement of dispersion, when the averages are same but the variations in the individual items are different

can better be illustrated with the following examples:

**Example : 1**

**PRODUCTION OF RICE (IN METRIC TONNES PER ACRE IN THREE FIELDS)**

| Year | Field 'A' ($X_1$) | Field 'B' ($X_2$) | Field 'C' ($X_3$) |
|---|---|---|---|
| 1974 | 10 | 8 | 4 |
| 1975 | 10 | 15 | 20 |
| 1976 | 10 | 10 | 32 |
| 1977 | 10 | 15 | 30 |
| 1978 | 10 | 17 | 12 |
| 1979 | 10 | 13 | 2 |
| 1980 | 10 | 9 | 0 |
| 1981 | 10 | 8 | 0 |
| 1982 | 10 | 2 | 0 |
| 1983 | 10 | 3 | 0 |
| Total: | $\Sigma X_1$=100 | $\Sigma X_2$=100 | $\Sigma X_3$ = 100 |

Arithmetic mean of :

Field 'A' $= \frac{\Sigma X_1}{n} = \frac{100}{10} = 10$

Field 'B' $= \frac{\Sigma X_2}{n} = \frac{100}{10} = 10$

Field 'C' $= \frac{\Sigma X_3}{n} = \frac{100}{10} = 10$

In this Example, the average production of rice in all the fields is same. But the yealry production figures differ widely from the average. In field 'A' the average production and yearly values of productilon for all the years is same. Hence, the average perfectly represents the values. In field 'B' except in 1976, the production figures for all other years show variation from the average. In field 'C' wide variations are observed in the production for all the years from the average. Hence, comparison among these three fields on the basis of their average would be misleading, because the degree of deviations of figures in field 'B' and field 'C' from their averages are altogether different in each case. Therefore, measurement of these deviations is necessary to make fruitful decisions. Sometimes, the average may be different, but the nature of variations may be same. This can be illustrated with the help of the example given below :

**Example : 2**

**PRODUCTION OF THREE FACTORIES IN '000 TONS**

| Years | | 1975 | 1976 | 1977 | 1978 | 1979 | 1980 | 1981 | 1982 | 1983 |
|---|---|---|---|---|---|---|---|---|---|---|
| FACTORIES | $X_1$ | 7 | 9 | 11 | 13 | 15 | 17 | 19 | 21 | 23 |
| | $X_2$ | 146 | 148 | 150 | 152 | 154 | 156 | 158 | 160 | 162 |
| | $X_3$ | 35 | 37 | 39 | 41 | 43 | 45 | 47 | 49 | 51 |

**Solution :**

| Year | FACTORIES | | | | | |
|---|---|---|---|---|---|---|
| | $X_1$ | Variations ($\overline{X_1} = 15$) | $X_2$ | Variations ($\overline{X}_2 = 154$) | $X_3$ | Variations ($\overline{X}_3 = 43$) |
| 1975 | 7 | −8 | 146 | −8 | 35 | −8 |
| 1976 | 9 | −6 | 148 | −6 | 37 | −6 |
| 1977 | 11 | −4 | 150 | −4 | 39 | −4 |
| 1978 | 13 | −2 | 152 | −2 | 41 | −2 |
| 1979 | 15 | 0 | 154 | 0 | 43 | 0 |
| 1980 | 17 | +2 | 156 | +2 | 45 | +2 |
| 1981 | 19 | +4 | 158 | +4 | 47 | +4 |
| 1982 | 21 | +6 | 160 | +6 | 49 | +6 |
| 1983 | 23 | +8 | 162 | +8 | 51 | +8 |
| Total | $\Sigma X_1 = 135$ | $= 0$ | $\Sigma X_2 = 1386$ | $= 0$ | $\Sigma X_3 = 387$ | $= 0$ |

Arithmetic mean production of

Factory $X_1$ $= \frac{\Sigma X_1}{N} = \frac{135}{9} = 15$

Factory $X_2$ $= \frac{\Sigma X_2}{N} = \frac{1386}{9} = 154$

Factory $X_3$ $= \frac{\Sigma X_3}{N} = \frac{387}{9} = 43$

In the above example,the average production of the factories P, Q and R is 15,000 tons, 1,54,000 tons and 43,000 tons respectively. Though the averages differ with one another, variations are similar in nature. If the nature of variations in the time series is ignored, one may conclude that the factory "Q" is efficient as it shows a higher average value. But it could be a hasty conclusion, as the variations in the production figures are occuring in the same manner for all the three factories. Hence, the measurement of the variations is significant to arrive at appropriate conclusions.

## 16.4 OBJECTIVES OF MEASURES OF DISPERSION

Measures of dispersion are studied with the following objectives:

*(i) To determine the reliability of an average :* An average obtained from homogeneous set of observations is considered to be representative and reliable. Measures of dispersion help to study the representativeness and the reliability of the average. When the dispersion is small, it can be considered that greater uniformity is ensured in the distribution and the average is considered to be fairly representative and reliable. But greater value of dispersion indicates that the average is unreliable and not a representataive one. Same opinion is expressed by W.A.Spurr and C.P.Bonini and according to them, "When dispersion is small, the average is a typical value in that it closely represents the individual values and it is reliable in that it is a good estimate of the corresponding average in the population. On the other hand, when the dispersion is great, the average is not so typical and unless the sample is very large, the average may be unreliable".

*(ii) To Analyse the nature and causes of dispersion:* Measurement of variation is an effective technique for analysing the nature of variation and detecting the causes for such variations in the given set of observations. It also helps to determine whether the variation is due to random causes or due to assignable causes in the original values. Thus it serves as a mechanism to control variations that arise due to assignable reasons.

*(iii) To compare the variability among the distributions:* Two or more sets of observations are comparable when they are consistent. Measures of dispersion serve as an effective tool for comparisons with regard to variability among the distributions. A high degree of dispersion shows lack of uniformity and the low degree of dispersion acounts for more uniformity and consistency. If, for example, the prices of a commodity over a period of time are to be compared, less variations in the prices denote more uniformity and representativeness and vice-versa.

*(iv) To facilitate the use of other statistical measures :* Dispersion or variation is very useful in the aplication of advanced statistical techniques like correlation, regression, analysis of time series, tests of significance, etc. It is also useful in determining the effective production control techniques.

## 16.5 CHARACTERISTICS OF A SATISFACTORY MEASURE OF DISPERSION

Since measurement of dispersion occupies an important place in various statistical investigations, it is necessary to use an appropriate and satisfactory measure of dispersion. For this it should possess the following characteristics :

*(i) It should be simple to understand and easy to calculate :* A satisfactory measure of dispersion should be so simple as to be readily comprehensible and should not require lengthy computations. As it is widely used on account of its simplicity.

*(ii) It must be regidly defined :* It should be rigidly defined so as to ensure the same interpretation at all times and places. This avoids confusion and ambiguity relating to its

meaning. If it is not so, it may be influenced by the personal bias of the statistician. As far as possible it must be algebraically defined to arrive at uniform computations.

*(iii) It should be based on all the values in the distribution :* Any measure of dispersion is considered to be fully representative, when its computations are based on all the values. Further, it should not be affected by the extreme values in the distribution. If small or large items influence the value of dispersion, its usefulness is lessened.

*(iv) It must be suitable for further algebraic treatment :* Since the measures of dispersion are used in various fields to facilitate the application of advanced statistical cost and production control tehniques, it must be amenable to further computations.

*(v) It must possess sampling stability :* Any value of dispersion which is not affected by the variations in sampling is considered to be a satisfactory measure. In other words, the values should be uniform for all the samples drawn from same population.

However, it is to be remembered that while these characteristics nevertheless guide to gauge the representativeness of the measures of dispersion, it must be compatible with the nature and scope of the investigation, the degree of accuracy desired, etc.

**Check Your progress-1**

Listout the characteristics of a satisfactory measure of dispersion.

## 16.6 TYPES AND METHODS OF MEASURES OF DISPERSION

There are two types of measures of dispersion. They are (i) distance measures and (ii) measures of deviations. The distance measures, also known as positional measures, study the dispersion in terms of distance between the value of selected observations. The measures of deviations are employed to study the variations from the respective measures of central tendency. The two types of dispersions - viz., distance measures and deviations can be studied by (a) numerical methods and (b) graphical methods. Numerical methods explain the dispersion in numbers. They include (i) Range, (ii) Quartile deviation, (iii) Mean deviation and (iv) Standard deviation. Graphic method presents the dispersion by graphic means. The important graphic method used to show the dispersion is Lorenz curve.

## 16.7 ABSOLUTE AND RELATIVE MEASURES OF DISPERSION

The measures of dispersion may be absolute or relative. Abosolute measures of dispersion are expressed in the same statistical unit in which the original values in the distribution are expressed. This is useful to assess the absolute magnitude of variability. It is used to compare two or more distributions which are expressed in homogeneous units. Relative measures of dispersion also known as 'coefficients' are considered to be pure numbers, independent of the

units of measurement. These measures are expressed as ratio of absolute dispersion to the appropriate average used for the measurement of variation. They can also be expressed in terms of percentages. Relative measures are used in comparing the distributions expressed in different statistical units of measurement and the relative accuracy of the data.

Detailed study of these methods of dispersion is dealt in the subsequent units.

## 16.8 SUMMING UP

Dispersion is the degree of variation of individual items in the distribution from the central value. The techniques used to measure such variations are called measures of dispersion or variation. They help to determine the reliability of an average and compare the variability among distributions. Despersion is studied with the help of measures like Range, Inter Quartile Range, Quartile Deviation, Mean Deviation, Standard Deviation and Lorenz Curve. Despersion can be measured both in absolute and relative terms (coefficients) while an absolute value is used to compare the distributions expressed in homogeneous statistical units, the relative measures are used to compare the distributions expressed in different statistical units. A low value of a measure of dispersion indicates uniformity and consistency of data and vice-versa.

## 16.9 CHECK YOUR PROGRESS : MODEL ANSWERS

1. The characteristics of a satisfactory, measure of dispersion are :

   i) It should be simple to understand and easy to calculate

   ii) It must be regidly defined

   iii) It should be based on all the values of distribution

   iv) It should be amenable for further algebraic treatment

   v) It must possess the sampling stability



## 16.10 MODEL EXAMINATION QUESTIONS

**A. short Questions**

1. Define dispersion.
2. What is relative measure of dispersion?
3. What is absolute measure of dispresion?
4. What are the types of measures of dispersion?
5. What is meant by distance measures of dispersion ?
6. Explain the need for measurement of dispersion ?
7. What are the objectives of measurement of dispersion ?
8. Explain the characteristics of dispersion.

**B. Essay Qestions**

9. What do you understand by absolute and relative measures of dispersion ? Explain their importance.

10. What are the requisites of a satisfactory measure of dispersion ? Examine in their light any two common measures of dispersion.

## 16.11 RECOMMENDED BOOKS


1. Gupta, S.P. : "Statistical Methods", sultan chand & Company, New Delhi.
2. Gupta, B.N. : "Statistics", sahitya Bhavan, Agra.
3. Gupta, S.C. : "Fundamentals of statistics", Himalaya Publishing House, Bombay.
4. Simpson and Kafka : "Basic Statistics", Oxford and I.B.H. publishing company, Calcutta.


## 16.12 GLOSSARY

| | | |
|---|---|---|
| Absolute Measures of dispersion | : | They express the dispersion in terms of original units of data. |
| Dispersion or variation | : | It refers to the degree of spread of each individual item or value from the central value in a given distribution. |
| Measure of dispersion or variation | : | A measure that is used to express the scatteredness or homogenity of data is called measure of dispersion or measure of variation. |
| Relative Measure of dispersion | : | They are pure numbers, independent of units of measurement. These are also called coefficients. |

# UNIT-17 : RANGE AND QUARTILE DEVIATION

**Contents**



## 17.0 AIMS AND OBJECTIVES

This unit aims at explaining the range, and also methods of its calculation, uses, merits and limitations. This unit also deals with quartile deviation, its calculation, merits and limitations.

After going through this unit, you should be able to:

- explain the meaning of range
- compute the range
- list the merits and limitations of range
- explain the utility of range
- calculate inter-quartile range
- describe the meaning of quartile deviation
- calculate quartile deviation
- list the merits and limitation of quartile deviation

## 17.1 INTRODUCTION

In unit 16 we have given you an overview of measures of dispersion. In this unit we describe about two such measures of dispersion viz., range and quartile deviation.

The simplest measure of variation is range. It measures the dispersion by taking the two extreme values i.e., highest and lowest values of a given series.

Unlike range, quartile deviation takes in to account the middle 50% of items. These 50% of items are represented by upper quartile ($Q_3$) and lower quartile ($Q_1$) of a given data. Then the difference between $Q_3$ and $Q_1$ is known as inter-quartile range and when such difference is divided by two, it is called quartile deviation. Let us see the theoretical and practical aspects of both range, and quartile deviation.

## 17.2 MEANING OF RANGE

Range is the simple and easy method of studying dispersion. Range can be defined as the absolute difference between the largest value and the smallest value in a given frequency distribution. Symbolically,

$R = L - S$

Where, R = Range

L = Largest value

S = Smallest value

The value of range computed with the help of the above formula is an absolute measure of dispersion. Such measure can be compared, only if all the sets of distributions are expressed in same statistical units, such as rupees, litres, meters, etc. But if different sets of distributions are expressed in different statistical units, the absolute measure of range cannot be effectively compared. For example, in Factory- 'A', the production is expressed as 'maunds' and in Factory- 'B' production is expressed as 'tons', the absolute value of range cannot be used for comparing both the distributions. For ensuring comparability, relative measure of range, known as co-efficient of range is used. This is obtained by dividing the absolute range by the sum of the largest and the smallest values (L + S) in the distribution. It is calculated by the following formula.

Coefficient of range $= \frac{L-S}{L+S}$

In the distributions having similar averages, range with a smaller value represents uniformity and consistency in the distribution. It means that the average is representative in the distribution. On the other hand, the higher value of range shows lack of uniformity and consistency in the distribution. It explains that the average is inadequate and not representative.

## 17.3 CALCULATION OF RANGE

The procedure for calculating the range and its Coefficient in individual and discrete series is explained below:

i) Locate the highest and the lowest values in the data.

ii) To obtain range, subtract the lowest value from the highest value i.e., $R = L - S$.

iii) For obtaining co-efficient of range, divide the range by the sum of the largest and the smallest values in the distribution. In the case of discrete series, the frequencies are to be ignored.

**Individual Series**

**Illustration - I:** The following are the hourly wages paid to the workers in two factories. Find out which factory is consistent in the payment of wages.

Factory-A: (Rs.) 8, 6, 12, 25, 16, 15, 18, 2.

Factory-B: (Rs.) 6, 5, 22, 28, 30, 10, 17, 13 .

**Solution:**

CALCULATION OF RANGE AND CO-EFFICIENT OF RANGE

| Factory-A (Rs.) | Factory-B (Rs.) |
|---|---|
| 8 | 6 |
| 6 | 5 |
| 12 | 22 |
| 25 | 28 |
| 16 | 30 |
| 15 | 10 |
| 18 | 17 |
| 2 | 13 |

Range $= L - S$

**Factory-A:** $L = 25, \quad S = 2$

Substituting the values in the formula,

Range (R) $= 25 - 2 = 23$

Co-efficient of Range $= \frac{L-S}{L+S}$

$$= \frac{25-2}{25+2}$$

$$= \frac{23}{27}$$

$$= 0.85$$

**Factory-B:** $L = 30$; $S = 5$

Substituting the values in the formula,

Range (R) $= 30 - 5 = 25$

Co-efficient of Range $= \frac{L-S}{L+S}$

$$= \frac{30-5}{30+5}$$

$$= \frac{25}{35}$$

$$= 0.71$$

Since the variation in hourly wages computed in terms of coefficient of Factory-'B' is smaller than that of Factory-'A', we can conclude that Factory-'B' is more consistent in the matter of payment of wages.

**Discrete Series**

**Illustration-II:** Calculate range and its coefficient from the following data relating to incomes of 400 doctors.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Income (Rs.): | 200, | 400, | 600, | 800, | 1000, | 1200, | 1400, | 1600. |
| Number of Doctors : | 20, | 25, | 26, | 27, | 80, | 88, | 94, | 40. |

**Solution:**

CALCULATION OF RANGE AND COEFFICIENT OF RANGE

| Income (Rs.) | No. of Doctors |
|---|---|
| 200 | 20 |
| 400 | 25 |
| 600 | 26 |
| 800 | 27 |
| 1000 | 80 |
| 1200 | 88 |
| 1400 | 94 |
| 1600 | 40 |

Range (R) = L − S

L = 1600

S = 200

Substituting the values in the formula.

R = 1600 − 200

= Rs. 1400

Co-efficient of Range = $\frac{L-S}{L+S}$

Substituting the values of Land S in the formula.

Co-efficient of Range = $\frac{1600-200}{1600+200}$

= $\frac{1400}{1800}$

= Rs. 0.778

Coefficient of range is high in the distribution. Hence, it is more variable and less consistent.

**Continuous Series**

In the case of continuous series, the following steps are involved in the calculation of range and co-efficient of range.

i) Take the upper limit of the class having highest value and lower limit of the class having lowest value. These two values represent the largest and the smallest values in the distribution. Ignore the frequencies.

ii) The value of range can be obtained by subtracting the lower limit of the class having lowest value from the upper limit of the class having the highest value. Thus,

Range = L − S

iii) For obtaining coefficient of range, divide the absolute value of range by the sum of largest and smallest values. Thus coefficient of range,

= $\frac{L-S}{L+S}$

**Illustration - III:** Calculate range and its co-efficient from the marks obtained by commerce students in Statistics.

| Marks in Statistics: | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 |
|---|---|---|---|---|---|---|
| No. of Students: | 6 | 8 | 7 | 11 | 10 | 17 |

**CALCULATION OF RANGE AND COEFFICIENT OF RANGE**

| Profits (in Rs. '000') | No. of Companies |
|---|---|
| 14.5 - 24.5 | 10 |
| 24.5 - 34.5 | 21 |
| 34.5 - 44.5 | 31 |
| 44.5 - 54.5 | 7 |
| 54.5 - 64.5 | 8 |

Range (R) = L − S

Class having maximum profits = 54.5 − 64.5

Class having minimum profits = 14.5 − 24.5

$$L = 64.5$$

$$S = 14.5$$

Substituting the values in the formula,

$R = 64.5 - 14.5 = 50$

Coefficient of Range $= \frac{L-S}{L+S}$

$$= \frac{64.5-14.5}{64.5+14.5}$$

$$= \frac{50}{79}$$

$$= 0.6329$$

As the value of coefficient of range is high, the distribution is more variable and less consistent.



## 17.4 MERITS AND LIMITATION OF RANGE

**(a) Merits**

i) Range is the simplest method of studying dispersion.

ii) Range can be easily calculated. It does not require lengthy calculations.

**(b) Limitations**

i) Since Range is computed by taking only two extreme values, it ignores all other values in the distribution.

ii) Its value is affected by extreme items as it does not describe all other items.

iii) It is considered to be unreliable measure of dispersion, because changes in the values of the items other than the extreme items do not alter the value of range.

iv) Often, range is influenced by the fluctuations in sampling. Its computation is not practicable in case of grouped observations which include open-end classes.

v) Range is not amenable for algebraic treatment.

On account of these limitations, the utility of range is limited. In the words of W.I King, Range is too indefinite to be used as a practical measure of dispersion.

## 17.5 UTILITY OF RANGE

Though range is a crude measure of dispersion, it has a specific significance due to its simplicity. It can be usefully applied in certain cases. For instance, it is most conveniently used in weather forecasts, where identifying the extreme values are enough for making appropriate decisions. Usually, meteorological department makes weather forecasts on the basis of minimum and maximum limits of temperature, rainfall, etc. In industries, the quality of the products manufactured will be determined by studying a sample out of the toal production. In such cases, the limits for variations will be fixed and test checks are conducted to see whether the quality is within the prescribed limits. If the variations or consistency in the pattern of production exceeds the limits, proper care will be taken to inspect the equipment. Range is normally applied in constructing quality control charts like 'R' charts and 'X' charts. Range also helps to understand the price behaviour of stocks and shares in stock markets.

## 17.6 INTER-QUARTILE RANGE

The Range, which takes into account only two extreme values, is considered a crude measure of dispersion. To overcome certain limitations of range another method known as 'inter-quartile range' has been developed. In simple words, inter-quartile range can be defined as the absolute difference between the upper and lower quartiles of a given frequency distribution. In other words, it includes only the middle 50% of the items in the distributions and ignores one quarter of the observations on either end of the distribution. The inter-quartile range is calculated with the help of the following formula.

Inter quartile range (I.R.) $= Q_3 - Q_1$

Where I.R. = Inter-quartile range

$Q_3$ = Upper Quartile

$Q_1$ = Lower Quartile

For example, if in a frequency distribution $Q_3$ and $Q_1$ are 59 and 18 respectively, its inter quartile range is as follows:

I.R. $= 59 - 18 = 41$

Inter quartile range is often used in the data pertaining to socio-economic problems and it studies the absolute difference between two quartiles.

## 17.7 MEANING OF QUARTILE DEVIATION

To overcome the limitations of range and inter quartile range, another method of dispersion called quartile deviation has been developed from inter quartile range. Quartile deviation, also known as 'semi-inter quartile range', can be defined as the average absolute difference between the upper and lower quartiles of a frequency distribution. It studies the range of spread of various items either side of the median. But, while studying the spread, it considers only 25% of the items on either side of the median and it ignores nearly 50% of the items scattered on the extreme ends of the distribution. A high degree of quartile deviation means low uniformity, and low degree of variation hence, less consistency in the formation of the distribution and vice-versa. Quartile deviation is calculated with the help of the following formula:

$$\text{Q.D} = \frac{Q_3 - Q_1}{2}$$

Where, Q.D. = Quartile deviation

$Q_3$ = Upper Quartile

$Q_1$ = Lower Quartile

Quartile deviation calculated as above is an absolute measure. This is useful in comparing the distributions expressed in homogeneous statistical units. But, for comparing the distributions expressed in different statistical units, relative measure of quartile deviation known as 'coefficient of quartile deviation' is computed with the help of the formula given below:

$$\text{Coefficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

## 17.8 CALCULATION OF QUARTILE DEVIATION

The procedure followed for the calculation of quartile deviation and the coefficient of quartile deviation is explained below:

i) Arrange the values in ascending or descending order.

ii) Obtain the values of third quartile ($Q_3$) and first quartile ($Q_1$) by applying the following formulae:

$Q_3$ = Size of 3 $\left(\frac{N+1}{4}\right)$th item

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

Where, N = Total Number of observations.

In the case of discrete and continuous series, obtain the cumulative frequencies for the location of upper and lower quartiles. Further, in case of continuous series, the following formulae are applied to find out the values of upper and lower quartiles.

$Q_3$ = Size of 3 $\left(\frac{N}{4}\right)$ th item

$Q_1$ = Size of $\left(\frac{N}{4}\right)$ th item

Since the values will be in continuous form, apply the following formula to locate the exact

value of $Q_3$

$$Q_3 = L + \frac{3(\frac{N}{4})-cf}{f} \times i$$

Where, $Q_3$ = Upper quartile

L = Lower limit of the upper quartile class

$3(\frac{N}{4})$ = Upper quartile size

cf = Cumulative frequency of the class preceeding the upper quartile class

f = Frequency of the upper quartile class

i = Class interval of upper quartile class

Apply the following formula to loacte the exact value of $Q_1$

$$Q_1 = L + \frac{(\frac{N}{4})-cf}{f} \times i$$

Where, $Q_1$ = Lower quartile

L = Lower limit of the lower quartile class

$(\frac{N}{4})$ = Lower quartile size

cf = Cumulative frequency of the class preceding the lower quartile class

f = Frequency of the lower quartile class

i = Class interval of lower quartile class

iii) To obtain quartile deviation, find out the difference between upper quartile and lower quartile and devide it by 2.

iv) For obtaining the coefficient of quartile deviation, the difference between upper quartile and lower quartile is divided by the sum of upper quartile and lower quartile.

**(A) Individual Series**

**Illustration - VI:** Below given are the passengers travelled in each bus of a travelling company. Calculate the quartile deviation and coefficient of quartile deviation.

Passengers
In each bus :43 52 37 61 41 58 72 35 65 69 75

**Solution:**

**CALCULATION OF QUARTILE DEVIATION AND COEFFICIENT OF QUARTILE DEVIATION**

The data when arranged in ascending order will read as follows:

| Number of passengers in each bus |
|---|
| 35 |
| 37 |
| 41 |
| 43 |
| 52 |
| 58 |
| 61 |
| 65 |
| 69 |
| 72 |
| 75 |

**Calculation of quartiles**

$Q_3$ = Size of 3 $(\frac{N+1}{4})$th item

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

Here N = 11

$Q_3$ = Size of 3 $(\frac{11+1}{4})$th item

= Size of 3 $(\frac{12}{4})$th item

= Size of 3(3) th item

= Size of 9th item

∴ Size of 9th item in the distribution is = 69.

$Q_1$ = Size of $(\frac{11+1}{4})$th item

= Size of $(\frac{12}{4})$th item

= Size of 3rd item

∴ Size of 3rd item in the distribution is = 41

Q.D = $\frac{Q_3-Q_1}{2}$

Here, $Q_3$ = 69 and $Q_1$ = 41

Substituting the values in the formula,

Q.D = $\frac{69-41}{2}$

= $\frac{28}{2}$

$= 14$

$$\text{Co-efficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{69-41}{69+41}$$

$$= \frac{28}{110}$$

$$= 0.255$$

Since the value of the coefficient of quartile deviation is low, there is greater uniformity in the distribution.

**Illustration - VII:** Following data relates to the number of defects per 100 trials in a manufacturing process of two machines. Find out which machine is more uniform by using quartile deviation and its coefficient.

Machine I : 2, 15, 12, 6, 16, 13, 7, 17, 13, 9, 18, 14, 10, 19, 15.

Machine II : 12, 5, 15, 12, 7, 16, 13, 8, 17, 14, 10, 19, 15, 11, 20.

**Solution:**

In order to calculate Quartile deviation and its coefficient the data should be arranged in ascending order.

**CALCULATION OF QUARTILE DEVIATION AND COEFFICIENT OF QUARTILE DEVIATION.**

| S.No | Machine I | Machine II |
|---|---|---|
| 1. | 2 | 5 |
| 2. | 6 | 7 |
| 3. | 7 | 8 |
| 4. | 9 | 10 |
| 5. | 10 | 11 |
| 6. | 12 | 12 |
| 7. | 13 | 12 |
| 8. | 13 | 13 |
| 9. | 14 | 14 |
| 10. | 15 | 15 |
| 11. | 15 | 15 |
| 12. | 16 | 16 |
| 13. | 17 | 17 |
| 14. | 18 | 19 |
| 15. | 19 | 20 |

**Machine I :**

Calculation of quartiles:

$Q_3$ = size of $3(\frac{N+1}{4})$th item

= size of $3(\frac{15+1}{4})$th item

= size of $3(\frac{16}{4})$th item

= size of 3(4)th item

therefore size of 12th item in the distribution is = 16

$Q_1$ = size of $\frac{15+1}{4}$ th item

= size of $(\frac{16}{4})$th item

therefore size of 4th item in the distribution is = 9

Q.D. = $(\frac{Q_3-Q_1}{2})$

Here,

$Q_3 = 16$ and $Q_1 = 9$

Substituting the values in the formula

Q.D. = $\frac{16-9}{2}$

= $\frac{7}{2}$

= 3.5

Coefficient of Quartile Deviation = $\frac{Q_3-Q_1}{Q_3+Q_1}$

= $\frac{16-9}{16+9}$

= $\frac{7}{25}$

= 0.28

**Machine - II:**

Calculation of quartiles:

$Q_3$ = size of $3\ \frac{N+1}{4}$ th item

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

Here, N = 15

$Q_3$ = size of $3\left(\frac{15+1}{4}\right)$ th item

= size of $3\left(\frac{16}{4}\right)$ th item

= size of 3(4)th item

∴ size of 12th item in the distribution is = 16

$Q_1$ = size of $(\frac{15+1}{4})$th item

= size of $(\frac{16}{4})$th item

∴ size of 4th item in the distribution is = 10

Q.D = $\frac{Q_3-Q_1}{2}$

Here,

$Q_3 = 16$ and $Q_1 = 10$,

Substituting the values in the formula,

Q.D = $(\frac{16-10}{2})$

= $\frac{6}{2}$

= 3

Coefficient of Quartile Deviation = $\frac{Q_3-Q_1}{Q_3+Q_1}$

= $\frac{16-10}{16+10}$

= $(\frac{6}{26})$

= 0.23

Since, the value of quartile deviation is small for both the machines, both the distributions show consistency. But, relatively Machine-II is considered to be more consistent, than Machine-I as revealed by the coefficient of quartile deviation.

**(B) Discrete Series**

**Illustration VIII:** From the following data, calculate Quartile Deviation and Coefficient of Quartile deviation.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Wages per day (in rupees) | : 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 |
| Number of persons | : 5 | 9 | 18 | 35 | 42 | 32 | 15 | 10 |

Calculation of quartiles:

$Q_3$ = size of 3 $(\frac{N+1}{4})$ th item

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

N = sum of frequencies.

Brand 'A':

Here N = 200

$Q_3$ = size of $3(\frac{200+1}{4})$ th item

= size of $3(\frac{201}{4})$ th item

= size of 3(50.25) th item

= size of 150.75th item

Since 150.75th item is located in the cumulative frequency 160.

$Q_3 = 250$.

$Q_1$ = size of $(\frac{200+1}{4})$ th item

= size of $\left(\frac{201}{4}\right)$ th item

= size of 50.25th item

Since 50.25th item is located in the cumulative frequency 60, $Q_1 = 100$.

Q.D. = $\frac{Q_3-Q_1}{2}$

Here,

$Q_3 = 250$ and $Q_1 = 100$

Substituting the values in the formula,

Q.D. = $\frac{250-100}{2}$

= $\frac{150}{2}$

= 75

Coefficient of quartile deviation,

= $\frac{Q_3-Q_1}{Q_3+Q_1}$

= $\frac{250-100}{250+100}$

= $\frac{150}{350}$

= 0.429

**Brand 'B' :**

$N = 286$

$Q_3 =$ size of $3(\frac{286+1}{4})$ th item

$=$ size of $3\left(\frac{287}{4}\right)$ th item

$=$ size of $3(71.75)$ th item

$=$ size of $215.25$ th item

Since, 215.25th item is located in the cumulative frequency 233,

$Q_3 = 250.$

$Q_1 =$ size of $(\frac{286+1}{4})$ th item

$=$ size of $\left(\frac{287}{4}\right)$ th item

$=$ size of 71.75th item

Since 71.75th item is located in the cumulative frequency 110,

$Q_1 = 150.$

$\text{Q.D.} = \frac{250-150}{2}$

$= \frac{100}{2}$

$= 50$

$\text{Coefficient of quartile deviation} = \frac{Q_3-Q_1}{Q_3+Q_1}$

$= \frac{250-150}{250+150}$

$= \frac{100}{400}$

$= 0.25$

Since the coefficient of quartile deviation of Brand 'B' bulbs low, it has more consistent lighting performance.

**(C) Continuous Series**

**Illustration-X :** Calculate quartile deviation and coefficient of quartile deviation of marks of 60 students and interpret the results.

| Marks | : | 0-15 | 15-30 | 30-45 | 45-60 | 60-75 | 75-90 |
|---|---|---|---|---|---|---|---|
| Number of Students | : | 8 | 5 | 13 | 15 | 5 | 4 |

**Solution :**

Marks shall be considered as 'X' and number of students as frequency.

**CALCULATION OF QUARTILE DEVIATION AND COEFFICIENT OF QUARTILE DEVIATION.**

| (X) | Frequency | Cumulative frequency (cf) |
|---|---|---|
| 0-15 | 8 | 8 |
| 15-30 | 5 | 13 |
| 30-45 | 13 | 26 |
| 45-60 | 15 | 41 |
| 60-75 | 5 | 46 |
| 75-90 | 4 | 50 |
| | N = 50 | |

Calculation of quartiles :

$Q_3$ = Size of $3(\frac{N}{4})$ th item

$Q_1$ = size of $(\frac{N}{4})$ th item

N = sum of the frequencies

N = 50

$Q_3$ = size of $3(\frac{50}{4})$ th item

= size of 3(12.5) th item

= size of 37.5th item

Since, size of 37.5th item is located in the cumulative frequency 41, $Q_3$ class is 45-60. To locate the exact value of $Q_3$ , the following formula is applied.

$$Q_3 = L + \frac{3(\frac{N}{4}) - cf}{f} \times i$$

L = 45

$3(\frac{N}{4}) = 37.5$

cf = 26

f = 15

i = 15

Substituting the values in the formula.

$$Q_3 = 45 + \frac{37.5-26}{15} \times 15$$

$$= 45 + \frac{11.5}{15} \times 15$$

$$= 45 + 11.5$$

$$= 56.5$$

$Q_1$ size of $\left(\frac{N}{4}\right)$ th item

$Q_1$ = size of $\left(\frac{50}{4}\right)$ th item

= size of 12.5th item

Since the size of 12.5th item is located in the cumulative frequency 13, $Q_1$ class is 15-30.

To find out the exact value of $Q_1$ the following formula is applied.

$$Q_1 = L + \frac{\frac{N}{4}-cf}{f} \times i$$

Here, L = 15

$(\frac{N}{4}) = 12.5$

cf = 8

f = 5

i = 15

Substituting the values in the formula,

$$Q_1 = 15 + \frac{12.5-8}{5} \times 15$$

$$= 15 + \frac{4.5}{5} \times 15$$

$$= 15 + 13.5$$

$$= 28.5$$

$$\text{Q.D.} = \frac{Q_3-Q_1}{2}$$

Here, $Q_3 = 56.5$ and $Q_1 = 28.5$

Substituting values in the formula.

$$\text{Q.D.} = \frac{56.5-28.8}{2}$$

$$= \frac{28}{2}$$

$$= 14$$

$$\text{Coefficient of Q.D.} = \frac{Q_3-Q_1}{Q_3+Q_1}$$

$$= \frac{56.5-28.5}{56.5+28.5}$$

$$= \frac{28}{85}$$

$$= 0.32$$

Since quartile deviation of marks (14) and coefficient of quartile deviation of marks (0.32) are small the distribution has consistency.

**Illustration XI:** From the following data obtain the quartile deviation and its coefficient.

| Class (X) | : | 18-21 | 22-25 | 26-35 | 36-45 | 46-55 |
|---|---|---|---|---|---|---|
| Frequency | : | 13 | 23 | 17 | 22 | 25 |

**Solution :**

Since class intervals are in inclusive method an adjustment of class intervals are necessary.

**CALCULATION OF QUARTILE DEVIATION AND**

**COEFFICIENT OF QUARTILE DEVIATION.**

| Class (X) | Frequency (f) | Cumulative frequency (cf) |
|---|---|---|
| 17.5 - 21.5 | 13 | 13 |
| 21.5 - 25.5 | 23 | 36 |
| 25.5 - 35.5 | 17 | 53 |
| 35.5 - 45.5 | 22 | 75 |
| 45.5 - 55.5 | 25 | 100 |
| | N = 100 | |

**Calculation of quartiles**

$Q_3$ = size of $3(\frac{N}{4})$th item

$Q_1$ = size of $(\frac{N}{4})$th item

N = 100

$Q_3$ = size of $3(\frac{100}{4})$th item

= size of 3(25)th item

= size of 75th item

Since 75th item is located in the cumulative frequency 75,

$Q_3$ class is 35.5 - 45.5.

As the $Q_3$ size is in continuous form, to locate its extact value, we must use the following formula.

$$Q_3 = L + \frac{3(\frac{N}{4}) - cf}{f} \times i$$

Here, L = 35.5

$3(\frac{N}{4}) = 75$

cf = 53

f = 22

i = 10

Substituting the values in the formula,

$Q_3 = 35.5 + \frac{75-53}{22} \times 10$

$= 35.5 + \frac{22}{22} \times 10$

$= 35.5 + 10$

$= 45.5$

$Q_1$ = size of ( $\frac{100}{4}$)th item

= size of 25th item

Since 25th item is located in the cumulative frequency 36,

$Q_1$ class is 21.5 − 25.5.

As the $Q_1$ size is continuous class, to obtain its exact value the following formula is used:

$Q_1 = L + \frac{\frac{N}{4}-cf}{f} \times i$

Here, L = 21.5

$(\frac{N}{4}) = 25$

cf = 13

f = 23

i = 4

Substituting the values in the formula.

$Q_1 = 21.5 + \frac{25-13}{23} \times 4$

$= 21.5 + (\frac{12}{23}) \times 4$

$= 21.5 + 2.08696$

$= 23.59$

Q.D.$= \frac{Q_3-Q_1}{2}$

Here,

$Q_3 = 45.5$ and $Q_1 = 23.59$

substituting the values in the formula,

Q.D.$= \frac{45.50-23.59}{2}$

$= \frac{21.91}{2}$

$= 10.95$

Coefficient of Quartile Deviation $= \frac{Q_3-Q_1}{Q_3+Q_1}$

$$= \frac{45.50-23.59}{45.50+23.59}$$

$$= \frac{21.91}{69.09}$$

$$= 0.317$$

Since quartile deviation (10.95) and coefficient of quartile deviation (0.317) are smaller, the distribution is considered to be consistent and less variable.

**Illustration XII:** Find out the consistency of the following distribution with the help of quartile deviation and coefficient of quartile deviation.

| Value Less than : | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 |
|---|---|---|---|---|---|---|---|---|
| Frequency : | 68 | 143 | 200 | 279 | 311 | 356 | 389 | 420 |

**Solution :**

Since distribution is in 'Less than' form it must be adjusted as below:

**CALCULATION OF QUARTILE DEVIATION AND COEFFICIENT OF QUARTILE DEVIATION**

| Class (X) | Frequency (f) | Cumulative Frequency (cf) |
|---|---|---|
| 0-10 | 68 | 68 |
| 10-20 | 75 | 143 |
| 20-30 | 57 | 200 |
| 30-40 | 79 | 279 |
| 40-50 | 32 | 311 |
| 50-60 | 45 | 356 |
| 60-70 | 33 | 389 |
| 70-80 | 31 | 420 |
| | N = 420 | |



**Calculation of quartiles**

$Q_3$ = size of $3(\frac{N}{4})$th item

$Q_1$ = size of $(\frac{N}{4})$th item

N = 420

$Q_3$ = size of $3\left(\frac{420}{4}\right)$ th item

= size of 3(105)th item

= size of 315th item

Since 315th item is located in the cumulative frequency 356, $Q_3$ class is 50-60. As $Q_3$ size is in continuous form, to locate its exact value the following formula is used.

$$Q_3 = L + \frac{3\frac{N}{4} - cf}{f} \times i$$

Here, L = 50

$3(\frac{N}{4}) = 315$

cf = 311

f = 45

i = 10

Substituting the values in the formula,

$$Q_3 = 50 + \frac{315-311}{45} \times 10$$

$$= 50 + \frac{4}{45} \times 10$$

$$= 50 + \frac{40}{45}$$

$$= 50 + 0.888$$

$$Q_3 = 50.888$$

$Q_1$ = size of $(\frac{420}{4})$ th item

= size of 105th item

= 105th item is located in the cumulative frequency in 143, $Q_1$ class = 10-20.

To findout the exact value of $Q_1$, the following formula is used.

$$Q_1 = L + \frac{\frac{N}{4} - cf}{f} \times i$$

Here, L = 10

$\frac{N}{4} = 105$

cf = 68

f = 75

i = 10

Substituting the values in the formula,

$$Q_1 = 10 + \frac{105-68}{75} \times 100$$

$$= 10 + \frac{37}{75} \times 10$$

$$= 10 + \frac{370}{75}$$

$$= 10 + 4.933$$

$$= 14.933$$

Q.D. $= \frac{Q_3 - Q_1}{2}$

Here, $Q_3 = 50.888$ and $Q_1 = 14.933$

Substituting the values in the formula,

$$\text{Q.D} = \frac{50.888-14.933}{2}$$

$$= \frac{35.955}{2}$$

$$= 17.98$$

$$\text{Coefficient of quartile deviation} = \frac{Q_3-Q_1}{Q_3+Q_1}$$

$$= \frac{50.888-14.933}{50.888+14.933}$$

$$= \frac{35.96}{65.82}$$

$$= 0.55$$

As the coefficient of quartile deviation is 0.55; it can be considered that the distribution is having a balanced spread of the items.

## Check your progress - 1

Calculate quartile deviatilon and coefficient of quartile deviation from the following data.

| X | 40-60 | 60- 80 | 80-100 | 100-120 | 120-140 | 140-160 |
|---|---|---|---|---|---|---|
| f | 8 | 16 | 14 | 22 | 38 | 12 |



## 17.9 MERITS AND LIMITATIONS OF QUARTILE DEVIATION

**(a) Merits**

i) Quartile deviatilon is simple to calculate and easy to understand. It is also easy to compute even in the case of distribution with open end classes.

ii) Since 50% of the observations in the distribution are taken into account it can be treated as a better measure of dispersion than range. As observed by Criffin 'The quartile deviation is a useful location based on measure of dispersion, because even though there is some departure from symmetry, approximately half the observations will be included in a range of plus and minus one quartile deviation around the median. This half of the observations

is the central half, and therefore excludes the extreme.

**(b) Limitations**

i) It does not depend upon every value in the distribution. So, it can not be considered as a satisfactory measure of dispersion.

ii) The value of quartile deviation is affected by sampling fluctuations.

iii) Quartile deviation canot be subjected to further statistical treatment.

iv) It does not reveal the extent of variability of the values.

Due to these limitations, quartile deviation is not commonly used as a measure of dispersion. To overcome some of the limitations of quartile deviation, certain other mesures of dispersion like, mean deviation, standard deviatilon have been developed. They are discussed in the subsequent units.

## 17.10 SUMMING UP

Range is the absolute difference between the largest and the smallest value in a distribution. It is generally used in weather forecast, quality control, stock matters, etc.

Interquartile range is the absolute difference between the upper and lower quartiles. The relative measure of quartile deviation is called coefficient of quartile deviation. Though it is simple to calculate, it is not a representative measure since it excludes 25% of the items on each of the extreme ends of the distribution.

## 17.11 CHECK YOUR PROGRESS : MODEL ANSWERS

1. Find out the two quartiles viz., $Q_1$ and $Q_3$ by using the following formulae.

$$Q_1 = L + \frac{\frac{N}{4} - cf}{f} \times i$$

$$Q_3 = L + \frac{3\frac{N}{4} - cf}{f} \times i$$

$Q_1 = 85$ and $Q_3 = 131.84$

$$Q.D = \frac{Q_3 - Q_1}{Q_3 + Q_1} = 0.21$$

## 17.12 MODEL EXAMINATION QUESTIONS

**A. Short Questions**

1. Define
   a) Range
   b) Inter-quartile range
   c) Quartile deviation
2. How do you compute range?
3. What are the uses of range?
4. What are the limitations of range?
5. How do you compute inter-quartile range?
6. What is Co-efficient of quartile deviation? How do you compute it?
7. How do you compute quartile deviation in discrete series?
8. Explain the merits and limitations of quartile deviation?

**EXERCISES**

9. Calculate Range and its Co-efficient from the following details of average prices relating to a commodity.

| Jan. | Feb. | Mar. | Apr. | May. | Jun. | July. | Aug. | Sep. | Oct. | Nov. | Dec. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 43 | 26 | 53 | 65 | 71 | 16 | 28 | 32 | 21 | 15 | 85 | 48 |

Ans : 0.7

10. Compute range and coefficient of range from the following data:

1150, 2436, 3261, 1656, 1763, 2556, 966

Ans : 0.54

11. Calculate Range and Coefficient of range of the marks of 100 boys

| Marks : | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 |
|---|---|---|---|---|---|---|---|---|
| Number of boys : | 8 | 11 | 16 | 25 | 36 | 2 | 1 | 1 |

Ans : 0.77

12. The following data relate to the yearly incomes of College teachers. Find out range and its coefficient.

Income
(Rs.) : 1200 1600 1800 2000 2200 2400

Number of
persons : 250 360 285 175 160 120

Ans : 0.33

13. Find out range and coefficient of range from the following data:

Class : 10-19 20-29 30-39 40-49 50- 59 60-69

Frequency : 16 30 45 56 23 10

Ans : 0.74

14. Calculate range and coefficient of range (use mid-valus):

Colar silze : 2-6 6-10 10-14 14-18 18-22 22-26

Number of
Persons : 6 5 3 9 8 10

Ans : 0.85

15. Calculate quartile deviation and coefficient of quartile deviation from the following data :

Annual 1450, 1470, 1360, 1210, 960, 1490
Income : 1530, 1660, 1780, 1800, 2000.
(Rs.)

Ans : 13.3

16. Find out Inter-quartile range from the following data:

16, 25, 18, 13, 9, 28, 36, 43, 52

Ans : 39.52

17. Find out quartile deviatilon and its coefficient from the series given below and compare the consistency of the distribution.

X : 23, 45, 36, 61, 75, 54, 18, 29, 77

Y : 1 07, 243, 324, 134, 127, 116, 282, 300, 189

**Ans : 44% and 41%**

18. Compute coefficient of quartile deviation.

15.2, 15.3, 15.8, 16.1, 17.2, 16.5, 14.9,
17.6, 18.0, 17.9, 16.9,

Ans : 6.9

19. Calculate quartile deviation and coefficient of quartile deviation.

Weight (Kgs) : 35 37 39 41 43 45 49 51 53 55

Number of
Students : 12 11 16 25 32 54 43 46 30 21

Ans : 6.5

20. Following data relate to the yearly profits of 1200 firms.

Profits
(in 000's) : 10 15 20 25 30 35 40 45 50 55

Number of firms : 56 126 148 138 110 76 156 175 136 79

Find out the quartile deviation and coefficient of quartile deviation.

Ans : 38.4

21. Use Quartile Deviation and its coeffilcient to compare the following distributions.

Size : 15 16 17 18 19 20 21 22 23 24 25

Frequency : 21 33 23 14 18 16 28 10 16 11 9

Frequency : 14 36 26 37 18 24 27 13 12 7 4

Ans : 31.5

22. Compute quartile deviation and its coefficient.

Class : Less than 50, 50-75, 75-100, 100-125, Above 125

Frequency : 12 14 11 17 15

Ans : 31.15 and 0.34

23. Calculate coefficient of quartile deviation.

Mid Values : 15 20 25 30 35 40 45 50 55

Frequency : 36 47 54 25 19 8 49 64 69

Ans : 0.37

24. Find out quartile deviation and its coefficient.

Class : 2-10 11-19 20-28 29-37 38-46

Frequency : 18 23 9 4 29

Ans 0.55

## 17.13 RECOMMENDED BOOKS


1. Gupta, S.P. : "Statistical Methods", Sultan Chand & Company, New Delhi.
2. Gupta, B.N. : "Statistics", Sahitya Bhavan, Agra.
3. Gupta, S.C. : "Fundamentals of Statistics", Himalaya Pub. House, Bombay.
4. Simpson and Kafka : "Basic Statistics", Oxford and IBH Publishing Company, Calcutta.


## 17.14 GLOSSARY

1. Inter-quartile Range : It is the absolute difference between the upper quartile and the lower quartile.
2. Quartile Deviation : It is the average absolute difference between upper quartile and lower quartile of a distribution.
3. Co-efficient of Quartile Deviation : The relative measure of quartile deviation.
4. Co-efficient of Range : The relative measure of range.
5. Range : It is the absolute difference between the highest value and the lowest value of a distribution.

# UNIT -18: MEAN DEVIATION

**contents**



## 18.0 AIMS AND OBJECTIVES

This unit aims at explaining the meaning, mathematical calculation, merits and limitations of mean deviation.

On completion of this unit, you should be able to :

- explain the meaning of Mean Deviation
- compute mean deviation for the given exercises
- list the merits and limitations of mean deviation.



## 18.1 INTRODUCTION

In unit seventeen we discussed the range and quartile deviation for measuring the dispersion. The cheif limitation of these two methods is that, their calculation is not based on all the values. To overcome this problem, another method of dispersion which is based on all the values, is used. This is called mean deviation. In this unit, we shall see the theoretical and the practical aspects of mean deviation.

## 18.2 Mean Deviation - Meaning

Mean deviation, also termed as 'average deviation' is defined as the absolute average of the deviations taken from a measure of central tendency. In the words of Clark and Schkade, "Average deviation is the average amount of scatter of the items in a distribution from either mean or median ignoring the signs of the deviation". Though deviations can be taken from any measure of central tendency, arithmetic mean, median and mode are generally used. Mean is preferred because it is algebraically sound. Theoretically, median is considered appropriate as

## CONTENTS

# BUSINESS STATISTICS

## BLOCKS III TO V



Dr. B. R. AMBEDKAR OPEN UNIVERSITY
HYDERABAD
1992

**COURSE TEAM**

Prof. A. Shankaraiah (Editor)

Dr. V . Gangadhar

Dr. R. Sudarshan

Sri. N . Hanumantha Rao

Sri. V.V. Subrahmanya Sarma

**ASSOCIATE EDITORS**

Prof. V. Nagaraja Naidu

Sri. P. Krishna Rao

Art

Chandra

DR. B. R. AMBEDKAR OPEN UNIVERSITY

Hyderabad

Frist Published in 1984

Revised in 1990

Re-Print - 1992 - 93.



This text forms part of Dr. B. R. Ambedkar Open University Course. The complete syllabus for the course appears at the end of the text.

Further information about the Dr. B. R. Ambedkar Open University Courses may be obtained from the D (Academic), Dr. B. R. Ambedkar Open University, Somajiguda, Hyderabad - 500 482 (A.P.).

Printed at M/s. Sruthi Graphics (P) Ltd. 8-3-231/G/4, Sri Krishna Nagar, Hyderabad - 500 045. Phone

the sum of deviations is minimum when + or − signs are ignored. Mode is not very widely used because it is difficult to define the modal value in all distributions. In practice, however, arithmetic mean is the most commonly used average. This is the reason why the measure is called 'Mean Deviation'. For calculation of mean deviation, the alegebraic signs + or − are to be ignored, because, if signs are taken into consideration, the sum of deviations from mean will be zero, and the sum of deviations from median or mode will be nearly zero.

Mean deviation is obtained by dividing the sum of absolute deviations taken from the average by the total number of observations. Generally, the symbol modulus denotes that deviations taken by ignoring algebraic signs. The formula of mean deviation, when the deviations are measured from arithmetic mean is :

Where,

$$M.D_{\overline{X}} = \frac{\Sigma(/d_{\overline{X}}/)}{n}$$

$M.D._{\overline{X}}$ = Mean deivation from mean

$\Sigma(/d_{\overline{X}}/)$ = Sum of deviations measured from mean ignoring + or − signs

$n$ = Total number of observations.

When Mean deviation is calculated from median, the formula is :

$$M.D._{Med} = \frac{\Sigma(/d_{Med.}/)}{n}$$

where,

$M.D._{Med}$ = Mean deviation from Median

$\Sigma(/d_{Med.}/)$ = Sum of the deviations taken from the median ignoring + or − signs.

when Mean deviation is taken from Mode :

$$M.D._{Z} = \frac{\Sigma/d_{Z}/}{n}$$

$M.D._{Z}$ = Mean deivation from mode

$\Sigma(/d_{Z}/)$ = Sum of the deviations taken from mode ignoring + or − signs.

Mean deviation computed with the help of the above formula is an absolute measure. The relative measures of mean deviation is called coefficient of mean deviation. It is computed by dividing the mean deviation by the average used in measuring such deviations. The formula for calculating the coefficient of mean deviation, when Arithmetic mean is used, is :

Co-efficient of M.D.$_{\overline{X}} = \frac{M.D._{\overline{X}}}{\overline{X}}$

Where:

$M.D._{\overline{X}}$ = Mean deviation from mean

$\overline{X}$ = Arithmetic mean

When median is used, the formula is :

Coefficient of Mean deviation (Median) $= \frac{M.D._{Med.}}{Median}$

Where $M.D._{Med}$ = Mean deviation from median

When Mode is used, the formula is :

Coeffieient of M.D.$_Z = \frac{M.D._Z}{Z}$

Where $M.D._Z$ = Mean Deviation from mode

Z = Mode

## 18.3 CALCULATION OF MEAN DEVIATION

The procedure, involved in the calculation of mean deviation in individual, discrete and continuous series, both by direct method and short-cut method, is explained below :

**I) Individual series : a) Direct Method :**

i) Arrange the observations in ascending or descending order, if median is used.

ii) Calculate the average to be used to take deviations.

iii) Measure the deviations from the average selected ignoring the plus(+) or (−) minus and obtain the sum of the deviations. ($\Sigma$ /d/)

iv) Divide the sum of absolute deviations by the number of observations for obtaining mean deviation. Thus the formula is :

$$\text{M.D.} = \frac{\Sigma/d/}{N}$$

v) Divide the mean deviation by the average used to calculate coefficient of mean deviation. Thus :

coefficient of M.D. $= \frac{M.D.}{Average}$

**Illustration-I**

From the following data relating to hourly out-put of two workers, X and Y, compute Mean deviation and its coefficient. Use arithmetic mean.

X : 34 48 41 36 44 37 43 39 40 38

Y : 44 46 26 53 30 27 50 48 29 47

**Solution :**

**CALCULATION OF MEAN DEVIATION AND COEFFICIENT OF MEAN DEVIATION**

| Worker X | $/d_{\overline{X}}/$ | Worker Y | $/d_{\overline{X}}/$ |
|---|---|---|---|
| 34 | 6 | 44 | 4 |
| 48 | 8 | 46 | 6 |
| 41 | 1 | 26 | 14 |
| 36 | 4 | 53 | 13 |
| 44 | 4 | 30 | 10 |
| 37 | 3 | 27 | 13 |
| 43 | 3 | 50 | 10 |
| 39 | 1 | 48 | 8 |
| 40 | 0 | 29 | 11 |
| 38 | 2 | 47 | 7 |
| $\Sigma X = 400$ | $\Sigma/d_{\overline{X}}/ = 32$ | $\Sigma$ Y = 400 | $\Sigma/d_{\overline{X}}/ = 96$ |

**WORKER 'X'**

Arithmetic Mean $= \frac{\Sigma X}{N}$

Here $\Sigma X = 400$

N = 10

Substituting the values in the formula

$$\overline{X} = \frac{400}{10}$$

$$= 40$$

$$M.D._{\overline{X}} = \frac{\Sigma/d_{\overline{X}}/}{N}$$

Here $\Sigma/d_{\overline{X}}/ = 32$

N = 10

Substituting the values in the formula

$$M.D._{\overline{X}} = \frac{32}{10}$$

$$= 3.2$$

Coefficient of M.D. $= \frac{M.D._{\overline{X}}}{\overline{X}}$

Here $M.D._{\overline{X}} = 3.2$

$= 40$

Substituting the values in the formula,

Coefficient of M.D. $= \frac{3.2}{40}$

$= 0.08$

**WORKER 'Y' :**

Arithmetic Mean $= \frac{\Sigma Y}{N}$

Here $\Sigma Y = 400$

N = 10

Substituting the values in the formula,

$= \frac{400}{10}$

$= 40$

M.D. $= \frac{\Sigma / d_{\overline{X}} /}{N}$

Here $\Sigma / d_{\overline{X}} / = 96$

N = 10

Substituting the values in the formula,

$= \frac{96}{10}$

$= 9.6$

Coefficient of M.D. $= \frac{M.D._{\overline{X}}}{\overline{X}}$

Here $M.D._{\overline{X}} = 9.6$

$\overline{X} = 40$

Substituting the values in the formula,

Coefficient of M.D. $= \frac{9.6}{40}$

$= 0.24$

Since the coefficient of mean deviation of hourly output of worker 'X' is low, it is more consistent and less variable.

**Individual Series : (b) Short-cut-method :**

i) Arrange the observations in ascending or descending order.

ii) Calculate the average to be used.

iii) Find out the sum of the observations above the value of average ($\Sigma X_A$) and below the value of average ($\Sigma X_B$) in the distribution.

iv) Count the number of observations having their value above and below the average.

v) Apply the formula to obtain Mean deviation

$$M.D. = \frac{\Sigma X_A - \Sigma X_B - (n_A - n_B)Average}{N}$$

Where M.D. = Mean deviation

$\Sigma X_A$ = Sum of the observations above the value of average.

$\Sigma X_B$ = Sum of the observations below the value of average.

$n_A$ = Number of observations above the value of average.

$n_B$ = Number of observations below the value of average.

**Illustration II :** Calculate mean deviation and coefficient of mean deviation from mean and median from the following series.

X : 17 26 19 22 11 32 13 15 9 24 26

**Solution :**

**CALCULATION OF MEAN DEVIATION AND COEFFICIENT OF MEAN DEVIATION**

| Sl. No. | 'X' Series |
|---|---|
| 1. | 9 |
| 2. | 11 |
| 3. | 13 |
| 4. | 15 |
| 5. | 17 |
| 6. | 19 |
| 7. | 22 |
| 8. | 24 |
| 9. | 26 |
| 10. | 26 |
| 11. | 32 |
| | $\Sigma X = 214$ |

**Solution**

Calculation of Mean deviation and coefficient of M.D. from arithmetic mean

$$\overline{X} = \frac{\Sigma X}{N}$$

Here $\Sigma X = 214$

$N = 11$

Substituting the values in the formula,

$$\overline{X} = \frac{214}{11}$$

$$= 19.45$$

$$\text{M.D.}_{\overline{X}} = \frac{\Sigma X_A - \Sigma X_B - (n_A - n_B)\overline{X}}{N}$$

Here, $\Sigma X_A = 130$

$\Sigma X_B = 84$

$n_A = 5$

$n_B = 6$

$N = 11$

$\overline{X} = 19.45$

Substituting the values in the formula,

$$\text{M.D.}_{\overline{X}} = \frac{130 - 84 - (5-6)19.45}{11}$$

$$= \frac{46 - (-1)19.45}{11}$$

$$= \frac{65.45}{11}$$

$$= 5.95$$

Coefficient of M.D. $= \frac{M.D._{\overline{X}}}{\overline{X}}$

$M.D._{\overline{X}} = 5.95$

$\overline{X} = 19.45$

Substituting the values in the formula

Coefficient of M.D. $= \frac{5.95}{19.45}$

$= 0.306$

Since, the coefficient of mean deviation is low, the distribution is considered as consistent.

Calculation Mean deviation and coefficient of Mean Deviation from median :

Med. = Size of $\left(\frac{N+1}{2}\right)$th item N = 11

Med. = Size of $\left(\frac{11+1}{2}\right)$th item

= Size of 6th item

= Size of 6th item in the distribution is 19.

$$M.D._{Med.} = \frac{(\Sigma X_A - \Sigma X_B - n_A - n_B)Med}{N}$$

Here $\Sigma X_A = 149$

$\Sigma X_B = 65$

$n_A = 6$

$n_B = 5$

N = 11

Median = 19

Substituting the values in the formula,

$$M.D._{Med.} = \frac{149-65-(6-5)19}{11}$$

$$= \frac{84-(19)}{11}$$

$$= \frac{65}{11}$$

$$= 5.91$$

Coefficient of M.D. $= \frac{M.D._{Med.}}{Median}$

$M.D._{Med.} = 5.91$

Median = 19

Substituting the values in the formula,

Coefficient of M.D. $= \frac{5.91}{19}$

$= 0.311$

As the coefficient of mean deviation is low, the series is consistent.

**II) Discrete Series : (a) Direct method :**

i) Calculate the average from which mean deviation is to be computed.

ii) Find out the absolute deviations from the average by ignoring + or − signs.

iii) Multiply each individual deviation by the respective frequency and obtain their total.

iv) Obtain the total of frequencies.

v) For obtaining mean deviation, divide the sum of the products of frequencies and absolute deviations by the sum of frequencies. Thus the formula is :

$$\frac{\Sigma f|d|}{N}$$

Where, $\Sigma f|d|$ = Sum of the products of frequencies and absolute deviations

N = Sum of frequencies

vi) For calculating coefficient of mean deviation, divide the mean deviation by the average.

**Illustration III** : From the following distribution, calculate coefficient of mean deviation from mean.

| X : | 30 | 50 | 70 | 90 | 110 | 130 | 150 | 170 | 190 |
|---|---|---|---|---|---|---|---|---|---|
| f : | 6 | 9 | 11 | 14 | 20 | 15 | 10 | 8 | 7 |

**Solution** : To get coefficient of mean deviation, first mean deviation is to be calculated.

CALCULATION OF MEAN DEVIATION AND COEFFICIENT OF MEAN DEVIATION

| (X) | (f) | (fx) | $/d_{\overline{X}}/$ | $f/d_{\overline{X}}/$ |
|---|---|---|---|---|
| 30 | 6 | 180 | 80 | 480 |
| 50 | 9 | 450 | 60 | 540 |
| 70 | 11 | 770 | 40 | 440 |
| 90 | 14 | 1260 | 20 | 280 |
| 110 | 20 | 2200 | 0 | 0 |
| 130 | 15 | 1950 | 20 | 300 |
| 150 | 10 | 1500 | 40 | 400 |
| 170 | 8 | 1360 | 60 | 480 |
| 190 | 7 | 1330 | 80 | 560 |
| | $\Sigma f = 100$ | $\Sigma fX = 11000$ | | $\Sigma f/d_{\overline{X}}/ = 3480$ |

Mean $= \frac{\Sigma fX}{N}$

Here $\Sigma fx = 11{,}000$ and $N = 100$

substituting the values in the formula,

$\overline{X} = \frac{11{,}000}{100}$

$= 110$

$$M.D. = \frac{\Sigma f(/d_{\overline{X}}/)}{N}$$

Here $\Sigma f/d_{\overline{X}}/ = 3{,}480$

$N = 100$

Substituting the values in the formula

$= \frac{3{,}480}{100}$

$= 34.8$

Coefficient of M.D. $= \frac{M.D._{\overline{X}}}{\overline{X}}$

Here $M.D._{\overline{X}} = 34.8$

$\overline{X} = 110$

substituting the values in the formula,

coefficient of M.D. $= \frac{34.8}{110}$

$= 0.32$

As the coefficient of mean deviation is low, the distribution is more consistent and less variable.

**Illustration IV :** Calculate mean deviation and coefficient of mean deviation from the distribution of heights (use median).

| Heights (inches) | 60 | 61 | 62 | 63 | 64 | 65 | 66 | 67 |
|---|---|---|---|---|---|---|---|---|
| Number of persons | 15 | 17 | 14 | 13 | 10 | 16 | 12 | 19 |

**Solution :**

**CALCULATION OF MEAN DEVIATION AND COEFFICIENT OF MEAN DEVIATION**

| Height (X) | No. of persons (f) | cf | $/d_{Med}/$ | $f/d_{Med}/$ |
|---|---|---|---|---|
| 60 | 15 | 15 | 3 | 45 |
| 61 | 17 | 32 | 2 | 34 |
| 62 | 14 | 46 | 1 | 14 |
| 63 | 13 | 59 | 0 | 0 |
| 64 | 10 | 69 | 1 | 10 |
| 65 | 16 | 85 | 2 | 32 |
| 66 | 12 | 97 | 3 | 36 |
| 67 | 19 | 116 | 4 | 76 |
| | $\Sigma f = 116$ | | | $\Sigma f/d_{Med}/ = 347$ |

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

N = 116

Median size of $\left(\frac{116+1}{2}\right)$th item

= size of 58.5th item

= Size of 58.5th item is located in the cumulative frequency 59.

Therefore median is 63.

$$M.D._{Med.} = \frac{\Sigma f|d_{Med}|}{N}$$

Here, $\Sigma f|d_{Med.}| = 247$

$N = 116$

Subsitituing the values in the formula

$M.D._{Med} = \frac{247}{116}$

$= 2.12$

Coefficient of M.D. $= \frac{M.D._{Med}}{Median}$

Here, $M.D._{Med.} = 2.13$

Median $= 63$

substituting the values in the formula

coefficient of M.D. $= \frac{2.13}{63}$

$= 0.03$

As the coefficient of mean deviation is low, the distribution of heights has greater consistency.

**Discrete Series : (b) Short-cut Method :**

i) Calculate the average to be used.

ii) Find out the sum of frequencies.

iii) Multiply each value with its frequency and find out the sum of the products above the value of average $\Sigma f X_A$ and below the value of average $\Sigma f X_B$ respectively.

iv) Obtain the sum of frequencies which are above the value of the average and below the value of the average.

v) Apply the formula.

$$M.D. = \frac{\Sigma f X_A - \Sigma f X_B(\Sigma f_A - \Sigma f_B)Average}{N}$$

Where M.D. = Mean deviation

$\Sigma f X_A$ = Sum of the products of 'X' and its respective frequencies which are above the value of the average.

$\Sigma f X_B$ = Sum of the products of the 'X' and its respective frequencies which are below the value of the average.

$\Sigma f_A$ = Sum of the frequencies above the value of average.

$\Sigma f_B$ = Sum of the frequencies below the value of average.

N = Sum of the frequencies.

**Illustration V :** Calcuaire the mean deviation and coeffcient of mean deviation from the following:

| X : | 12 | 15 | 18 | 21 | 24 | 27 | 30 | 33 | 36 |
|---|---|---|---|---|---|---|---|---|---|
| f : | 8 | 12 | 14 | 26 | 32 | 36 | 18 | 9 | 2 |

**Solution:**

**CALCULATION OF MEAN DEVIATION AND COEFFICIENT OF MEAN DEVIATION**

| X | f | fX |
|---|---|---|
| 12 | 8 | 96 |
| 15 | 12 | 180 |
| 18 | 14 | 252 |
| 21 | 26 | 546 |
| 24 | 32 | 768 |
| 27 | 36 | 972 |
| 30 | 18 | 540 |
| 33 | 9 | 297 |
| 36 | 2 | 72 |
| | $\Sigma f = 157$ | $\Sigma fX = 3{,}732$ |

Mean $= \overline{X} = \frac{\Sigma fX}{N}$

Here, $\Sigma fx = 3{,}723$

N $= 157$

Substituting the values in the formula:

$$\overline{X} = \frac{3{,}723}{157}$$

$$= 23.71$$

M.D.$\overline{x} = \frac{\Sigma fX_A - \Sigma fX_B - (\Sigma f_A - \Sigma f_B)\overline{X}}{N}$

Here, $\Sigma fx_A = 2{,}649$

$\Sigma fx_B = 1{,}074$

$\Sigma f_A = 97$

$\Sigma f_B = 60$

N $= 157$

$\overline{X} = 23.71$

substituting the values in the formula

$$M.D._{\overline{X}} = \frac{2649-1074-(97-60)23.71}{157}$$

$$= \frac{1,575-(37)23.71}{157}$$

$$= \frac{1575-877.27}{157}$$

$$= 4.44$$

coefficient of M.D. $= \frac{M.D._{\overline{X}}}{\overline{X}}$

Here, $M.D._{\overline{X}} = 4.44$

$\overline{X} = 23.71$

substituting the values in the formula :

Coefficient of M.D. $= \frac{4.44}{23.71}$

$= 0.187$

As the coefficient of mean deviation is low, the distribution is highly consistent.

**III) continuous Series : (a) Direct Method :**

i) Obtain the mid values of each class (m)

ii) Calculate the average and take the absolute deviations of the mid values from the average.

iii) Multiply each individual absolute deviation with the respective frequency (f/d/) and get their total ($\Sigma$f/d/)

iv) to obtain mean deviation, divide the sum of the products of frequencies and absolute deviations by the sum of frequencies. Thus the formula is

$$\frac{\Sigma f/d/}{N}$$

Where $\Sigma f/d/$ = sum of the products of the frequencies and absolute deviations.

N = Sum of frequecies.

v) Coefficient of mean deviation can be obtained by dividing the mean deviation by the average used.

**Illustration VI** : Calculate the mean deviation and its coeffcient from the following data

| Marks : | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 | 70-80 |
|---|---|---|---|---|---|---|---|---|
| No.of students: | 2 | 3 | 4 | 2 | 4 | 6 | 8 | 1 |

**Solution :**

**CALCULATION OF MEAN DEVIATION AND COEFFICIENT OF MEAN DEVIATION**

| Class | MV | f | fX | $/d_{\overline{X}}/$ | $f/d_{\overline{X}}/$ |
|---|---|---|---|---|---|
| 0-10 | 5 | 2 | 10 | 40 | 80 |
| 10-20 | 15 | 3 | 45 | 30 | 90 |
| 20-30 | 25 | 4 | 100 | 20 | 80 |
| 30-40 | 35 | 2 | 70 | 10 | 20 |
| 40-50 | 45 | 4 | 180 | 0 | 0 |
| 50-60 | 55 | 6 | 330 | 10 | 60 |
| 60-70 | 65 | 8 | 520 | 20 | 160 |
| 70-80 | 75 | 1 | 75 | 30 | 30 |
| | | $\Sigma f = 30$ | $\Sigma fX = 1{,}330$ | | $\Sigma f/d_{\overline{X}}/ = 520$ |

Arithmetic mean $= \overline{X} = \frac{\Sigma fX}{N}$

Here $\Sigma fX = 1330$

$N = 30$

substituting the values in the formula

$\overline{X} = \frac{1330}{30}$

$= 44.33$

$$M.D._{\overline{X}} = \frac{\Sigma f/d_{\overline{X}}/}{N}$$

Substituting the values in the formula,

$= \frac{520}{30}$

$= 17.33$

coefficient of M.D. $= \frac{M.D._{\overline{X}}}{\overline{X}}$

Here $M.D._{\overline{X}} = 17.33$

$\overline{X} = 44.33$

Substituting thevalues in the formula

Coefficient of M.D. $= \frac{17.33}{44.33}$

$= 0.39$

The distribution is f airly consistent as the coefficient of mean deviation is low

**Illustration VII :** Calculate mean deviation and its coefficient from the data pertaining to intelligence rating of 200 school boys (use Median).

Intelligence rating : 2-6 6-10 10-14 14-18 18-22 22-26 26-30

No. of boys : 9 17 12 56 24 39 43

**Solution:**

Intelligence rating 'X', Number of boys 'f'

**CALCULATION OF MEAN DEVIATION AND COEFFICIENT OF MEAN DEVIATION.**

| X | MV | f | cf | $/d_{Med}/$ | $f/d_{Med}/$ |
|---|---|---|---|---|---|
| 2-6 | 4 | 9 | 9 | 15 | 135 |
| 6-10 | 8 | 17 | 26 | 11 | 187 |
| 10-14 | 12 | 12 | 38 | 7 | 84 |
| 14-18 | 16 | 56 | 94 | 3 | 168 |
| 18-22 | 20 | 24 | 118 | 1 | 24 |
| 22-26 | 24 | 39 | 157 | 5 | 195 |
| 26-30 | 28 | 43 | 200 | 9 | 387 |
| | | $\Sigma f = 200$ | | | $\Sigma f/d_{Med}/ = 1180$ |

Median = Size of $(\frac{N}{2})$th item

N = 200

Med. = Size of $(\frac{200}{2})$th item

= Size of 100th item

= Size of 100th item is located in the cumulative frequency 188. So Median class is 18-22.

The exact value of the median can be located by using the following formula.

$$Med. = L + \frac{\frac{N}{2} - cf}{f} \times i$$

Here L = 18

$(\frac{N}{2}) = 100$

cf = 94

f = 24

i = 4

Substituting the values in the formula,

$$\text{Med.} = 18 + \frac{100-94}{24} \times 4$$

$$= 18 + \frac{6 \times 4}{24}$$

$$= 18 + 1$$

$$= 19$$

$$M.D._{Med.} = \frac{\Sigma f / d_{Med.} /}{N}$$

Here, $\Sigma f / d_{Med.} / = 1180$

$N = 200$

Substituting the values in the formula

$$\text{M.D.}_{Med.} = \frac{1180}{200}$$

$$= 5.9$$

$$\text{Coefficient of M.D.} = \frac{M.D._{Med.}}{Median}$$

Here, $M.D._{Med.} = 5.9$

Median = 19

Substituting the value in the formula,

$$\text{Coefficient of M.D.} = \frac{5.9}{19}$$

$$= 0.311$$

since the value of coefficient of mean deviation of intelligence rating is low, the distribution is constistent.

**Continuous Series (b) Short-cut method.**

i) Find the mid values of the class.

ii) Obtain the sum of frequencies.

iii) Calculate the average to be used.

iv) Multiply each mid value with respective frequency to get (fx). Obtain the products above the value of the average and below the value of the average respectively.

v) Find out the sum of the frequencies above and below the value of the average respectively.

vi) Use the following formula to obtain mean deviation

$$\text{M. D.} = \frac{\Sigma m f_A - \Sigma m f_B - (\Sigma f_A - \Sigma f_B) Average}{N}$$

Where, $\Sigma m f_A$ = Sum of the products of the mid values and frequencies which are above the value of the average.

$\Sigma mf_B$ = Sum of the products of the mid values and the frequencies which are below the value of the average.

$\Sigma f_A$ = Sum of the frequencies below the value of the average.

$\Sigma f_B$ = Sum of the frequencies above the value of the average.

N = Sum of the frequencies.

**Illustration VIII :** Find out the mean deviation and its coefficient from the following data (use arithmetic average)

| Class : | 0-9 | 10-19 | 20-29 | 30-39 | 40-49 | 50-59 |
|---|---|---|---|---|---|---|
| Frequency: | 14 | 67 | 28 | 104 | 10 | 5 |

**Solution :**

CALCULATION OF MEAN DEVIATION AND COEFFICIENT OF MEAN DEVIATION

| Class | f | m | mf |
|---|---|---|---|
| 0-9 | 14 | 4.5 | 63.0 |
| 10-19 | 67 | 14.5 | 971.5 |
| 20-29 | 28 | 24.5 | 686.0 |
| 30-39 | 104 | 34.5 | 3588.0 |
| 40-49 | 10 | 44.5 | 445.0 |
| 50-59 | 5 | 54.5 | 272.0 |
| | $\Sigma f = 228$ | | $\Sigma mf = 6026.0$ |

Arithmetic average $\overline{X} = \frac{\Sigma mf}{N}$

Here, $\Sigma mf = 6026.0$

$N = 228$

Substituting the values in the formula,

$$\overline{X} = \frac{6026}{228}$$

$$= 26.43$$

$$M.D._{\overline{X}} = \frac{\Sigma mf_A - \Sigma mf_B - (\Sigma f_A - \Sigma f_B)\overline{X}}{N}$$

Here $\Sigma mf_A = 4305.5$

$\Sigma mf_B = 1720.5$

$\Sigma f_A = 119$

$\Sigma f_B = 109$

$N = 228$

$\overline{X} = 26.43$

Substituting the values in the formula,

$$M.D._{\overline{X}} = \frac{4305.5 - 1720.5 - (119-109)26.43}{228}$$

$$= \frac{2585 - (10)26.43}{228}$$

$$= \frac{2585 - 264.3}{228}$$

$$= \frac{2320.7}{228}$$

$$= 10.18$$

Coefficient of M.D. $= \frac{M.D._{\overline{X}}}{\overline{X}}$

Here, $M.D._{\overline{X}} = 10.18$

$\overline{X} = 26.43$

Substituing the values in the formula,

Coefficient of M.D. $= \frac{10.18}{26.43}$

$= 0.385$

As the coefficient of mean deviation is low, the distribution is more consistent and less variable.

**Check your progress-1**

calculate Mean Deviation from Median for the following data.

| Class : | 0-10 | 10-20 | 2-30 | 30-40 | 40-50 | 50-60 |
|---|---|---|---|---|---|---|
| Frequency: | 4 | 5 | 10 | 6 | 5 | 4 |



## 18.4. MERITS AND LIMITATIONS OF MEAN DEVIATION

**(A) Merits**

i) Mean deviation is simple to calculate and easy to understand. As 'Wessell, Willet and Simone' point out, "The outstanding advantage of the average deviation is its relative simplicity. Anyone familiar with the concept of average can readily appreciate the meaning of the average deviation. If a situation requires a measure of dispersion that will be presented to the general public or any group not.thoroughly grounded in statistics, the average deviation is very useful".

ii) Since the mean deviation takes into consideration the average of the differences most of the irregularities in the distribution are dispensed with.

iii) The computation of mean deviation is based on every item in the distribution. Hence, it is more representattive measure of dispersion than the distance measures like range, quartile deviation, etc.

iv) It is very useful in small samples with extreme values, as it is least affected by it.

**(B) Limitations**

i) The serious limitation of the mean deviation is that it ignores + or − signs while measuring the deviations. This prevents further algebraic use of the measure. It studies only the extent of variability but not the direction of variation.

ii) Mean deviation may not give accurate results with any of the three average, viz., mean, median, and mode. Though median is considered the appropriate average, it gives unsatisfactory results when the variations are high in the distribution. Mean is also not a correct measure because the sum of deviations ignoring + or − signs is higher. Mode is inappropriate because its exact value cannot always be found.

iii) Calculation of mean deviation is not simple. Particularly in case of short-cut method its calculation involves lengthy procedure.

iv) It is not considered a well defined measure as it alters with different averages.

## 18.5. SUMMING UP

Mean deviation is the average of the deviations measured from a measure of central tendency (mean, median or mode) by ignoring + or − signs. Though deviations can be obtained by applying any measure of central tendency, arithmetic mean is commonly used. Mean deviation is considered a better measure of dispersion since its computation is based on all the observations of the distribution. Though it is commonly used in various socio-economic surveys, it is not amenable for further algebraic treatment.



## 18.6 CHECK YOUR PROGRESS: MODEL ANSWERS

| Class | M.V | f | c.f | /d/ | f/d/ |
|---|---|---|---|---|---|
| 0-10 | 5 | 4 | 4 | 23 | 92 |
| 10-20 | 15 | 5 | 9 | 13 | 65 |
| 20-30 | 25 | 10 | 19 | 3 | 30 |
| 30-40 | 35 | 6 | 25 | 7 | 42 |
| 40-50 | 45 | 5 | 30 | 17 | 85 |
| 50-60 | 55 | 4 | 34 | 27 | 108 |
| | | | | | 422 |

$$Md = 20 + \frac{17-9}{10} \times 10$$

$$= 28$$

$$M.D. = \frac{422}{34} = 12.41$$

## 18.7 MODEL EXAMINATION QUESTIONS

**A. Short Qeustions**

1. What is mean deviation?
2. What is coefficient of mean deviation?
3. Why do you ignore algebraic signs to calculate mean deviation?
4. Why do you prefer arithmetic mean to calculate mean deviation?
5. Explain the procedure for calculating mean deviation in discrete series?
6. How do you calculate mean deviation by short cut method in continuous series?

**B. Essay Questions**

7. Explain the merits and limitations of mean deviation.

**EXERCISES**

8. Find out mean deviation and its coeffcient.

   Size : 43; 45; 63; 41; 73; 34; 93; 38; 83

   (Ans : 18.7 and 0.33)

9. Calculate coefficient of mean deviation for income and expenditure of a family.

   Income (Rs.) : 410, 320, 500, 350, 205, 475, 525, 270, 360 475

   Expenditure: 350, 270, 430, 275, 195, 405, 490, 260, 310, 425

   (Ans : 0.21)

10. Compute mean deviation and coefficient of mean deviation from the annual output of a factory (Use Median).

    Annual output (in '000 tonnes) : 22, 38, 40, 51, 24, 37, 49, 61, 74, 77, 80

    (Ans : 18.4 and 0.49)

11. Calculate mean deviation and its coefficient.

    | Sl.No. : | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
    |---|---|---|---|---|---|---|---|---|---|---|
    | 'X' : | 29, | 21, | 28, | 22, | 29, | 25, | 27, | 24, | 20, | 30 |

    (Ans : 2.57 and 0.46)

12. Compute Mean deviation of marks obtained in Accountancy by B.Com. students. Calculate coefficient of mean deviation and interpret the results.

| | | | | | | |
|---|---|---|---|---|---|---|
| Marks : | 25 | 30 | 35 | 40 | 45 | 50 |
| No.of students: | 8 | 16 | 24 | 20 | 18 | 4 |

(Ans : 5.2 and 0.14)

13. Calculate coefficient of mean deviation and compare the efficiency of two transport lorries (Use Median)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Number of liters of fuel : | 50 | 75 | 100 | 125 | 150 | 175 | 200 |
| Number of kilometers: | | | | | | | |
| Lorry - A: | 300 | 450 | 600 | 625 | 1050 | 1313 | 1600 |
| Lorry - B: | 250 | 600 | 700 | 625 | 600 | 1400 | 1200 |

(Ans :0.25 and 0.27)

14. Compute coefficient of mean deviation from the following data.

| | | | | | | |
|---|---|---|---|---|---|---|
| X : | 10 | 16 | 22 | 28 | 34 | 40 |
| f : | 15 | 28 | 35 | 17 | 24 | 21 |

(Ans : 0.33)

15. Calculate mean deviation and its coefficient from the details given below.

| | |
|---|---|
| Sum of the products absolute deviation and frequencies | = 2410 |
| Sum of frequecies | = 50 |
| Arithmetic mean used to measure the absolute deviation | = 65 |

(Ans : 48.2 and 0.75)

16. Calculate mean deviation and coefficient of mean deviation of sales in a cloth shop.

| | | | | | | |
|---|---|---|---|---|---|---|
| X : | 25 | 30 | 35 | 40 | 45 | 50 |
| f : | 7 | 16 | 23 | 35 | 52 | 12 |

(Ans : 5.24 and 0.13)

17. Find out coefficient of mean deviation from the following data. (Use Median).

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Number of persons per family : | 2 | 5 | 8 | 11 | 14 | 17 | 20 |
| Number of families : | 5 | 11 | 4 | 5 | 3 | 2 | 1 |

(Ans : 0.79)

18. Calcualte mean deviation by short-cut method.

| X : | 4.2 | 4.3 | 4.7 | 4.8 | 5.0 | 5.1 | 5.8 | 7.5 |
|---|---|---|---|---|---|---|---|---|
| f : | 86 | 52 | 49 | 43 | 35 | 31 | 30 | 11 |

(Ans : 1.2)

19. Calculate mean deviation (Use Median).

| Class : | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 |
|---|---|---|---|---|---|---|---|
| Frequency: | 4 | 8 | 12 | 16 | 20 | 24 | 28 |

(Ans : 16.11)

20. Compute mean deviation from the following data.

| Age : | 0-6 | 6-12 | 12-18 | 18-24 | 24-30 | 30-36 | 36-42 |
|---|---|---|---|---|---|---|---|
| Number of persons : | 15 | 22 | 26 | 18 | 30 | 46 | 17 |

(Ans : 17.7)

21. Calculate mean deviation. Use short-cut method.

| Height (cms) : | 150-155 | 155-160 | 160-165 | 165-170 | 170-175 | 175-180 |
|---|---|---|---|---|---|---|
| Number of persons : | 30 | 47 | 69 | 32 | 18 | 4 |

(Ans : 5.46)

22. Compute mean deviation and its coefficient from the following data.

| Class : | 0-3 | 4-7 | 8-11 | 12-15 | 16-19 | 20-23 | 24-27 | 28-31 |
|---|---|---|---|---|---|---|---|---|
| Size : | 9 | 14 | 11 | 8 | 10 | 16 | 18 | 5 |

(Ans : 8.7 and 0.83)

23. Calculate mean deviation and coefficient of mean deviation.

| Midvalues : | 12 | 22 | 32 | 42 | 52 | 62 | 72 |
|---|---|---|---|---|---|---|---|
| Frequency : | 14 | 43 | 84 | 33 | 61 | 57 | 49 |

(Ans : 14.4 and 0.19)

24. Calcualte mean deviation and its coeeficient from the following data.

| Marks (less than): | 10 | 20 | 30 | 40 | 50 | 60 | 70 |
|---|---|---|---|---|---|---|---|
| Number of : | 5 | 12 | 17 | 25 | 38 | 41 | 50 |

(Ans : 89.1 and 0.59)

## 18.8 RECOMMENDED BOOKS


1. Gupta, S.P : "Statistical Methods", Sultan Chand & Company, New Delhi.
2. Gupta, B.N. : "Statististics", Sahitya Bhavan, Agra.
3. Gupta, S.C : "Fundamentals of statistics", Himalaya Pub. House, bombay.
4. Simpson and Kafka : "Basic statistics', Oxford and I.B.H. Publishing company, Calcutta.


## 18.9 GLOSSARY

1. Co-efficient of Mean deviation : It is the relative measure of mean deviation.
2. Mean Deviation : It is obtained by dividing the sum of deviations, taken from an average, by the total number of observations.

## UNIT - 19 STANDARD DEVIATION AND LORENZ CURVE

**Contents**





## 19.0 AIMS AND OBJECTIVES

The aim of this unit is to explain standard deviation, its calculation and other measures related to it, such as the coefficient of variation and variance and to deal with the graphic method of studying dispersion.

On completion of this unit, you should be able to :

- explain the terms standard deviation, coefficient of variation, variance and combined standard deviation.
- calculate standard deviation, coeffiecient of variation, variance and combined deviation.
- differentiate between mean deviation and standard deviation
- list the merits and limitations of standard deviation
- establish the relationship among various measures of dispersion
- explain dispersion through lerenz curve.

## 19.1 INTRODUCTION

The cheif defect of mean deviation is that it ignores algebraic signs of deviations. To remove the drawback of all the methods discussed so far, standard deviation is used. Like mean deviation, the calculation of standard deviation is also based on all the observations. It is a popularly used method. Let us discuss its related measures and computational aspects.

Further, we are also introducing lorenz curve which is a method of studying dispersion through graph.

## 19.2. STANDARD DEVIATION-MEANING

To overcome the limitation of ignoring the algebraic signs in mean deviation, the concept of standard deviation was developoed by "Karl Pearson" in 1893. Standard deviation may be defined as the square root of the mean of the squared deviations measured from arithmetic average. It is also called the "Root-mean square deviation". Symbolically it is expressed by the small Greek letter $\sigma$ (sigma). For calculating the standard deviation the deviations are always taken from the mean, because, the sum of squared deviations are minimum when measured from the arithmetic average. High value of standard deviation denotes greater variation and less uniformity. Whereas the lower value indicates greater representativeness of the averages.

While the standard deviation is an absolute measure, the relative measure is known as coefficient of standard deviation.

## 19.3 DIFFERENCE BETWEEN MEAN DEVIATION AND STANDARD DEVIATION

Though theoretically both the mean deviation and standard deviation are computed by taking all the observations in the distribution, the following differences can be noted between them.

| | Mean deviation | | Standard deviation |
|---|---|---|---|
| (i) | Deviations are to be taken by ignoring + or − signs. | (i) | Deviations are taken by taking + or − signs into account. |
| (ii) | It can be computed from any of the three averages namely mean, median and mode. | (ii) | It should be always computed from the arithmetic mean only. |
| (iii) | The deviations need not be squared. | (iii) | The deviations are to be squared. |

## 19.4 CALCULATIONS OF STANDARD DEVIATION

Methods of calculating the standard deviation in individual, discrete and continuous series are explained below.

**I (a) Individual series: When the Deviations are measured from actual arithmetic mean.**

i) Calculate the arithmetic mean of the given distribution.

ii) Take the deviations from the actual mean $(X - \overline{X})$

iii) Square the individual deviations and find out their total.

iv) Divide the squared deviations by the number of items and find out the square root to obtain the S.D. Thus the formula:

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}}$$

Where,

$\sigma$ = standard deviation

$\sigma x^2$ = sum of the squared deviations of "X" values measure from the actual arithmetic mean.

N= Number of observations

v) For obtaining coefficient of standard deviation divide the standard deviation by arithmetic mean. Thus coefficient of standard deviation

$$= \frac{\sigma}{\overline{X}}$$

**Illustration :** I. Find the standard deviation from the following

(X) 22, 17, 15, 23, 25, 24, 16, 18, 26, 14

**Solution:**

**COMPUTATION OF STANDARD DEVIATION AND COEFFICIENT OF STANDARD DEVIATION**

| 'X' | $x$ $(X-\overline{X})$ | $x^2$ |
|---|---|---|
| 22 | 2 | 4 |
| 17 | −3 | 9 |
| 15 | −5 | 25 |
| 23 | 3 | 9 |
| 25 | 5 | 25 |
| 24 | 4 | 16 |
| 16 | −4 | 16 |
| 18 | −2 | 4 |
| 26 | 6 | 36 |
| 14 | −6 | 36 |
| $\Sigma X = 200$ | | $\Sigma x^2 = 180$ |

Arithmetic mean $= \overline{X} = \frac{\Sigma X}{N}$

Here, $\Sigma X = 200$, N = 10

Substituting the values in the formula,

$$\overline{X} = \frac{200}{10}$$

$$= 20$$

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}}$$

Here,

$$\Sigma x^2 = 180$$

$$N = 10$$

Substituting the values in the formula

$$= \sqrt{\frac{180}{10}}$$

$$= \sqrt{18}$$

$$= 4.24$$

Coefficient of standard deviation $= \frac{\sigma}{\bar{X}}$

Coefficient of $\sigma = \frac{4.24}{20}$

$= 0.212$

As the coefficient of standard deviation is low, the distribution is less variable and more homogeneous.

**Individual series: When the deviations are taken from Assumed mean**

i) Take any value as the assumed mean and measure the deviations of 'X' from the assumed mean and find its total.

ii) Square these deviations and find out their total

iii) Apply the formula

$$\sigma = \sqrt{\frac{\Sigma dx^2}{N} - \left(\frac{\Sigma dx}{N}\right)^2}$$

Where,

$\Sigma dx$ = sum of the deviations measured from assumed mean

$\Sigma dx^2$ = Sum of the squared deviations measured from the assumed mean

N = total number of observations

**Illustration : II.** Find the standard deviation and coefficient of standard deviation from the following data.

(X): 458 459 450 448 457 462 439 454 451

**Solution :** Calculation of standard deviation and coefficient of standard deviation.

| X | $dx$ A = 457 | $dx^2$ |
|---|---|---|
| 458 | +1 | 1 |
| 459 | +2 | 4 |
| 450 | −7 | 49 |
| 448 | −9 | 81 |
| 457 | 0 | 0 |
| 462 | +5 | 25 |
| 439 | −18 | 324 |
| 454 | −3 | 9 |
| 451 | −6 | 36 |
| $\Sigma X = 4078$ | $\Sigma dx = -35$ | $\Sigma dx^2 = 529$ |

Arithmetic mean $\overline{X} = \frac{\Sigma X}{N}$

Here $\Sigma X = 4078$ and N = 9

Substituting the values in the formula,

$$= \frac{4078}{9}$$
$$= 453.11$$

Standard deviation $(\sigma) = \sqrt{\frac{\Sigma dx^2}{N} - (\frac{\Sigma dx}{N})^2}$

Here,

$$\Sigma dx^2 = 529$$
$$\Sigma dx = -35$$
$$N = 9$$

Substituting the values in the formula

$$= \sqrt{\frac{529}{9} - \left(\frac{-35}{9}\right)^2}$$
$$= \sqrt{58.78 - 15.12}$$
$$= \sqrt{43.66}$$
$$= 6.61$$

Coefficient of standard deviation $= \frac{\sigma}{\overline{X}}$

Coefficient of $\sigma = \frac{6.61}{453.11} = 0.015$

Since the coeffcient of standard deviation is low, the distribution is highly consistent.

**Illustration: III.** Compare the variability of the following two distributions with the help of standard deviation and its coefficient.

Distributions

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| A | 12 | 11 | 19 | 17 | 18 | 20 | 23 | 26 |
| B | 5 | 8 | 10 | 13 | 16 | 19 | 25 | 30 |

**Solution:** To compare the variability of the distribution coefficient of standard deviation is calculated.

| Distribution 'A' | | | Distribution 'B' | | |
|---|---|---|---|---|---|
| | (X−17) | | | (X−13) | |
| X | $dx$ | $dx^2$ | X | $dx$ | $dx^2$ |
| 12 | −5 | 25 | 5 | −8 | 64 |
| 11 | −6 | 36 | 8 | −5 | 25 |
| 19 | 2 | 4 | 10 | −3 | 9 |
| 17 | 0 | 0 | 13 | 0 | 0 |
| 18 | 1 | 1 | 16 | 3 | 9 |
| 20 | 3 | 9 | 19 | 6 | 36 |
| 23 | 6 | 36 | 25 | 12 | 144 |
| 26 | 9 | 81 | 30 | 17 | 289 |
| $\Sigma X$ =146 | $\Sigma dx = 10$ | $\Sigma dx^2$ =192 | $\Sigma X$ =126 | $\Sigma dx = 22$ | $\Sigma dx^2 = 576$ |

**Distribution'A'**

Arithmetic mean= $\overline{X} = \frac{\Sigma X}{N}$

$\Sigma X = 146$ and N=8

$\overline{X} = \frac{146}{8} = 18.25$

standard deviation $= \sqrt{\frac{\Sigma dx^2}{N} - \left(\frac{\Sigma dx}{N}\right)^2}$

$\Sigma dx^2 = 192$

$\Sigma dx = 10$

N = 8

Substituting the values in the formula

$$\sigma = \sqrt{\frac{192}{8} - \left(\frac{10}{8}\right)^2}$$
$$= \sqrt{24 - 1.56}$$
$$= \sqrt{22.44}$$
$$\sigma = 4.74$$

Coefficient of standard deviation $= \frac{\sigma}{\overline{X}}$

$= \frac{4.74}{18.25}$

$= 0.259$

**Distribution : 'B'**

Arithmetic mean $\overline{X} = \frac{\Sigma X}{N}$

$\Sigma X = 126$ and $N = 8$

$\overline{X} = \frac{126}{8}$

$= 15.75$

Standard deviation $= \sqrt{\frac{\Sigma dx^2}{N} - \left(\frac{\Sigma dx}{N}\right)^2}$

$\Sigma dx^2 = 576$

$\Sigma dx = 22$

$N = 8$

substituting the values in the formula

$= \sqrt{\frac{576}{8} - \left(\frac{22}{8}\right)^2}$

$= \sqrt{72 - 7.56}$

$= \sqrt{64.44}$

$= 8.03$

Coefficient of standard deviation $= \frac{\sigma}{\overline{X}}$

$= \frac{8.03}{15.75}$

$= 0.509$

Distribution 'A' is more consistent and less variable as its coefficient of standard deviation is low.

**II.(b) Discrete series:** ***When deviations are measured from Actual arithmetic mean :***

i) Calculate the arithmetic mean

ii) Obtain the deviations from the mean and square each individual deviation.

iii) Multiply the squared deviations with the respective frequencies and obtain their total.

iv) Divide the sum of the product of squared deviations and frequencies by the sum of frequencies and find the square root.

Thus the formula is,

$$\sigma = \sqrt{\frac{\Sigma f d^2}{N}}$$

Where, $\Sigma fd^2$ = sum of the products of the squared deviations and the frequencies

N = Sum of the frequencies

v) For obtaining coefficient of standard deviation divide the standard deviation by arithmetic mean. Thus, coefficient of standard deviations :

$$= \frac{\sigma}{\overline{X}}$$

**Illustration :IV** . Find the standard deviation and its coefficient from the data relating to the bursting pressure of packing paper.

| bursting pressure | 10 | 12 | 14 | 16 | 18 | 20 | 22 | 24 | 26 |
|---|---|---|---|---|---|---|---|---|---|
| Number of samples: | | | | | | | | | |
| Brand'A' | 2 | 3 | 2 | 4 | 1 | 8 | 3 | 1 | 2 |
| Brand'B' | 4 | 10 | 8 | 6 | 2 | 0 | 4 | 2 | 0 |

**Solution :** CALCULATION OF STANDARD DEVIATION AND COEFFICIENT OF STANDARD DEVIATION.

| BRAND 'A' | | | | | | BRAND 'B' | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Bursting pressure | Number of samples | fx | $(X - \overline{X})$ d | $d^2$ | $fd^2$ | Number of samples | fx | d | $d^2$ | $fd^2$ |
| X | f | | | | | f | | | | |
| 10 | 2 | 20 | −8 | 64 | 128 | 4 | 40 | −5 | 25 | 100 |
| 12 | 3 | 36 | −6 | 36 | 108 | 10 | 120 | −3 | 9 | 900 |
| 14 | 2 | 28 | −4 | 16 | 32 | 8 | 112 | −1 | 1 | 8 |
| 16 | 4 | 64 | −2 | 4 | 16 | 6 | 96 | 1 | 1 | 6 |
| 18 | 1 | 18 | 0 | 0 | 0 | 2 | 36 | 3 | 9 | 18 |
| 20 | 8 | 160 | 2 | 4 | 32 | 0 | 0 | 5 | 25 | 0 |
| 22 | 3 | 66 | 4 | 16 | 48 | 4 | 88 | 7 | 49 | 196 |
| 24 | 1 | 24 | 6 | 36 | 36 | 2 | 48 | 9 | 81 | 162 |
| 26 | 2 | 52 | 8 | 64 | 128 | 0 | 0 | 11 | 121 | 0 |
| N=26 | $\Sigma fX = 468$ | | | | $\Sigma fd^2 = 528$ | N = 36 | $\Sigma fX = 540$ | | | $\Sigma fd^2 = 1390$ |

Arithmetic mean $= \overline{X} = \frac{\Sigma fX}{N}$

$\Sigma fX = 468$ and $N = 26$

$\overline{X} = \frac{468}{26} = 18$

Standard deviation $= \sqrt{\frac{\Sigma fd^2}{N}}$

Here,

$$\Sigma fd^2 = 528$$

$$N = 26$$

Substituting the values in the formula

$$\sigma = \sqrt{\frac{528}{26}}$$

$$= \sqrt{20.31}$$

$$= 4.51$$

Coefficient of standard deviation $= \frac{\sigma}{\overline{X}}$

$$= \frac{4.51}{18}$$

$$= 0.251$$

**Brand 'B'**

Arithmetic mean $= \overline{X} = \frac{\Sigma fX}{N}$

$\Sigma fx = 540$ and $N = 36$

Substituting the values in the formula

$$\overline{X} = \frac{540}{36}$$

$$= 15$$

Standard deviation $= \sqrt{\frac{\Sigma fd^2}{N}}$

$$\Sigma fd^2 = 1390$$

$$N = 36$$

substituting the values in the formula,

$$= \sqrt{\frac{1390}{36}}$$

$$= \sqrt{38.61}$$

$$= 6.21$$

Coefficient of standard deviation $= \frac{\sigma}{\overline{X}}$

$$= \frac{6.21}{15}$$

$$= 0.414$$

As the coefficient of standard deviation of bursting pressure of Brand 'A' paper samples is low, Brand 'A' is considered more consistent.

**Discrete Series :** ***When deviations are taken from assumed mean***

1) Take any value in the distribution as the assumed mean and measure the deviatioins of 'X' from the assmued mean.

ii) Multiply these deviations with their respective frequencies and add them

iii) Square the deviations and multiply them with the respective frequencies, and add them.

iv) Apply the following formula.

$$= \sqrt{\frac{\Sigma f dx^2}{N} - \left(\frac{\Sigma f dx}{N}\right)^2}$$

Where

$\Sigma f dx^2$ = Sum of the products of squared deviations and frequencies.

$\Sigma f dx$ = Sum of the products of deviations and their respective frequencies.

N = Sum of frequencies

**Illustration :** V. The following data related to the weights of a group of boys. Compute standard deviation and its coefficient.

| Weights (in Kgs) : | 40 | 42 | 44 | 46 | 48 | 50 | 52 | 54 | 56 | 58 | 60 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| No. of boys: | 12 | 8 | 4 | 10 | 5 | 9 | 5 | 3 | 2 | 1 | 1 |

**Solution:** CALCULATION OF STANDARD DEVIATION AND COEFFICIENT OF STANDARD DEVIATION.

| Weights (X) | Number of boys (f) | $dx$ | $dx^2$ | $fdx$ | $fdx^2$ |
|---|---|---|---|---|---|
| 40 | 12 | −10 | 100 | −120 | 1200 |
| 42 | 8 | − 8 | 64 | − 64 | 512 |
| 44 | 4 | − 6 | 36 | − 24 | 144 |
| 46 | 10 | − 4 | 16 | − 40 | 160 |
| 48 | 5 | − 2 | 4 | − 10 | 20 |
| 50 | 9 | 0 | 0 | 0 | 0 |
| 52 | 5 | 2 | 4 | 10 | 20 |
| 54 | 3 | 4 | 16 | 12 | 48 |
| 56 | 2 | 6 | 36 | 12 | 72 |
| 58 | 1 | 8 | 64 | 8 | 64 |
| 60 | 1 | 10 | 100 | 10 | 100 |
| | $\Sigma f$ =60 | | | $\Sigma fdx$ = + 52 = − 206 − 258 | $\Sigma fdx^2$ =2340 |

Arithmetic mean $= \overline{X} = A + \frac{\Sigma fdx}{N}$

Here,

$$A = 50$$

$$\Sigma fdx = -206$$

$$N = 60$$

substituting the values in the formula

$$= 50 + \frac{-206}{60}$$

$$= 50 - 3.43$$

$$= 46.57$$

Standard deviation $= \sqrt{\frac{\Sigma fdx^2}{N} - \left(\frac{\Sigma fdx}{N}\right)^2}$

Here, $\Sigma fdx^2 = 2340$

$$\Sigma fdx = -206$$

$$N = 60$$

substituting the values in the formula:

$$= \sqrt{\frac{2340}{60} - \left(\frac{-206}{60}\right)^2}$$

$$= \sqrt{39 - 11.76}$$

$$= \sqrt{27.24}$$

$$= 5.22$$

Coefficient of standard deviation

$$= \frac{\sigma}{\overline{X}}$$

$$= \frac{5.22}{46.57} = 0.112$$

Since the coefficient of standard deviation of the weights is low, the distribution is consistent.

**III. a) Continuous series:** ***When deviations are measured from actual arithmetic mean.***

i) Find the mid value of the classes of the distribution.

ii) Calculate the arithmetic mean and take the deviation of mid values from the actual mean.

iii) Square the deviations and multiply them with their respective frequencies, and obtain their totals.

iv) Divide the sum of the products of frequencies and squared deviations by the sum of the frequencies, and obtain the standard deviation.

Symbolically,

$$\sigma = \sqrt{\frac{\Sigma f d^2}{N}}$$

Where, $\Sigma f d^2$ = Sum of the products of frequencies and squared deviations taken from the mean.

N = Sum of the frequencies

**Illustration :** VI calculate standard deviation and its coefficient from the following data.

| Class | : | 1-5 | 5-9 | 9-13 | 13-17 | 17-21 | 21-25 |
|---|---|---|---|---|---|---|---|
| Frequency: | | 15 | 12 | 24 | 16 | 10 | 20 |

**Solution :** CALCULATION OF STANDARD DEVIATION AND COEFFICIENT OF STANDARD DEVIATION

| Class | MV X | f | fX | d | $d^2$ | $fd^2$ |
|---|---|---|---|---|---|---|
| 1-5 | 3 | 15 | 45 | −8 | 64 | 960 |
| 5-9 | 7 | 12 | 84 | −4 | 16 | 192 |
| 9-13 | 11 | 20 | 220 | 0 | 0 | 0 |
| 13-17 | 15 | 16 | 240 | 4 | 16 | 256 |
| 17-21 | 19 | 10 | 190 | 8 | 64 | 640 |
| 21-25 | 23 | 2 | 46 | 12 | 144 | 288 |
| | | $\Sigma f$ =75 | $\Sigma fX$ =825 | | | $\Sigma fd^2 = 2336$ |

Arithmetic mean $\overline{X} = \frac{\Sigma fX}{N}$

Here,

$$\Sigma fX = 825$$

$$N = 75$$

substituting the values in the formula

$$\overline{X} = \frac{825}{75} = 11$$

Standard deviation $= \sqrt{\frac{\Sigma f d^2}{N}}$

Here,

$$\Sigma f d^2 = 2336$$

$$N = 75$$

Substituting the values in the formula

$$= \sqrt{\frac{2336}{75}}$$

$$= \sqrt{31.15}$$

$$= 5.58$$

Coefficient of Standard deviation $= \frac{\sigma}{\bar{X}}$

$$= \frac{5.58}{11} = 0.507$$

As the coefficient of standard deviation is high, the distribution is less uniform or more variable.

**b) Continuous Series : When deviations are measured from assumed mean**

i) Find the midvalues of classes of the distribution.

ii) Take any value from the midvalues as the assumed mean and measure the deviations of mid values from the assumed mean.

iii) Multiply these deviations with the frequencies and obtain the total.

iv) Square the deviations and multiply them with their respective frequencies and find out their total.

v) Use the following formula for calculating standard deviation.

$$\sigma = \sqrt{\frac{\Sigma f dx^2}{N} - \left(\frac{\Sigma f dx}{N}\right)^2}$$

Where, $\Sigma f dx$ = Sum of the products of the deviations measured from assumed mean and their respective frequencies

$\Sigma f dx^2$ = Sum of the products of squared devations measured from assumed mean and their respective frequencies.

N = Sum of frequencies.

vi) Coefficient of standard deviation is obtained by dividing the standard deviation by the arithmetic mean.

Symbolically,

Coefficient of $\sigma = \frac{\sigma}{\bar{X}}$

**Illustration : VII** Calculate standard deviation and its coefficient from the following data.

| Class | 0 - 6 | 6 - 12 | 12 - 18 | 18 - 24 | 24 - 30 | 30 - 36 | 36 - 40 |
|---|---|---|---|---|---|---|---|
| frequency | 5 | 8 | 4 | 6 | 10 | 9 | 3 |

**Solution :**

**CALCULATION OF STANDARD DEVIATION**

| X | MV | f | (X−A)dx (A = 21) | $dx^2$ | $fdx$ | $fdx^2$ |
|---|---|---|---|---|---|---|
| 0-6 | 3 | 5 | −18 | 324 | −90 | 1620 |
| 6-12 | 9 | 8 | −12 | 144 | −96 | 1152 |
| 12-18 | 15 | 4 | − 6 | 36 | −24 | 144 |
| 18-24 | 21 | 6 | 0 | 0 | 0 | 0 |
| 24-30 | 27 | 10 | 6 | 36 | 60 | 360 |
| 30-36 | 33 | 9 | 12 | 144 | 108 | 1296 |
| 36-40 | 39 | 3 | 18 | 324 | 54 | 972 |
| | $\Sigma f = 45$ | | | | $\Sigma fdx = 12$ | $\Sigma fdx^2 = 5544$ |

Arithmetic mean : $\overline{X} = A + \frac{\Sigma fdx}{N}$

$A = 21$

Here, $\Sigma fdx = 12$

$N = 45$

Substituting the values in the formula

$\overline{X} = 21 + \frac{12}{45}$

$\overline{X} = 21 + 0.2$

$\overline{X} = 21.27$

Standard deviation $= \sqrt{\frac{\Sigma fdx^2}{N} - \left(\frac{\Sigma fdx}{N}\right)^2}$

Here,

$\Sigma fdx^2 = 5544$

$\Sigma fdx = 12$

$N = 45$

Substituting the values in the formula

$$\sigma = \sqrt{\frac{5544}{45} - \left(\frac{12}{45}\right)^2}$$

$$= \sqrt{123.2 - 0.07}$$

$$= \sqrt{123.13}$$

$$\sigma = 11.10$$

Coefficient of stqandard deviation $= \frac{\sigma}{\bar{X}}$

$= \frac{11.10}{21.27}$

$= 0.522$

As the coeffficient of standard deviation is 0.522 the distribution is considered less consistent.

**c) Step deviation method :** In this method the deviations are divided by a common factor to make the calculations easier. Calculation of standard deviation, under this method,involves the following correction.

$$d^1 = \frac{(m - A)}{C}$$

Here,

The deviations taken from the midvalues are divided by the common factor.

Normally the class interval is taken as the common factor.

**Illustration: VIII.** Calculate standard deviation and its coefficient from the following data.

| Class | : | 25-35 | 35-45 | 45-55 | 55-65 | 65-75 | 75-85. |
|---|---|---|---|---|---|---|---|
| Frequency : | | 10 | 15 | 20 | 21 | 18 | 16 |

**Solution :** CALCULATION OF STANDARD DEVIATION AND COEFFICIENT OF STANDARD DEVIATION.

| Class | Mv | f | $\frac{(m-A)}{C}$ A = 50 $d^1$ | $d^{1^2}$ | $fd^1$ | $fd^{1^2}$ |
|---|---|---|---|---|---|---|
| 25-35 | 30 | 10 | −2 | 4 | −20 | 40 |
| 35-45 | 40 | 15 | −1 | 1 | −15 | 15 |
| 45-55 | 50 | 20 | 0 | 0 | 0 | 0 |
| 55-65 | 60 | 21 | 1 | 1 | 21 | 21 |
| 65-75 | 70 | 18 | 2 | 4 | 36 | 72 |
| 75-85 | 80 | 16 | 3 | 9 | 48 | 144 |
| | | N = 100 | | | $\Sigma fd^1 = 70$ | $\Sigma fd^{1^2} = 291$ |

Arithmetic mean $= \bar{X} = A + \frac{\Sigma fd^1}{N} \times C$

Here,

$A = 50$

$\Sigma fd^1 = 70$

$$N = 100$$

$$C = 10$$

Substituting the values in the formula

$$\overline{X} = 50 + \frac{70}{100} \times 10$$

$$= 50 + 7$$

$$= 57$$

Standard deviation $= \sqrt{\frac{\Sigma f d^{1^2}}{N} - \left(\frac{\Sigma f d^1}{N}\right)^2} \times C$

Here, $\Sigma f d^{1^2} = 291$

$$\Sigma f d^1 = 70$$

$$N = 100$$

$$C = 10$$

Substituting the values in the formula

$$\sigma = \sqrt{\frac{291}{100} - \left(\frac{70}{100}\right)^2} \times 10$$

$$= \sqrt{2.91 - 0.49} \times 10$$

$$= \sqrt{2.42} \times 10$$

$$= 1.5556 \times 10$$

$$\sigma = 15.56$$

Coefficient of standard deviation $= \frac{\sigma}{\overline{X}}$

$$= \frac{15.56}{57}$$

$$= 0.273$$

As the coefficient of standard deviation is low the distribution is consistent.

**Check your progress - 1**

In a distribution $\Sigma dx = 50$, $\Sigma dx^2 = 1820$ and $N = 10$

Find out the standard deviation.

## 19.5 COEFFICIENT OF VARIATION

Standard deviation is an absolute measure and the relative measure of standard deviation is called the coefficient of variation . 'Karl Parson' introduced this measure in 1895. According to him the the coefficient of variation is "Percentage variation in the mean the standand deviation beling treated as the total variation in the mean". The coefficient of variation is used to compare the consistency and uniformity of two or more distributions. Lower the value of coefficient of variation means higher the consistency, more uniformity and lower variablity. Symbolically, coefficient of variation can be expressed as

$$C.V = \frac{\sigma}{\overline{X}} \times 100$$

Where C.V. = Coefficient of variation,

$\sigma$ = Standard deviati on

$\overline{X}$ = arithmetic me an

Both coefficient standard deviation and the coefficient of variation are the relative measures of dispersion. The distinguishng feature between them is that, while coefficient of standard deviation is the ratio of standard deviation to mean, the coefficient of variation is its percentage.

**Illustration :** IX. Calcualte the coefficient of variation and compare the consistency of the following distributions.

| Class | 0-50 | 50-100 | 100-150 | 150-200 | 200-250 | 250-300 |
|---|---|---|---|---|---|---|
| Distributions'A' | 2.3 | 1.7 | 2.0 | 3.3 | 2.7 | 4.7 |
| 'B' | 1.8 | 3.2 | 4.0 | 4.8 | 5.0 | 2.4 |

**Solution :** CALCULATION OF STANDARD DEVIATION AND COEFFICIENT OF STANDARD DEVIATION

| | | Distribution 'A' | | | | Distribution 'B' | | |
|---|---|---|---|---|---|---|---|---|
| Class | MV | f | $d^1$ $(\frac{MV-A}{C})$ | $fd^1$ | $fd^{1^2}$ | f | $fd^1$ | $fd^{1^2}$ |
| 0 - 50 | 25 | 2.3 | −2 | −4.6 | 9.2 | 1.8 | −3.6 | 7.2 |
| 50 - 100 | 75 | 1.7 | −1 | −1.7 | 1.7 | 3.2 | −3.2 | 3.2 |
| 100 - 150 | 125 | 2.0 | 0 | 0 | 0 | 4.0 | 0 | 0 |
| 150 - 200 | 175 | 3.3 | 1 | 1 | 3.3 | 4.8 | 4.8 | 4.8 |
| 200 - 250 | 225 | 2.7 | 2 | 5.4 | 10.8 | 5.0 | 10.0 | 20.0 |
| 250 - 300 | 275 | 4.7 | 3 | 14.1 | 42.3 | 2.4 | 7.2 | 21.6 |
| | | N = 16.7 | | $\Sigma fd^1 = 16.5$ | $\Sigma fd^{1^2} = 67.3$ | N = 21.2 | $\Sigma fd^1 = 15.2$ | $\Sigma fd^{1^2} = 56.8$ |

Note : Common factor is used while taking the deviations.

**Distribution : 'A'**

Arithmetic Mean $\overline{X} = A + \frac{\Sigma fd^1}{N} \times C$

Here,

$$A = 125$$

$$\Sigma fd^1 = 16.5$$

$$C = 50$$

$$N = 16.7$$

Substituting the values inthe formula,

$$\overline{X} = 125 + \frac{16.5}{16.7} \times 50$$

$$= 125 + \frac{825.0}{16.7} \times 50$$

$$= 125 + 49.40$$

$$\overline{X} = 174.40$$

Standard deviation $\sigma = \sqrt{\frac{\Sigma fd^{1^2}}{N} - \left(\frac{\Sigma fd^1}{N}\right)^2} \times C$

Here,

$$\Sigma fd^{1^2} = 67.3$$

$$\Sigma fd^1 = 16.5$$

$$C = 50$$

$$N = 16.7$$

Substituting the values in the formula,

$$\sigma = \sqrt{\frac{67.3}{16.7} - \left(\frac{16.5}{16.7}\right)^2} \times 50$$

$$= \sqrt{4.03 - 0.98} \times 50$$

$$= \sqrt{3.05} \times 50$$

$$= 1.75 \times 50$$

$$\sigma = 87 .5$$

Coefficient of variation (CV)$= \frac{\sigma}{X} \times 100$

$$= \frac{87.25}{174.40} \times 100$$

$$= 0.5003 \times 100 = 50.03\ \%$$

**Distribution 'B'**

Arithmetic Mean $\overline{X} = A + \frac{\Sigma fd^1}{N} \times C$

Here,

$$A = 125$$

$$\Sigma f d^1 = 15.2$$

$$N = 21.2$$

$$C = 50$$

Substituting the values in the formula

$$\overline{X} = 125 + \frac{15.2}{21.2} \times 50$$

$$= 125 + 35.85$$

$$\overline{X} = 160.85$$

$$\text{Standard deviation} = \sqrt{\frac{\Sigma f d^{1^2}}{N} - \left(\frac{\Sigma f d^1}{N}\right)^2} \times C$$

Here, $\Sigma f d^{1^2} = 56.8$

$$\Sigma f d^1 = 15.2$$

$$N = 21.2$$

$$C = 50$$

Substituting the values in the formula

$$\sigma = \sqrt{\frac{56.8}{21.2} - \left(\frac{15.2}{21.2}\right)^2} \times 50$$

$$= \sqrt{2.68 - 0.51} \times 50$$

$$= \sqrt{2.17} \times 50$$

$$= 1.47 \times 50$$

$$\sigma = 73.5$$

$$\text{Coefficient of variation (CV)} = \frac{\sigma}{\overline{X}} \times 100$$

$$\text{C.V.} = \frac{73.5}{160.85} \times 100$$

$$= 0.4569 \times 100$$

$$= 45.69\%$$

Since the coefficient of variation is low for distributioin 'B' it is more consistent and less variable.

**Illustration:** X. Following data relate to the daily turnover of two factories find out which factory is more consistent.

| | | Factory A | Factory B |
|---|---|---|---|
| Standard deviation | = | 29 | 31 |
| Mean | = | 45 | 50 |
| Number of observations | = | 100 | 100 |

**Solution**: CALCULATION OF COEFFICIENT OF VARIATION

**Facotry 'A':**

$$C.V. = \frac{\sigma}{\bar{X}} \times 100$$

$$= \frac{29}{45} \times 100$$

$$= 0.6444 \times 100$$

$$= 64.44\ \%$$

**Factory 'B':**

$$C.V. = \frac{\sigma}{\bar{X}} \times 100$$

$$= \frac{31}{50} \times 100$$

$$= 0.62 \times 100$$

$$= 62\ \%$$

Factory 'B' is more consistent in the matter of daily turnover as its coeffficient of variation is lower than that of Factory 'A'.

**Check your progress - 2**

In a distribution arithmetic mean is 120 and standard deviation is 8. Find out the coefficient of variation.

## 19.6 VARIANCE

The term variance was used by R.A. Fisher in 1913, according to him "it is the square of the standard deviation ($\sigma^2$)." In the words of william Greenwald the variance is "The mean of the squared deviations about the mean of a series". Thus symbolically variance may be expressed as:

$$V = \frac{\Sigma(X-\overline{X})^2}{N}$$

Where,

V = variance

$\Sigma(X - \overline{X})$ = Sum of deviations taken from the mean.

When deviations are taken from assumed mean, variance can be calculated by the following formula.

$$\text{variance} = \frac{\Sigma fd^{1^2}}{N} - \left(\frac{\Sigma fd^1}{N}\right)^2$$

Variance and standard deviation are closely related to each other, since, variance is the square of standard deviation and standard deviation is the square root of variance. In a given frequency distribution, a low degree of variance indicates high uniformity, homogenity and conversly less variable.

## 19.7 COMBINED STANDARD DEVIATION

When standard deviation for two or more distributions are combined, it is called the combined standard deviation. The procedure for its calculation is explained below.

(i) Compute the combined mean by adopting the following formula.

$$\overline{X}_{12.....n} = \frac{N_1\overline{X}_1 + N_2\overline{X}_2.....N_n\overline{X}_n}{N_1 + N_2.....N_n}$$

Where $\overline{X}_{12.....n}$ = Combined mean

N = Number of observations

$\overline{X}$ = Arithmetic average

(ii) Find the deviations from the combined mean, by subtracting combined mean from the mean of the series.

Thus,

$$d_1 = (\overline{X}_1 - \overline{X}_{12}) \quad \text{and} \quad d_2 = (\overline{X}_2 - \overline{X}_{12})$$

(iii) Combined standard deviation can be calculated by the following formula

$$\sigma_{12.....n} = \frac{N_1{\sigma_1}^2 + N_2{\sigma_2}^2.....N_n{\sigma_n}^2 + N_1{d_1}^2 + N_2{d_2}^2}{N_1 + N_2.....N_n}$$

Where, $\sigma_{12.....n}$ = Combined standard deviation.

N = Total number of observations

$\sigma$ = Standard deviation

$d^2$ = Square of the difference between actual mean and combined mean.

**IllustrationXI :** Find the combined standard deviation of two groups

| | Group I | Group II |
|---|---|---|
| Mean $(\overline{X})$ | 12 | 15 |
| standard deviation $(\sigma)$ | 6 | 8 |
| Number of observations | 25 | 25 |

**Solution :** CALCULATION OF COMBINED STANDARD DEVIATION.

COMBINED MEAN OF TWO GROUPS :

$$\overline{X}_{12} = \frac{N_1\overline{X}_1 + N_2\overline{X}_2}{N_1 + N_2}$$

Here,

$N_1 = 25$ and $N_2 = 25$

$\overline{X} = 12$ and $\overline{X}_2 = 15$

substituting the values in the formula

$$\overline{X}_{12} = \frac{(25)(12)+(25)(15)}{25+25}$$

$$= \frac{300+375}{50}$$

$$= \frac{675}{50}$$

$$= 13.5$$

combined standard deviation :

$$\sigma_{12} = \frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2d_2^2}{N_1 + N_2}$$

$$d_1 = (\overline{X}_1 - \overline{X}_{12})$$

$$= (12 - 13.5)$$

$$= -1.5$$

$$d_2 = (\overline{X}_2 - \overline{X}_{12})$$

$$= (15 - 13.5)$$

$$= 1.5$$

Here, $N_1 = 25$ and $N_2 = 25$

$\sigma_1 = 6$ and $\sigma_2 = 8$

$d_1 = -1.5$ and $d_2 = 1.5$

substituting the values in the formula

$$\sigma_{12} = \frac{(25)(6)^2+(25)(8)^2+(25)(-1.5)^2+(25)(1.5)^2}{25+25}$$

$$= \frac{(25)(36)+(25)(64)+(25)(2.25)+(25)(2.25)}{50}$$

$$= \frac{900+1600+56.25+56.25}{50}$$

$$= \frac{2612.5}{50}$$

$$\sigma_{12} = 52.25$$

Hence the combined standard deviation of two groups is 52.25.

## 19.8 RELATIONSHIP AMONG VARIOUS MEASURES OF DISPERSION

Various measures of dispersion are independent of each other. In case of normal distribution of the data these measures show the following relationship in terms of coverage of items.

| | | |
|---|---|---|
| Mean $\pm$ Q.D | Covers | 50% of the items |
| Mean $\pm$ M.D. | Covers | 57.5% " " |
| Mean $\pm \sigma$ | Covers | 68.27% " " |
| Mean $\pm 2\sigma$ | Covers | 95.45% " " |
| Mean $\pm 3\sigma$ | Covers | 99.73% " " |

Mean + Quartile Deviation covers only 50% of the items in the distribution. On the other hand, Mean + Standard Deviation covers 68.27% of the items. Further, Standard Deviation together with mean value explains greater percentage of items in the given distribution than any other combination of measures of dispersion with mean. As such, standard deviation ensures more reliable results and quartile deviation gives the least reliable results.

## 19.9 MERITS AND LIMITATIONS OF STANDARD DEVIATION

**Merits**

i) Standard deviation is a well defined measure of dispersion as compared to other measures such as Quartile deviatioin and mean deviation.

ii) It is amenable to further algebraic treatment

iii) It is based on all observations in the distribution. Therefore it can be treated as more representative measure.

iv) As it yields almost the same value for different samples drawn from the same population, it is considered to be least affected by the sampling fluctuations.

v) This measure has greater reliability and therefore it is commonly used in small samples.

**Limitations**

a) It is not easy to calculate the standard deviation as it involves calculation of squares, square roots etc.

b) Its value is affected by extreme items in the distribution.

Despite these limitations it is widely used in comparing the variability of different distributions. It is also used in different statistical techniques like skewness, kurtosis, correlation analysis, regression analysis, tests of significance, analysis of variance etc. It is also considered as the unit of measurement in normal distribution.

## 19.10. GRAPHIC METHOD OF STUDYING DISPERSON

The Dispersion in a distribution can also be studied by graphic methods. Lorenz curve is a popular method of presenting variability of a distribution on a graph. It has been first developed by Max O.Lornez, who used it to measure the variability in the distribution of income and wealth. This technique can also be used to study the distribution of profits, wages, turnover, etc.

The Lorenz curve is a cumulative percentage curve. A diagonal line drawn across 'O' of 'X' axis and '100' of 'Y' axis is called line of equal distribution. This is shown in fig.19.1. The departure of frequency curves from this line shows the degree of inequality of variability in the distribution. Greater departure of curve from the line of equal distributiion reveals greater variablity and less consistency and vice-versa.

The following procedure is adopted to construct Lorenez curve.

i) Find out cumulative totals and cumulative percentages for the size of the items (values) and for the frequencies. The percentages can be obtained by taking the cumulative total as 100 and obtaining the percentages of each value. In case of continuous classes, mid values are taken to find out the percentages.

ii) Plot the cumulative percentages of frequencies starting from 100 to 0 on 'X' axis whereas on 'Y'axis the cumulative percentage of variables are to be shown.

iii) Draw a diagonal line joining 0 and 100 . This line is called the line of equal distribution.

iv) Join the percentage values of each variable plotted on the graph to obtain the frequency curves. Dispersion is measured by taking the extent of departure of these curves from the line of equal distribution.

**Illustration XII:** Following data relates to wages paid in two factories Draw Lorenz curve to measure the dispesion of wages.

| Wages (in Rs.) : | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 |
|---|---|---|---|---|---|---|---|---|
| No of persons | | | | | | | | |
| Factory : A | 10 | 22 | 30 | 50 | 40 | 20 | 20 | 8 |
| Factory : B | 8 | 12 | 30 | 50 | 40 | 26 | 12 | 22 |

Solution : Wages 'X' Number of Persons 'f'

PREPARATION OF DATA FOR LORENZ CRUVE

| | | | FACTORY 'A' | | | FACTORY 'B' | | |
|---|---|---|---|---|---|---|---|---|
| (X) | Cumulative totals | Cumulative percentage | f | Cumulative totals | Cumulative percentages | f | cumulative totals | Cumulatve percentages |
| 5 | 5 | 2.8 | 10 | 10 | 5 | 8 | 8 | 4 |
| 10 | 15 | 8.3 | 22 | 32 | 16 | 12 | 20 | 10 |
| 15 | 30 | 16.7 | 30 | 62 | 31 | 30 | 50 | 25 |
| 20 | 50 | 27.8 | 50 | 112 | 56 | 50 | 100 | 50 |
| 25 | 75 | 41.7 | 40 | 152 | 76 | 40 | 140 | 70 |
| 30 | 105 | 58.3 | 20 | 172 | 86 | 26 | 166 | 83 |
| 35 | 140 | 77.8 | 20 | 192 | 96 | 12 | 178 | 89 |
| 40 | 180 | 100.0 | 0 | 200 | 100 | 22 | 200 | 100 |

Fig. 19.1

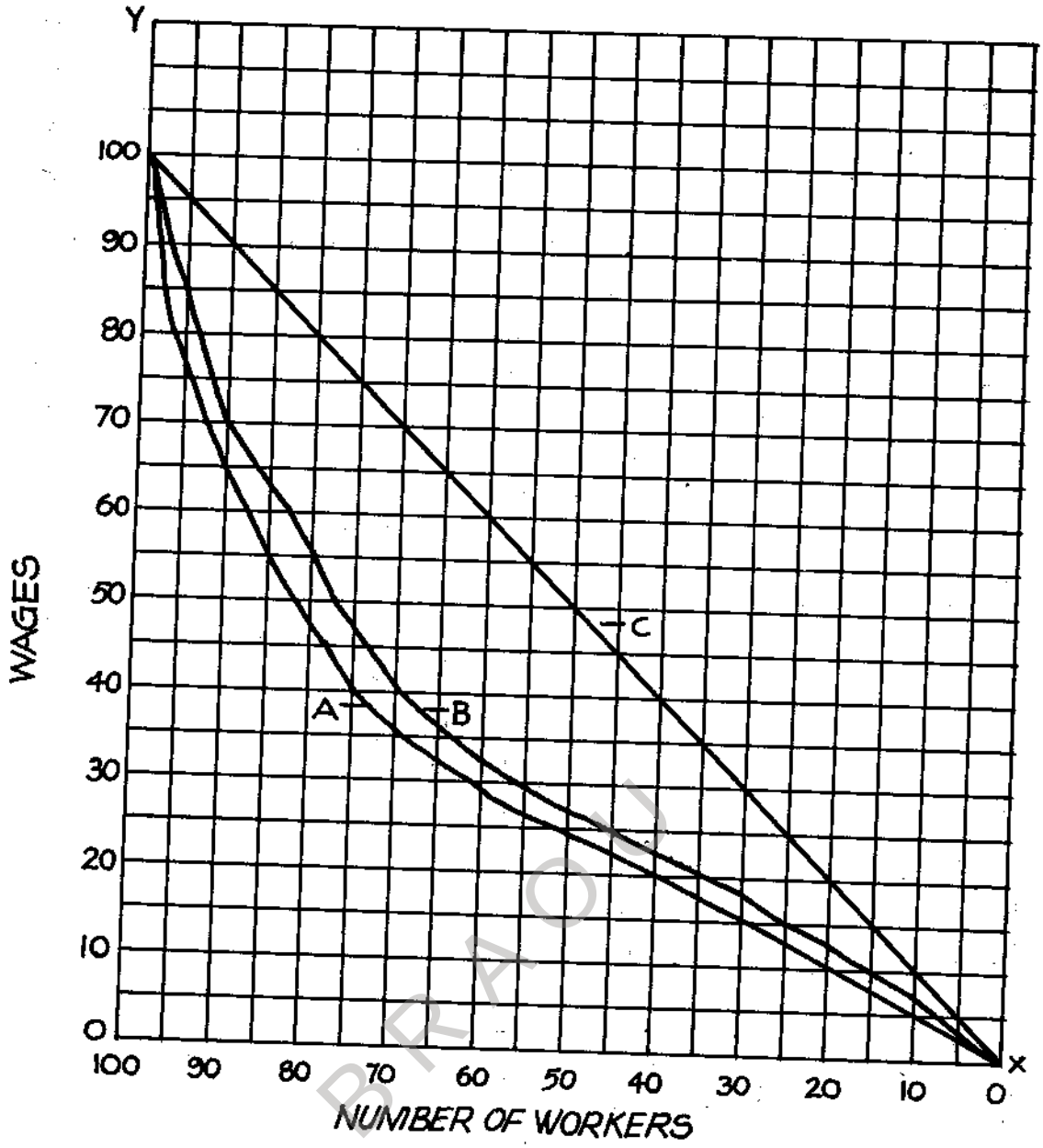

Y
100
90
80
70
60
50
40
30
20
10
0
WAGES
A
B
C
100
90
80
70
60
50
40
30
20
10
0
X
NUMBER OF WORKERS

Fig 19.1 Showing lorenz curve of Wages and number of persons.

oy: 1 cm = Rs.10/-

ox: 1 cm = 10 persons

A = Factory

B = Factory

C = Line of equal distribution.

As the cumulative percentage curve of wages paid in factory 'B' is nearer to the line of equal distribution, it can be concluded that factory 'B' is consistent in the matter of payment of wages.

The Lorenz curve is proved to be quite advantageous becuase the variations can be easily understood as this is visual aid. It is preferred in socioeconomic data relating to wealth and income of the people. But, the main limitation of this technique is that the variability cannot be directly qualified.

## 19.11 SUMMING UP

Standard deviation is an improved measure of dispersion. It is the square root of the average of squared deviations from the arithmetic mean. Co-efficient of variation is the ratio of standard deviation to arithmetic mean expressed as percentage. It is commonly used to make comparisons effectively. Though standard deviation and co-efficient of variation are widely used as appropriate measures of variation, their compilation is difficult and time consuming.

Lorenz curve is the graphic method of studying dispersion in which the variablity is compared with reference to the distance of the curve to the line of equal distribution.

## 19.12 CHECK YOUR PROGRESS : MODEL ANSWERS

1. Appy the following formula

$$\text{Standard deviation} = \sqrt{\frac{\Sigma dx^2}{N} - \left(\frac{\Sigma dx}{N}\right)^2}$$

$$= 12.53$$

2. Coefficient of Variations $= \frac{\sigma}{\bar{X}} \times 100$

$$= 6.67\%$$

## 19.13 MODEL EXAMINATION QUESTIONS

**A. Short Questions**

1.What is standard deviation?

2. How do you calculate step deviations?

3. What is coefficient of variation?

4. What is variance ?

5. What is combined standard deviation?

6. What is Lorenz curve?

7. Why do you call standard deviation as 'Root mean Square deviation'?

8. What are the limitations of standard deviation?

9. Distinguish between mean deviation and standard deviation.

10. Distinguish between coefficient of standard deviation and coefficient of variation.

11. How do you calculate combined standard deviation for two series?

12. Explain the merits and limiations of standard deviation.

**B. Essay Questions**

13. Discuss the relative merits and limitations of various measures of dispersion.

14. Why do you consider standard deviation superior to other measures of dispersion?

15. "Measures of dispersion and central tendency are complementary to each other in hghlighting the characteristics of a frequency distribution". Discuss with suitable examples.

16. Explain the significance of Lorenz curve? Explain the procedure for construction of Lorenz curve.

## EXERCISES

17. Calcualte standard deviation.

Weekly incomes (Rs.) : 135, 175, 129, 141, 109, 101, 182, 92

(Ans : 30.68)

18. Find out standard deviation and its coefficient.

60, 72 84 75 69 50 80 105 102 115

(Ans : 19.6 and 0.24)

19. Compute standard deviation.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Exports (in Rs. 000's) | : | 50 | 60 | 70 | 80 | 90 | 100 |
| Number of Companies | : | 14 | 25 | 19 | 16 | 22 | 6 |

(Ans : 15.10 and 0.21)

20. Given below are weekly wages paid to workers of two factories. Find out which factory is consistent.

| Weekly wages (Rs.) | 100 | 120 | 140 | 160 | 180 | 200 | 220 | 240 |
|---|---|---|---|---|---|---|---|---|
| Number of workmen | | | | | | | | |
| Factory - A : | 60 | 75 | 45 | 150 | 110 | 120 | 130 | 160 |
| Factory - B : | 126 | 156 | 142 | 136 | 113 | 135 | 181 | 172 |

(Ans : Factory B)

21. Calculate standard deviation of marks obtained by 100 students in statistics

| Marks : | 30 | 35 | 40 | 45 | 50 | 55 | 60 | 65 | 70 |
|---|---|---|---|---|---|---|---|---|---|
| Number of students : | 12 | 15 | 17 | 16 | 15 | 10 | 13 | 18 | 1 |

(Ans : 11.61)

22. Compute coefficient of standard deviation.

| Price (Rs.) : | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 |
|---|---|---|---|---|---|---|---|---|
| Number of units sold : | 23 | 68 | 77 | 72 | 41 | 24 | 34 | 11 |

(Ans : 0.47)

23. Calculate standard deviation and coefficient of standard deviation.

| Class : | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 |
|---|---|---|---|---|---|---|
| Frequency: | 6 | 10 | 12 | 21 | 13 | 8 |

(Ans : 14.39 and 0.44)

24. Find out standard deviation.

| Class : | 500-600 | 600-700 | 700-800 | 800-900 | 900-1000 |
|---|---|---|---|---|---|
| Frequency : | 5 | 8 | 15 | 6 | 6 |

(Ans : 1.204)

25. Compute coefficient of standard deviation.

| Age : | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 | 70-80 |
|---|---|---|---|---|---|---|---|
| Frequency: | 37 | 17 | 18 | 25 | 27 | 36 | 32 |

(Ans : 0.51)

26. Calculate standard deviation and coefficient of variation.

| Class | : 1000-2000 | 2000-3000 | 3000-4000 | 4000-5000 | 5000-6000 |
|---|---|---|---|---|---|
| frequency: | 3 | 9 | 12 | 10 | 6 |

(Ans : 1090 and 0.29)

27. Two brands of electronic type writers are tested and the following results are obtained.

| | Brand-'A' | Brand- 'B' |
|---|---|---|
| Mean | 65 | 80 |
| Standard deviation | 8 | 12 |
| Total number of observations | 200 | 200 |

Compare the consistency.

(Ans : A = 12.3% and B = 15%)

28. From the following data calculate combined standard deviation.

| | Sample-I | Sample-II |
|---|---|---|
| Mean | 16 | 14 |
| Variance | 25 | 9 |
| Total Number of observations | 50 | 50 |

(Ans : 8 and 2.5)

29. Construct lorenz curve from the following data.

| Class | : 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 |
|---|---|---|---|---|---|---|
| frequency | | | | | | |
| A | : 20 | 30 | 50 | 60 | 20 | 20 |
| B | : 10 | 30 | 20 | 50 | 40 | 50 |

30. The coefficent of variation of a series is 80%. The standard deviation is 16. Find out the arithmetic mean.

(Ans : 20)

31. Find out standard deviation and coefficient of variation.

$$\Sigma f dx = 50$$

$$\Sigma f dx^2 = 600$$

$$N = 100$$

Assumed Mean $= 35.5$

(Ans 2.4 and 0.07)

## 19.4 RECOMMENDED BOOKS


1. Gupta, S.P : "Statistical methods", Sultan chand & company, New Delhi.
2. Gupta, B.N. : "Statistics", Sahitya Bhavan, Agra.
3. Gupta, C.B : "Fundamentals of statistics", Himalaya publishing house, Bombay.
4. Simpson and Kafka : "Basic Statistics", Oxford and I.B.H. Publishing Company Calcutta.


## 19.15 GLOSSARY

1. Co-effcient of variation : It is the relative measure of standard deviation.
2. Combined standard deviation : When standard deviations of two distributions are combined it is called combined standard deviation.
3. Standard deviation : It is the square root of the mean of the squared deviations measured from arithmetic mean.
4. Variance : It is the square of the standard deviation.

## Unit-20 : CONCEPT OF SKEWNESS

**Contents**



## 20.0 AIMS AND OBJECTIVES

The aim of this unit is to explain the nature, scope and importance of skewness.

After going through this unit, you should be able to :

- difine skewness
- identify the objectvies of skewness
- identify the types of skewness
- recognise the nature of skewness
- distinguish between dispersion and skewness.



## 20.1 INTRODUCTION

An average is a single figure which gives a representative value of the series and depicts the characterstics of the whole group. Averages alone cannot adequately describe the whole set of observations as all the observations in a frequency distribution are seldom alike. Measures of dispersion are useful in examining the scatteredness of the mass of figures in a series from an average value. However, neither the averages nor the measures of dispersion reveal the entire story of a frequency distribution. For example, two distributions having the same mean and standard deviation may differ widely in their overall appearance which can be seen from the following (figs : 20.1 and 20.2) :

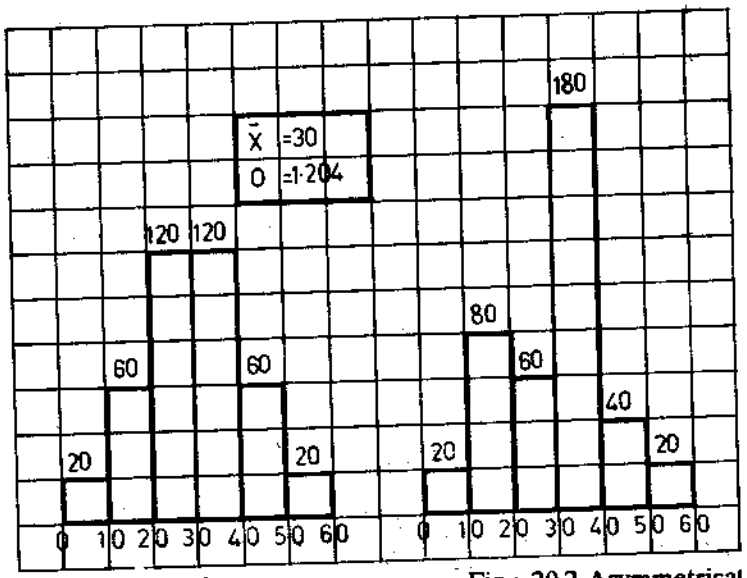



Fig : 20.1 Symmetrical distribution

Fig : 20.2 Asymmetrical distribution

Though the values of mean and standard deviation of both the distributions are same (Mean=30, Standard deviation=1.204), it does not mean that both the distributions are alike. While the frequency distribution in Fig.20.1 is symmetrical, the frequency distribution in Fig.20.2 is asymmetrical or skewed. Measures of Skewness help in understanding the nature and shape of the distribution.

## 20.2. DEFINITION OF SKEWNESS

Some of the important definitions of Skewness are given below :

(i) According to Riggleman and Frisbee', "Skewness is the lack of symmetry. When a Frequency distribution is plotted on a chart, skewness present in the items tends to be dispersed more on one side of the mean than on the other."

(ii) According to Garrett, "A distribution is said to be skewed when the mean and the median fall at different points in the distribution, and the balance (or centre of gravity) is shifted to one side or the other-to left or right".

(iii) 'Wessel, Willett and Simone'. defined skewness as "lack of symmetry. Any measure of skewness indicates the difference between the manner in which items are distributed in a particular distribution compared with a normal distribution".

(iv) According to Paden and Lindquist', "A distribution is said to be skewed, if it is lacking symmetray, that is, measures tend to pile up at one end or the other."

(v) According to 'Simpson and Kafka',"Measures of skeness tell us the direction and the extent of skewness. In symmetrical distribution the mean, median and mode are indentical. The more the mean moves away from the mode, the larger the asymmetry or skewness."

An analysis of the above definitions reveals the following:

(i) Skewness is the reslut of lack of symmetry.

(ii) If frequency distribution is plotted on a graph, items tend to disperse more on one side of the mean.

(iii) Measures of skewness indicate the direction and the extent of skewness.

(iv) In a skewed distribution, the values of mean, mode and median are not equal.

(v) In a distribution, the more the mean moves away form the mode, the larger would be the skewness.

(vi) Skewness indicates the deviation of items of a particular distribution from that of a normal distribution.

## 20.3. OBJECTIVES OF SKEWNESS

The concept of skewness has gained the importance due to the fact that statistical theory is often based upon the assumption of the normal distribution. The significance of skewness will be clear from the following objectives which it accomplishes :

(i) Skewness helps us to understand the nature and degree of concentration of values in a frequency distribution. We can also know whether the concentration of values is on the higher side or the lower.

(ii) Skewness helps us to ascertain the empirical relations of various averages such as mean, median, mode, etc. They are based on a moderately skewed distribution.

(iii) Measures of skewness helps to ascertain the extent to which a given distribution,departs from normality. Since many of the statistical measures are based on the assumption of a normal distribution, the study of measures of skewness is imperative.

## 20.4. TYPES OF FREQUENCY DISTRIBUTIONS AND SKEWNESS

Frequency distributions may be either symmetrical or asymmetrical

i) **Symmetrical distribution :** A frequency distribution is said to be symmetrical when it fulfils the following conditions :

a) Mean, median and mode are equal.

b) Data when plotted on a graph paper give the normal bell shaped curve.

c) Sum of the positive deviations which forms the median or mode is equal to the sum of the negative deviations.

d) Lower and upper quartiles, decile one and nine, and', percentile ten and ninety are equi-distant from the median.

e) Frequencies on either side of the mode are equal.

f) Frequencies on either side of the mode go on increasing upto a point and then begin to decreage in the same fashion.

While some of the frequency distributions are symmetrical and bell shaped or normal,

others are symmetrical but not bell-shaped. A special kind of symmetrical distribution is the normal distribution which is not only symmetrical but also gives a perfect bell shaped curve. The following are some of the possible patterns of the symmetrical distribution (fig: 20.3 to 20.7).

## Check your Progress - 1

Explain the properties of symmetrical distribution.

Fig : 20.3

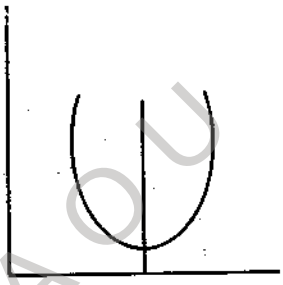


Fig : 20.4

Figs : 20.3 and 20.4 Symmetrical but not the bell-shaped or normal distribution



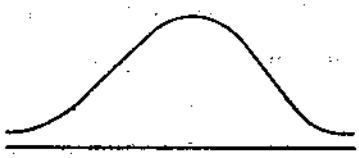


Fig : 20.5 : Symmetrical and bell-shaped (Normal distribution)

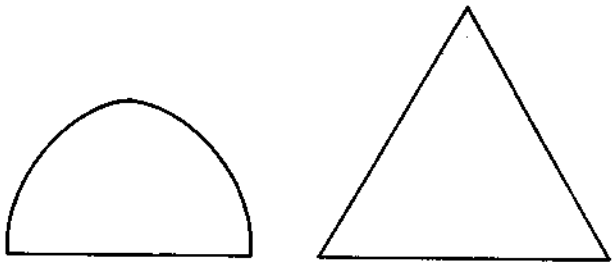


Fig : 20.6 Fig : 20.7

figs : 20.6 and 20.7 : Symmetrical but not the bell-shaped (normal distribution)

ii) **Skewed or asymmetrical distribution :** Any distribution that departs from normality is called a skewed distribution. Skewed or asymmetrical distributions may be either positively skewed or negatively skewed.

a) **Positively skewed distribution :** In a positively skewed distribution the value of mean is maximum, and that of the mode is minimum. Median lies in between these two extremes. In a positively skewed distribution, more than half of the area under the curve falls on the right side of the mode and its right tail is longer than its left tail. under such a distribution mean is greater than the median, and the median is greater than the mode (Mean > Median > mode). The difference between upper quartile and median is greater than the difference between median and lower quartile $(Q_3 -$ Med) > (Med $- Q_1)$. The following figure illustrates the properties of positively skewed distribution.



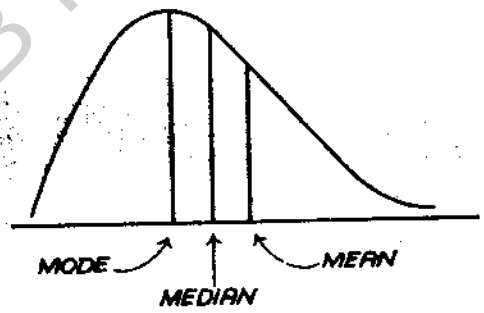



Fig : 20.8 Positively skewed distribution

b) **Negatively skewed distribution :** In a negatively skewed frequency distribution, mean is less than median, and median is less than mode (Mean < Median < Mode ). In other words, the value of mode is maximum and that of mean is minimum. The value of median lies in between these two values. More than half of the area under the distribution curve falls on the left side of the mode. Thus, the excess tail be on

the left-handside of the curve. The difference between median and lower quartile $(Q_3 - M) < (M - Q_1)$. The following figure depicts the negatively skewed distribution.

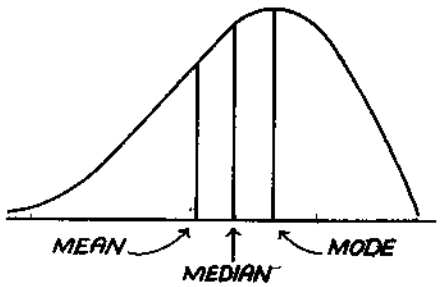



Fig : 20.9 Negatively skewed distribution

## 20.5. TESTS OF SKEWNESS

The following tests can be applied to ascertain whether a frequency distribution is symmetrical or skewed :

(i) In an asymmetrical or skewed distribution mean, median and mode are not equal. The greater the difference between mean and mode, the more would be the skewness in the distribution.

(ii) In a moderately skewed distribution Mean = Mode + 2/3 of (Median − Mode).

(iii) The data of an asymmetrical distribution when plotted on a graph paper do not give a bell - shaped curve. Thus, the vertical line through the centre does not cut the curve into two equal halves.

(iv) In a skewed distribution the sum of positive deviations from median or mode is not equal to the sum of negative deviations. Thus, the degree of skewness depends upon the difference between the sum of positive and negative deviations from median or mode.

(v) In a skewed distribution, frequencies on either side of the mode are not equal.

(vi) In a skewed distribution the values of lower quartile and upper quartile, decile one to decile nine and percentile ten to percentile ninety are not equi-distant from median.

The tests desicribed will be clear from the following Tables :

**TABLE -1**

**VARIOUS FREQUENCY DISTRIBUTIONS AND THE POSITION OF AVERAGE**

| Size | I Frequency | II Frequency | III Frequency |
|---|---|---|---|
| 20 | 2 | 2 | 3 |
| 40 | 6 | 18 | 4 |
| 60 | 10 | 10 | 7 |
| 80 | 14 | 8 | 7 |
| 100 | 10 | 6 | 10 |
| 120 | 6 | 4 | 18 |
| 140 | 2 | 2 | 2 |
| Skewness | Symmetry | Positive Skewness | Negative Skewness |
| Average based Computations of Skewness | $\overline{X}$ = Med = $M_0$<br>80 = 80 = 80 | $\overline{X}$ > Med > $M_0$<br>67.2 > 60 > 40 | $\overline{X}$ < Med < $M_0$<br>92.4 < 100 < 120 |
| Quartile based Computations of Skewness | ($Q_3$ − Med) = (Med − $Q_1$)<br>(100 − 80) = (80 − 60) | ($Q_3$ − Med) > (Med − $Q_1$)<br>(100 − 67.2) > (67.2 − 40) | ($Q_3$ − Med) < (Med − $Q_1$)<br>(120 − 100) < (100 − 60) |
| Shape of Frequency Curves | Normal | Skewed to the right | Skewed to the left |

**Note :** $\overline{X}$ = Mean; Med. = Median; $M_0$ = Mode; $Q_3$ =Upper Quartile; $Q_1$ = Lower Quartile.



## 20.6. DISTIONCTION BETWEEN DISPERSION AND SKEWNESS

**The following are some of the important points of difference between dispersion and skewness.**

| | Despersion | Skewness |
|---|---|---|
| (i) | It is a measure of variation of the items in a frequency distribution from a central value. | It is a measure of symmetry of distribution of values on both sides of the central value. |

| | | |
|---|---|---|
| (ii) | It is the 'average of second order 'because it is an average of deviations around a central value. | It is not an average but its measurment is based on various types of averages such as mean, median, mode, etc. |
| (iii) | It measures the degree of variability in data. | It helps to find out whether the concentration of values is on positive or negative side. |
| (iv) | It is concerned with the variability of values in general. | It is concerned with the symmetry of distribution on either side of the mode. |
| (v) | It is concerned with the general shape of frequency distribution. | It is concerned with the dispersion on the two sides of the mode with regard to the arrangement of frequencies. |
| (vi) | It indicates the extent to which a mean is representative of individual values in a frequency distribution. | It helps in judging the extent to which a given frequency distribution departs from normality. |

## 20. 7. SUMMING UP

Skewness is the result of lack of symmetry. It indicates the extent to which a given distribution departs from normality. In a skewed distribution mean, median and mode are not equal. The more the mean moves away from the mode, the larger would be the skewness. In a positively skewed distribution, mean is greater than median and median is greater than mode. On the other hand, in a negatively skewed distribution mode is greater than median and median is greater than the mean. Further, the difference between the upper quartile and median is lesser than the difference between median and lower quartile. In a moderately skewed distribution :

Mean = Mode + 2/3 (Median − Mode)

## 20.8 CHECK YOUR PROGRESS : MODEL ANSWERS

1. The properties of symmetrical distribution are :

i) Mean, Median and Mode are equal.

ii) When data are plotted on a graph paper, it gives a bell shaped curve.

iii) Sum of positive deviations from median or mode will be equal to the sum of negative deviations.

iv) $Q_1$ and $Q_3$, $D_1$ and $D_9$ and $P_{10}$ and $P_{90}$ are equi-distant from the median.

v) Frequencies on either side of mode are equal.

vi) Frequencies on either side of mode go on increasing up to a point and then start decreasing in the same manner.

## 20.9. MODEL EXAMINATION QUESTIONS

**A. Short Questions**

1. Give the meaning of 'Skewness'.
2. What is meant by 'Positive Skewness' ?
3. What do you mean by 'Negative Skewness' ?
4. Bring out the relationship among various averages in the case of a positlvely skewed distribution.
5. Bring out the relationship among various averages in the case of a negatively skewed distribution.
6. How are the frequencies spread in a skewed distribution.
7. Bring out the relationship among mean, mode and medain in the case of a moderately skewed distribution.

**B. Eassy Questions**

8. Define skewness and explain its salient features.
9. Expalin the significance of the study of skewness.
10. Distinguish between dispersion and skewness.
11. Expalin the various tests applied to find out the presence or absence of skewness in a distribution.



## 20.10. RECOMMENDED BOOKS


1. Gupta, S.P. : "Statistical Methods", Sultan chand & Comapny, New Delhi.
2. Gupta, B.N. : "Statistics", sahitya Bhavan, Agra.
3. Gupta, S.C. : "Fundamentls of Statistics", Himalaya Publishing House, Bombay.
4. Simpson and Kafka : " Basic statistics", Oxford and IBH publishing Company, Calcutta.

## 20.11. GLOSSARY

1 Asymmetrical or skewed distribution : A distribution that departs from symmetry and the mean, median and mode of such distribution will be unequal. The frequency curve of such data will extend to either right or left side of the X axis.

2 Negatively skewed distribution : In a negatively skewed distribution, Mean will be less than Median and Median will be less than Mode. The frequency curve of such distribution will be longer on the left hand side of 'X' axis.

3 Positively skewed distribution : In a positively skewed distribution Mean will be greater than Median and Median will be greater Mode. The frequency curve of such distribution will be longer on the right hand side of 'X' axis.

4 Skewness : Lack of symmetry in a distribution

5 Symmetrical distribution : A distribution is said to be symmetrical, whem mean, median and mode of that distribution are equal. If such data are plotted on a graph paper, it forms a bell shaped curve.

## UNIT - 21 : MEASURES OF SKEWNESS

Contents



## 21.0 AIMS AND OBJECTIVES

This unit aims at explaining the various measures of skewness with the help of suitable illustrations. After going through this unit, you should be able to :

– calculate absolute measures of skewness

– Calculate relative measures of skewness



## 21.1 INTRODUCTION

In unit 20 we dealt with the theoretical and graphical presentation of skewness. In this unit, an attempt is made to calculate the cofficient of skewness with the help of Karl person's and Bowley's formulae.

The extent and direction of asymmetry in a serics is calculated with the help of statistical measures of skewness. These measures may be either absolute or relative.

## 21.2 ABSOLUTE MEASURES OF SKEWNESS

Absolute measures of skewness tells us the extent of asymmetry in a series in absolute terms. According to this, skewness is calculated simply by obtaining the difference between the mean and mode of the given series.

Symbolically,

Absolute skewness $= \overline{X} - Z$

Accoridng to this method, if the value of mean is greater than mode, the distribution is said to be positively skewed. On the other hand, if the value of mode is greater than mean, the

**ANALYSIS TABLE**

| Col. No. | 0-5 | 5-10 | 10-15 | 15-20 | 20-25 | 25-30 | 30-35 | 35-40 | 40-45 |
|---|---|---|---|---|---|---|---|---|---|
| I | | | | | | 1 | | | |
| II | | | | | 1 | 1 | | | |
| III | | | | | | 1 | 1 | | |
| IV | | | | 1 | 1 | 1 | | | |
| V | | | | | 1 | 1 | 1 | | |
| VI | | | | | | 1 | 1 | 1 | |
| | | | | 1 | 3 | 6 | 3 | 1 | |

it is clear from the anlysis table that mode lies in the class 25-30.

Mode $= L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$

$L = 25, \quad f_1 = 60, \quad f_0 = 30, \quad f_2 = 30, \quad i = 5$

Substituting the values in the formula,

$$\text{Mode} = 25 + \frac{60-30}{2\times60-30-30} \times 5$$

$$= 25 + \frac{30}{120-60} \times 5$$

$$= 25 + \frac{150}{60}$$

$$= 25 + 2.5$$

$$= 27.5$$

$$SKp = \frac{Mean - Mode}{Standard\ Deviation}$$

Mean = 28.30, Mode = 27.50, S.D. = 10.015

Substitutin the values in the formula,

$$SKp = \frac{28.30-27.50}{10.015}$$

$$= \frac{0.80}{10.015}$$

$$= 0.08$$

Hence the distribution is positively skewed with very little degree of skewness.

**Illustration - 3**

Calculate Karl Person's coefficient of skewness from the following data :

| Class | : 10-14 | 15-19 | 20-24 | 25-29 | 30-34 | 35-39 |
|---|---|---|---|---|---|---|
| Frequency : | 3 | 10 | 20 | 35 | 18 | 4 |

**Solution :**

**CALCULATION OF KARL PESRSON'S COEFFICIENT OF SKEWNESS**

| Class | f | m.v. m | (m-22)/5 $d^1$ | $fd^1$ | $fd^{1^2}$ |
|---|---|---|---|---|---|
| 10-14 | 3 | 12 | − 2 | − 6 | 12 |
| 15-19 | 10 | 17 | − 1 | − 10 | 10 |
| 20-24 | 20 | 22 | 0 | 0 | 0 |
| 25-29 | 35 | 27 | 1 | 35 | 35 |
| 30-34 | 18 | 32 | 2 | 36 | 72 |
| 35-39 | 4 | 37 | 3 | 12 | 36 |
| | N = 90 | | | $\Sigma fd^1$ =67 | $\Sigma fd^{1^2}$ =165 |

Calculation of Mean :

$$\overline{X} = A + \frac{\Sigma fd^1}{N} \times C$$

$A = 22$, $\Sigma fd^1 = 67$, $N = 90$, $C = 5$

Substituting the values in the formula,

$$\overline{X} = 22 + \frac{67}{90} \times 5$$

$$= 22 + 3.72 = 25.72$$

Culculation of Standard Deviation :

$$\sigma = \sqrt{\frac{\Sigma fd^{1^2}}{N} - \left(\frac{\Sigma fd^1}{N}\right)^2} \times C$$

$\Sigma fd^{1^2} = 165$, $\Sigma fd^1 = 67$, $N = 90$, $C = 5$

Substituting the values in the formula

$$\sigma = \sqrt{\frac{165}{90} - \left(\frac{67}{90}\right)^2} \times 5$$

$$= \sqrt{1.83 - (0.744)^2} \times 5$$

$$= \sqrt{1.83 - 0.55} \times 5$$

$$= \sqrt{1.28} \times 5$$

$$= 1.13137 \times 5 = 5.6558$$

Calculation of Mode.

**GROUPING TABLE**

| Class | Col.I | Col.II | Col.III | Col.IV |
|---|---|---|---|---|
| 10-14 | 3 | | | |
| | | 13 | | |
| 15-19 | 10 | | | 33 |
| | | | 30 | |
| 20-24 | 20 | | | |
| | | 55 | | |
| 25-29 | 35 | | | |
| | | | 53 | |
| 30-34 | 18 | | | 57 |
| | | 22 | | |
| 35-39 | 4 | | | |

**ANALYSIS TABLE**

| Col.No. | 10-14 | 15-19 | 20-24 | 25-29 | 30-34 | 35-39 |
|---|---|---|---|---|---|---|
| I | | | | 1 | | |
| II | | | 1 | 1 | | |
| III | | | | 1 | 1 | |
| IV | | | | 1 | 1 | |
| | | | 1 | 4 | 2 | 1 |

It is clear from the analysis table that Mode lies in the class 25-29, the class boundaries of which are 24.5 - 29.5.

$$\text{Mode} = L + \frac{f_1 - f_o}{2f_1 - f_o - f_2} \times i$$

$L = 24.5, \quad f_1 = 35, f_o = 20, f_2 = 18, i = 5$

Substituting the values in the formula,

$$\text{Mode} = 24.5 + \frac{35-20}{2\times35-20-18} \times 5$$

$$= 24.5 + \frac{75}{32}$$

$$= 24.5 + 2.34 = 26.84$$

Calculation of skewness :

$$SKp = \frac{Mean - Mode}{Standard\ Deviation}$$

Mean = 25.72, Mode = 26.84, S.D. = 5.6568

Substituting the values in the formula,

$$SKp = \frac{25.72-26.84}{5.6568}$$

$$= \frac{-1.12}{5.6568}$$

$$= -0.198$$

Thus, the series is negatively skewed.

**Illustration - 4**

Calculate the coefficient of skewness for the following distribution of wages of 30 workers in a factroy :

| Wages more than (Rs.) : | 200 | 210 | 220 | 230 | 240 | 250 | 260 | 270 | 280 |
|---|---|---|---|---|---|---|---|---|---|
| No. of earners : | 30 | 26 | 24 | 21 | 18 | 14 | 8 | 3 | 0 |

**Solution :**

COMPUTATION OF KARL PEARSON'S COEFFICIENT OF SKEWNESS

| Wages (Rs.) | f | (m.v.) m | (m−255)/10 $d^1$ | $fd^1$ | $fd^{1^2}$ |
|---|---|---|---|---|---|
| 200-210 | 4 | 205 | −5 | −20 | 100 |
| 210-220 | 2 | 215 | −4 | − 8 | 32 |
| 220-230 | 3 | 225 | −3 | − 9 | 27 |
| 230-240 | 3 | 235 | −2 | − 6 | 12 |
| 240-250 | 4 | 245 | −1 | − 4 | 4 |
| 250-260 | 6 | 255 | 0 | 0 | 0 |
| 260-270 | 5 | 265 | 1 | 5 | 5 |
| 270-280 | 3 | 275 | 2 | 6 | 12 |
| 280-290 | 0 | 285 | 3 | 0 | 0 |
| N = 30 | | | | $\Sigma fd^1 = -36$ | $\Sigma fd^{1^2} = 192$ |

Calculation of Mean ;

$$\overline{X} = A + \frac{\Sigma fd^1}{N} \times C$$

$A = 255$, $\Sigma fd^1 = 36$, $N = 30$, $C = 10$

Substituting the values in the formula,

$$\overline{X} = 255 + \frac{-36}{30} \times 10$$

$$= 225 - 12 = 243$$

Calculation of standard Deviation :

$$\sigma = \sqrt{\frac{\Sigma f d^{1^2}}{N} - \left(\frac{\Sigma f d^1}{N}\right)^2} \times C$$

$\Sigma f d^{1^2} = 192$, $\Sigma f d^1 = -36$, N = 30, C = 10

Substituting the values in the formula,

$$\sigma = \sqrt{\frac{192}{30} - \left(\frac{-36}{30}\right)^2} \times 10$$

$$= \sqrt{6.4 - 1.44} \times 10$$

$$= \sqrt{4.96} \times 10$$

$$= 2.227 \times 10 = 22.27$$

Calculation of Mode :

**GROUPING TABLE**

| Class | Col. I | Col. II | Col. III | Col. IV | Col. V | Col. IV |
|---|---|---|---|---|---|---|
| 200-210 | 4 | | | | | |
| | | 6 | | | | |
| 210-220 | 2 | | | | | |
| | | | 5 | | | |
| 220-230 | 3 | | | 9 | | |
| | | | | | 8 | |
| | | 6 | | | | |
| 230-240 | 3 | | | | | 10 |
| | | | 7 | | | |
| 240-250 | 4 | | | 13 | | |
| | | 10 | | | | |
| 250-260 | 6 | | | | 15 | |
| | | | 11 | | | |
| 260-270 | 5 | | | | | 14 |
| | | 8 | | | | |
| 270-280 | 3 | | | 8 | | |
| | | | 3 | | | |
| 280-290 | 0 | | | | | |

ANALYSIS TABLE

| Col. No. | Wages in Rupees | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 200-210 | 210-220 | 220-230 | 230-240 | 240-250 | 250-260 | 260-270 | 270-280 |
| I | | | | | | 1 | | |
| II | | | | | 1 | 1 | | |
| III | | | | | | 1 | 1 | |
| IV | | | | 1 | 1 | 1 | | |
| V | | | | | 1 | 1 | 1 | |
| VI | | | | | | 1 | 1 | 1 |
| | | | | 1 | 3 | 6 | 3 | 1 |

It is clear from the analysis table that the mode lies in the class 250-260.

$$Mode = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

$$L = 250, \quad f_1 = 6, \quad f_0 = 4, \quad f_2 = 5, \quad i = 10$$

Substituting the values in the formula,

$$\text{Mode} = 250 + \frac{6-4}{2\times6-4-5} \times 10$$

$$= 250 + \frac{20}{3}$$

$$= 256.67$$

Calculation of coefficient of skewness :

$$SKp = \frac{Mean - Mode}{Standard\ Deviation}$$

Mean = 243, Mode = 256.67, Standard Deviation = 22.27.

Substituting the values in the formula,

$$SKp = \frac{243-256.67}{22.27}$$

$$= \frac{-13.67}{22.27}$$

$$= -0.61$$

Thus the distribution is negatively skewed

**Illustration - 5**

Calculate Karl Person's coefficient of skewness from the following data :

Marks below : 10 20 30 40 50 60 70 80

No. of students : 5 25 30 35 40 60 65 70

**Solution :**

Close observation of the distribution reveals that mode lies in 10-20 group as well as in 50-60. Hence the series is a bi-modal series and for calculating skewness the following formula is used.

$$SKp = \frac{3(mean - Median)}{Standard\ Deviation}$$

CALCULATION OF MEAN, MEDIAN AND STANDARD DEVIATION

| Marks | f | m m.v | (m − 35)/10 $d^1$ | $fd^1$ | $fd^{1^2}$ | c.f |
|---|---|---|---|---|---|---|
| 0-10 | 5 | 5 | −3 | −15 | 45 | 5 |
| 10-20 | 20 | 15 | −2 | −40 | 80 | 25 |
| 20-30 | 5 | 25 | −1 | − 5 | 5 | 30 |
| 30-40 | 5 | 35 | 0 | 0 | 0 | 35 |
| 40-50 | 5 | 45 | 1 | 5 | 5 | 40 |
| 50-60 | 20 | 55 | 2 | 40 | 80 | 60 |
| 60-70 | 5 | 65 | 3 | 15 | 45 | 65 |
| 70-80 | 5 | 75 | 4 | 20 | 80 | 70 |
| N = 70 | | | | $\Sigma fd^1 = 20$ | $\Sigma fd^{1^2} = 340$ | |

Calculation of Mean :

$$\overline{X} = A + \frac{\Sigma fd^1}{N} \times C$$

$A = 35,\ \Sigma fd^1 = 20,\ N = 70,\ C = 10$

Substituting the values in the formula,

$$\overline{X} = 35 + \frac{20}{70} \times 10$$

$$= 35 + 2.857 = 37.857$$

Calculation of Median :

$$\text{Median} = \text{Size of } \frac{N}{2}\text{th item}$$

$$= \frac{70}{2} = 35\text{th item}$$

Hence, Median lies in the class 30-40

$$\text{Median} = L + \frac{\frac{N}{2} - c.f.}{f} \times i$$

$L = 30,\ \frac{N}{2} = 35,\ c.f. = 30,\ f = 5,\ i = 10$

Substituting the values in the formula,

$$\text{Median} = 30 + \frac{35-30}{5} \times 10$$

$$= 30 + 10 = 40$$

**calculation of Standard Deviation :**

$$\sigma = \sqrt{\frac{\Sigma f d^{1^2}}{N} - \left(\frac{\Sigma f d^1}{N}\right)^2} \times C$$

$$\Sigma f d^{1^2} = 340, \quad \Sigma f d^1 = 20, \quad N = 70, \quad C = 10$$

Substituting the values in the formula,

$$\sigma = \sqrt{\frac{340}{70} - \left(\frac{20}{70}\right)^2} \times 10$$

$$= \sqrt{4.857 - 0.082} \times 10$$

$$= 21.85 \times 10$$

$$= 21.85$$

Calculation of skewness :

$$SKp = \frac{3(Mean - Median)}{Standard\ Deviation}$$

Mean = 37.857, Median = 40, S.D. = 21.85

Substituting the values in the formula,

$$SKp = \frac{3(37.857 - 40)}{21.85}$$

$$= \frac{3(-2.143)}{21.85}$$

$$= \frac{-6.429}{21.85}$$

$$= -0.294$$

Thus the distribution is negatively skewed.

**Illustration - 6**

Calculate the coefficient of skewness based on mean, median and standard deviation for the following distribution :

| Class : | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 | 70-80 | 80-90 |
|---|---|---|---|---|---|---|---|---|
| f : | 1 | 4 | 9 | 13 | 16 | 10 | 8 | 3 |

**Solution :**

**CALCULATION OF KARL PEARSON'S COEFFICIENT OF SKEWNESS**

| Class | f | m.v. m | $(m - 45)/10$ $d^1$ | $fd^1$ | $fd^{1^2}$ | c.f |
|---|---|---|---|---|---|---|
| 10-20 | 1 | 15 | −3 | −3 | 9 | 1 |
| 20-30 | 4 | 25 | −2 | −8 | 16 | 5 |
| 30-40 | 9 | 35 | −1 | −9 | 9 | 14 |
| 40-50 | 13 | 45 | 0 | 0 | 0 | 27 |
| 50-60 | 16 | 55 | 1 | 16 | 16 | 43 |
| 60-70 | 10 | 65 | 2 | 20 | 40 | 53 |
| 70-80 | 8 | 75 | 3 | 24 | 72 | 61 |
| 80-90 | 3 | 85 | 4 | 12 | 48 | 64 |
| | | | | $\Sigma fd^1 = 52$ | $\Sigma fd^{1^2} = 210$ | |

Calculation of Mean :

$$\overline{X} = A + \frac{\Sigma fd^1}{N} \times C$$

$A = 45$, $\Sigma fd^1 = 52$, $N = 64$, $C = 10$

Substituting the values in the formula,

$$\overline{X} = 45 + \frac{52}{64} \times 10$$
$$= 45 + 8.1$$
$$= 53.1$$

Calculation of standard Deviation :

$$\sigma = \sqrt{\frac{\Sigma fd^{1^2}}{N} - \left(\frac{\Sigma fd^1}{N}\right)^2} \times C$$

$\Sigma fd^{1^2} = 210$, $\Sigma fd^1 = 52$, $N = 64$, $C = 10$

Substituting the values in the formula,

$$\sigma = \sqrt{\frac{210}{64} - \left(\frac{52}{64}\right)^2} \times 10$$
$$= \sqrt{3.28 - 0.6600} \times 10$$
$$= \sqrt{2.6200} \times 10$$
$$= 1.62 \times 10$$
$$= 16.2$$

Calculation of Median :

$$\text{Median} = \text{Size of } \frac{N}{2}\text{th item}$$

$$= \frac{64}{2} = 32\text{nd item}$$

Median lies in the class 50-60

$$Median = L + \frac{\frac{N}{2} - c.f}{f} \times i$$

$$L = 50, \quad \frac{N}{2} = 32, \quad c.f = 27, \quad f = 16, \quad i = 10$$

Substituting the values in the formula,

$$\text{Median} = 50 + \frac{32-27}{16} \times 10$$

$$= 50 + \frac{5}{16} \times 10$$

$$= 50 + 3.1$$

$$= 53.1$$

Calculation of skewness :

$$SKp = \frac{3(Mean - Median)}{Standard\ Deviation}$$

Mean = 53.1, Median = 53.1, Standard deviation = 16.2

Substituting the values in the formula,

$$SKp = \frac{3(53.1-53.1)}{16.2}$$

$$= 0$$

Hence the distribution is symmetrical and not skewed.

**Check your progress - 1**

In a distribution Mean is 233.88, Median is 236.25 and standard deviation is 63.28. Find out the coefficient of skewnes.

**Illustration - 7**

The following facts relate to the workers of a factory before and after the settlement of an industrial dispute. Comment on the gains and losses from the point of view of workers and management.

| | | Before the settlement of dispute | After the settlement of dispute |
|---|---|---|---|
| No. of workers | : | 120 | 117 |
| Mean wages (Rs.) | : | 90 | 95 |
| Median wages (Rs.) | : | 96 | 90 |
| Standard Deviation(Rs.) | : | 6 | 5 |

**Solution:**

On the basis of the information given, the following comments can be made:

(1) **Comparison of total wage bill :**

Total wage bill before the settlement of the dispute $= 120 \times 90$

$= \text{Rs.}10,800$

Total wage bill after the settlement of the dispute $= 177 \times 95$

$= \text{Rs.}11,115$

It is clear that the total wage bill after the settlement of the industrial dispute has gone upto Rs.11,115 from Rs.10,800 before the settlement. But the number of workers has decreased from 120 to 117. Thus we can conclude that it is a gain to the workers and loss to the management unless the increased wages result in higher production and increased efficieny.

(ii) **Comparison of Median Wage :**

Before the settlement of the dispute, 50% of the workers used to get wages above Rs.96 and after the settlement, they get only above Rs.90.

(iii) **Comparison of variation :**

$$\left(\frac{\sigma}{\overline{X}} \times 100\right)$$

$$\sigma = 6, \quad \overline{X} = 90$$

Coefficient of variation before the settlement of the dispute

$$= \frac{6}{90} \times 100 = 6.67$$

Coefficient of variation after the settlement of the dispute

$$= \frac{5}{95} \times 100 = 5.26$$

Since the value of the coefficient of variation has decreased from 6.67 to 5.26 it can be concluded that wages are more uniformly distributed after the settlement of the dispute.

(iv) **Comparison of skewness :**

$$\text{SKp} = \frac{3(Mean - Median)}{Standard\ deviation}$$

SKp before the settlement of the dispute :

Mean = 90, Median = 96, S.D. = 6

Substituting the values in the formula,

$$= \frac{3(90-96)}{6} = \frac{-18}{6} = -3$$

SKp after the settlemtn of the dispute:

Mean = 95, Median = 90, Standard Deviation = 5

Substituting the values in the formula,

$$= \frac{3(95-90)}{5} = \frac{15}{5} = 3$$

While the distribution before the settlement of the dispute is negatively skewed, the distribution after the settlement of the dispute is positively skewed. This reveals that the number of workers getting low wages has increased considerably and that of workers getting high wages has fallen. This is despite the fact that the actual wage of workers has increased.

**(v) Position of averages :**

Mode = 3 Median − 2 Mean

Mode before settlement = $3 \times 96 - 2 \times 90$ = Rs.108

Mode after settlement = $3 \times 90 - 2 \times 95$ = Rs. 80

Thus, we conclude that after the dispute, concentration of individual wages is around a much smaller value than before. Most of the workers get near about Rs. 80 after the settlement as against Rs. 108 before the settlement of dispute.

**Illustration - 8.**

For a moderately skewed distribution the arithmetic mean is 50. The coefficient of variation is 2 and Karl Pearson's ceofficient of skewness is 0.3.

Find the mode and the median.

**Solution :** We are given that

$\overline{X} = 50$, C.V. = 2, SKp = 0.3

Calculation of standard Deviation:

$$\text{C.V} = \frac{\sigma}{\overline{X}} \times 100 = 2$$

$$= \frac{\sigma}{50} \times 100 = 2$$

$$= 2\sigma = 2$$

$$\sigma = \frac{2}{2} = 1$$

Thus standard deviation = 1

**Calculation of mode:**

$$\text{SKp} = \frac{Mean - Mode}{Standard\ Deviation}$$

$$= \frac{50 - Mode}{1} = 0.3$$

$$50 - \text{Mode} = 0.3$$

$$-\text{Mode} = -50 + 0.3$$

$$\text{Mode} = 49.7$$

**Calculation of Median:**

$$\text{Mode} = 3\,\text{Median} - 2\,\text{Mean}$$

$$49.7 = 3\,\text{Median} - 2 \times 50$$

$$3\,\text{Median} - 100 = 49.7$$

$$3\,\text{Median} = 149.7$$

$$\text{Median} = \frac{149.7}{3}$$

$$\text{Median} = 49.9$$

**Illustration - 9**

For a group of 100 items, $\Sigma X = 45{,}200$, $\Sigma X^2 = 2{,}42{,}70{,}000$, Mode = 437. Find the Pearsonian coefficient of skewness.

**Solution:**

Calculation of Mean :

$$\text{Mean} = \frac{\Sigma X}{N} = \frac{45200}{100} = 452$$

Calculation of Standard Deviation :

$$\Sigma X^2 = 2{,}42{,}70{,}000$$

$$\Sigma X = 45{,}200$$

$$N = 100$$

Substituting the value in the formula,

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2}$$

$$= \sqrt{\frac{2{,}42{,}7000}{100} - \left(\frac{45{,}200}{100}\right)^2}$$

$$= \sqrt{2,42,700 - 2,04,304}$$

$$= \sqrt{38,396}$$

$$\sigma = 195.95$$

Calculation of Karl Pearson's skewness:

$$SKp = \frac{Mean - Mode}{Standard\ Deviation}$$

Mean = 452, Mode = 437 and Standard Deviation = 195.95

$$SKp = \frac{452-437}{195.95}$$

$$= \frac{15}{195.95} = 0.0766$$

## 21.5 BOWLEY'S COEFFICIENT OF SKEWNESS

A.L. Bowley propounded an laternative method of the relative measure of skewness which is based on quartiles.

In an asymmetrical distribution the quartiles will not be equi-distant from the median. In a positively skewed distribution $Q_3$ will be far away from the median than $Q_1$. On the other hand in a negatively skewed distribution $Q_1$ will be far away from median than $Q_3$. In such a case skewness is measured with the help of the following formula:

Bowley's coefficient of skewness,

$$SK_B = \frac{(Q_3-Med)-(Med-Q_1)}{(Q_3-Med)+(Med-Q_1)}$$

or $$\frac{Q_3+Q_1-2Med}{Q_3-Q_1}$$

This measure is also called the quartile measure of skewness and the value of coefficient vaires between $\pm 1$. This method is useful specially in open end distributions and also in case of distributions which contain extreme values.

**Illustration : 10**

Calculate Bowley's coefficient of skewness for the following frequency distribution:

| Marks : | 0 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 |
|---|---|---|---|---|---|---|---|---|---|
| No. of Students: | 10 | 15 | 8 | 12 | 30 | 35 | 20 | 18 | 3 |

**Solution :**

CALCULATION OF BOWLEY'S COEFFICIENT OF SKEWNESS

| Marks X | f | c.f |
|---|---|---|
| 0 | 10 | 10 |
| 10 | 15 | 25 |
| 20 | 8 | 33 |
| 30 | 12 | 45 |
| 40 | 30 | 75 |
| 50 | 35 | 110 |
| 60 | 20 | 130 |
| 70 | 18 | 148 |
| 80 | 3 | 151 |

Calculation of Median :

Median = Size of $\frac{N+1}{2}$th item $= \frac{151+1}{2} =$ 76th item

Size of 76th item = 50. Hence Median = 50

Calculation of Quartiles:

$$Q_1 = \text{Size of } \frac{N+1}{4}\text{th item} = \frac{152}{4} = 38\text{th item}$$

Size of 38th item = 30, Hence $Q_1 = 30$.

$$Q_3 = \text{Size of } 3\left(\frac{N+1}{4}\right)\text{th item}$$

$$= 3 \times \frac{152}{4} = 114\text{th item}$$

Size of 114th item = 60, Hence $Q_3 = 60$.

Calculation of Bowley's coefficient of skewness :

$$SK_B = \frac{Q_3+Q_1-2Med}{Q_3-Q_1}$$

$Q_3 = 60$, $Q_1 = 30$, Median = 50

Substituing the values in the formula,

$$SK_B = \frac{60+30-2\times50}{60-30}$$

$$= \frac{90-100}{30}$$

$$= \frac{-10}{30}$$

$$= -0.33$$



**Illustration - 11**

Calculate coefficient of skweness based on quartiles and median from the following data

| Income (Rs.): | 100-200 | 200-300 | 300-400 | 400-500 |
|---|---|---|---|---|
| No. of persons : | 10 | 15 | 20 | 25 |

| Income (Rs.) : | 500-600 | 600-700 | 700-800 | 800-900 | 900-1000 |
|---|---|---|---|---|---|
| No. of persons : | 30 | 8 | 7 | 4 | 1 |

**Solution :**

**CALCULATION OF BOWLEY'S COEFFICIENT OF SKEWNESS**

| Income (Rs) | No. of persons | |
|---|---|---|
| | f | c.f |
| 100-200 | 10 | 10 |
| 200-300 | 15 | 25 |
| 300-400 | 20 | 45 |
| 400-500 | 25 | 70 |
| 500-600 | 30 | 100 |
| 600-700 | 8 | 108 |
| 700-800 | 7 | 115 |
| 800-900 | 4 | 119 |
| 900-1000 | 1 | 120 |

Calculation of Median:

$$\text{Median} = \text{Size of } \frac{N}{2}\text{th item} \quad = \frac{120}{2} = 60\text{th item.}$$

Median lies in the class 400-500

$$Median = L + \frac{\frac{N}{2} - cf}{f} \times i$$

L = 400, N = 60, c.f = 45, f = 25, i = 100

Substituting the values in the formula,

$$\text{Median} = 400 + \frac{60-45}{25} \times 100$$

$$= 400 + 15 \times 4$$

$$= 400 + 60 = 460$$

Calculation of Quartiles:

$Q_1$ = Size of $\frac{N}{4}$th item $= \frac{120}{4} = 30$th item

$Q_1$ lies in the class 300-400

$$Q_1 = L + \frac{\frac{N}{4} - c.f}{f} \times i$$

L = 300, $\frac{N}{4} = 30$, c.f = 25, f = 20, i = 100

Substituting the values in the formula,

$$Q_1 = 300 + \frac{30-25}{20} \times 100$$

$= 300 + 5 \times 5$

$= 325$

$Q_3 =$ Size of $\frac{3N}{4}$th item

$= \frac{3\times120}{4} = \frac{360}{4} = 90$th item

$Q_3$ lies in the class 500-600

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

L = 500, $\frac{3N}{4} = 90$, c.f = 70, f = 30, i = 100

Substituting the values in the formula,

$= 500 + \frac{90-70}{30} \times 100$

$= 500 + \frac{200}{3}$

$= 566.7$

Calculation of Bowley's Coefficient of Skewness:

$$SK_B = \frac{Q_3+Q_1-2Med}{Q_3-Q_1}$$

$Q_3 = 566.7$, $Q_1 = 325$, Median = 460

Substituting the values in the formula,

$SK_B = \frac{566.7+325-2\times460}{566.7-32.5}$

$= \frac{891.7-920}{241.7}$

$= \frac{-28.3}{241.7}$

$= -0.117$



**Illustration - 12**

In a frequency distribution the coefficient of skewness based on quartiles is 0.3. If the sum of the upper and the lower quartiles is 50 and the median is 19, find the value of upper quartile.

**Solution:**

We are given $SK_B = 0.3$

$Q_3 + Q_1 = 50$, Median = 19

$$SK_B = \frac{Q_3 + Q_1 - 2\ median}{Q_3 - Q_1}$$

Substituting the given values in the formula,

$0.3 = \frac{50-2\times19}{Q_3-Q_1}$

$$0.3 = \frac{50-38}{Q_3-Q_1}$$

$$0.3\,(Q_3 - Q_1) = 50 - 38$$

$$Q_3 - Q_1 = \frac{50-38}{0.3}$$

$$Q_3 - Q_1 = \frac{12}{0.3}$$

$$Q_3 - Q_1 = 40$$

$Q_3 + Q_1 = 50$ .. (i) as given in the problem.

$Q_3 - Q_1 = 40$ .. (ii) as calculated above.

If we solve the two equations

$$Q_3 + Q_1 = 50 \quad ...(i)$$

$$Q_3 - Q_1 = 40 \quad ...(ii)$$

$$2Q_3 = 90$$

$$Q_3 = \frac{90}{2} = 45$$

Hence the value of the upper quartile is 45.

## 21.6 SUMMING UP

Measures of skewness indicate the direction and degree of skewness in a series. Measures of skewness are of two types, viz., (i) absolute measures of skewness and (ii) relative measures of skewness. While absolute measures measure the skewness in absolute terms by simply taking the difference between the mean and mode, in relation to some measure of dispersion. The relative measures of skewness are useful for comparing two or more distributions as the coe-fficient obtained is a pure number which is independent of the measure of units in which the values of distribution are expressed.



## 21.7 CHECK YOUR PROGRESS:MODEL ANSWERS

1. The following formula may be applied :

$$SKp = \frac{3(\overline{X} - Md)}{\sigma}$$

The asnwer is — 0.112.

## 21.8 MODEL EXAMINATION QUESTIONS

A. Short Questions

1. What is meant by the term 'Skewness'?
2. What are the various methods of measuring Skewness?
3. How do relative measures of skewness differ from absolute measures of skewness?

4. Explain the merits of relative measures.

## EXERCISES

5. Calculate Pearson coeffficent of skewness from the following data:

| Wages (Rs.) | 100 | 200 | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 |
|---|---|---|---|---|---|---|---|---|---|---|
| No. of: workers | 50 | 20 | 25 | 10 | 30 | 30 | 40 | 15 | 15 | 2 |

(Ans : 0.75)

6. Compute Karl Pearson's coefficient of skewness from the following data.

| Class | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 | 70-80 | 80-90 |
|---|---|---|---|---|---|---|---|---|
| Frequency: | 2 | 12 | 30 | 10 | 57 | 43 | 15 | 6 |

(Ans : 0.82)

7. Calculate Karl Pearson's coefficient of skewness from the following data :

| Wages above : Rs. | 100 | 150 | 200 | 250 | 300 | 350 | 400 | 450 |
|---|---|---|---|---|---|---|---|---|
| No. of workers: | 175 | 150 | 72 | 57 | 42 | 23 | 15 | 0 |

(Ans: 0.5)

8. Find out Pearsonian coefficient of skewness from the following data.

| Age : | 20- 29 | 30-39 | 40-49 | 50-59 | 60-69 | 70-79 |
|---|---|---|---|---|---|---|
| No. of persons: | 25 | 35 | 40 | 90 | 75 | 60 |

(Ans : 0.41)

9. Find Bowley's coefficient of skewness from the following data:

| X: | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 |
|---|---|---|---|---|---|---|---|---|---|
| f: | 8 | 12 | 18 | 10 | 7 | 4 | 11 | 10 | 6 |

(Ans : 0.5)

10. Calculate the measure of skewness based on Quartiles and median from the following data:

| Variable: | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 | 70-80 | 80-90 | 90-100 |
|---|---|---|---|---|---|---|---|---|---|---|
| frequency: | 15 | 12 | 27 | 46 | 33 | 11 | 9 | 7 | 6 | 4 |

(Ans : 0.58)

11. For moderately skewed data the arithmetic mean is 300, the coefficient of variation is 12 and Pearson coefficient of skewness is 0.4. Find the mode and median.

(Ans : 9)

12. For a group 40 items $\Sigma X$ = 2840, $\Sigma X$= 284180 and mode = 524. Find the Personian coefficient of skewness.

(Ans : 0.08)

13. From the following data, calculate Karl Person's coefficient of Skewness and also Bowley's coefficient of skewness.

| | Place 'M' | Place 'N' |
|---|---|---|
| Mean | 900 | 840 |
| Median | 852 | 930 |
| Standard Deviation | 180 | 330 |
| First Quartile | 372 | 480 |
| Third Quartile | 1170 | 1560 |

(Ans : 0.8 and 0.203; and 0.8 and 0.16)

14. Calculate Bowley's coefficient of skewness from the following data;

| Amount spent on Advertisement: | 5-9 | 10-14 | 15-19 | 20-24 | 25-29 | 30-34 | 35-39 | 40-44 | 45-49 |
|---|---|---|---|---|---|---|---|---|---|
| No. of Companies : | 12 | 17 | 20 | 45 | 37 | 30 | 17 | 13 | 14 |

(Ans : 0.13)

15. In a certain distribution the following results were obtained : $\overline{X}$ = 180, Med = 192, coefficient of skewness = − 0.4. On the basis of the available data find out the value of Standard Deviation.

(Ans : 90)

16. In a frequency distribution Bowley's coefficient of skewness is 0.3. If the sum of the Upper and Lower Quartiles is 150 and the median is 19, find the value of the Upper Quartile.

(Ans : 261)

17. Pearsonian coefficient of skewness = 0.4 Mean = 42 and mode = 25. Find the Standard Deviation of the distribution.

(Ans : 127.5)

## 21.9 RECOMMENDED BOOKS


1. Gupta, S.P. : "Statistical Methods", Sultan Chand & Company, New Delhi.
2. Gupta, B.N : "Statistics", Sahitya Bhavan, Agra.
3. Gupta, S.C. : "Fundamentals of statistics" Himalaya Pub. House, Bombay.
4. Simpson and Kafka : "Basic statistics", Oxford and I.B.H. publishing Company, Calcutta.


## 21.10 GLOSSARY

| | | |
|---|---|---|
| Absolute Measure of skewness | : | It expresses the extent of skewness in the form in which the data have been expressed. Simply it is the difference between mean and mode or it can also be computed with the help of quartiles and median. |
| Relative Measures of skewness | : | It gives a pure number, which is independent of the units in which the data are expressed. It facilitates comparison between two or more distributions. |

## BLOCK- IV : CORRELATION AND REGRESSION ANALYSIS

## UNIT-22 : CORRELATION

**Contents**



## 22.0 AIMS AND OBJECTIVES

The aim of this unit is to explain the nature, scope and importance of correlation analysis. After going through this unit, you should be able to :

- define correlation
- explain correlation and causation
- describe the importance of correlation
- identify the various types of correlation
- recognise the degree of correlation between the variables.



## 22.1 INTRODUCTION

In the earlier units, we have discussed the descriptive statistics which described the quantitative characteristics of data. The scope of those units was strictly confined to the various values of one variable only. For example, measures of central tendency, measures of dispersion and skewness deal with the various values of a single variable. These statistical measures are useful for comparison and analysis, but they are not useful in finding out the quantitative relationship between the variables. In our practical life, we come across different sets of data that deal with more than one variable which are interrelated and interdependent. For example, there is close relationship between the price of a commodity and the quantity demanded, capital invested and

profits earned, advertisement expenditure and sales generated, yield of paddy, amount of rainfall and fertilizers used, etc. Further, the changes in the quantities of one variable are accompanied by the quantities of other variables. This is referred to as co-variation. Quantification of such interrelations and co-variation between different sets of data necessitates the use of analytical statistics. Correlation analysis is such an analytical statistical measure that helps in finding out the direction and extent of relationship between two or more variables. The measure of correlation known as correlation co- efficient or correlation index summarises the direction and degree of correlation in one figure. Thus, correlation analysis refers to the statistical technique that measures the closeness of the relationship between the variables. The process of correlation analysis helps us to find out whether the variables under study are related, and if related, the degree or extent of relationship and the direction of co-variation between variables.

## 22.2 DEFINITION OF CORRELATION

Some of the important definitions of correlation are given below:

(i) Simpson and Kafka defined correlation as "an analysis that deal with the association between two or more variables.

(ii) According to Ya-Lun-Chou "Correlation as "an analysis attempts to determine the degree of relationship between variables."

(iii) A.M.Tuttle defined correlation as "an analysis of the co-variation between two or more variables.

(iv) In the words of 'Croxton and Cowden', "When the relationship is of quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in brief formula is known as correlation.

(v) According to W.I.King, "Correlation means that between two series or groups of data there exists some causal connection". He further states, "if it is proved true that in a large number of instances two variables tend always to fluctuate in the same or in opposite directions, we consider that the fact is established and that a relationship exists. The relationship is called correlation."

An analysis of the above definitions reveals the following:

(i) Correlation analysis involves the study of two or more variables which have some causal relationship.

(ii) Correlation is an analysis of co-variation between the variables.

(iii) It is a statistical technique which measures the degree and direction of co-variation between the variables.

(iv) It is concerned with the measurement of the quantitative relationship between the variables.

**Check your progress -1**

Describe correlation analysis.

## 22.3 CORRELATION AND CAUSATION

Change in the values of one variable may be often instrumental for the change in the values of other variables. But correlation does not mean cause and effect relationship between any variables. To be more clear, it does not reveal which variable is the cause and which variable is the effect. Further, correlation analysis does not specify the causal relationship between the variables. High degree of correlation coefficient does not always mean that there is close relationship between the variables studied. For example the high degree of correlation coefficient between pigs and the production of pig-iron does not mean that there is close relationship between them. Thus, the presence of corelation betwen two variables does not necessarily imply the existence of direct causation, though causation will always result in correlation. The significant degree of correlation coefficient may be due to any one or the combination of the following factors:

*(i) One variable being the cause of the other*

The cause and effect relationship between the variables is often determined on the basis of the circumstances of the case. For example, rainfall causes the growth of agricultural production but agricultural production does not cause the rainfall. Since rainfall is independent of the agricultural production, it is called an independent variable and agricultural production which is subject to the influence of rainfall is called a dependent variable.

*(ii) Both the variables may act upon each other*

We may come across two or more variables which may effect each other. For example, in the case of high degree of correlation between price and demand of a commodity, it is possible that both the variables-price and demand, may affect each other. In such cases, it is not possible to identify exactly which variable is the cause and which variable is the effect.

*(iii) Both the variables being the result of a common cause*

Often, both the variables under study may be closely related to each other because of the fact that they are subject to the strong influence of certain other external factors. For example, the high degree of correlation between agricultural production and quantity of fertilizers used may be due to outside factors such as rainfall, irrigation facilities, quality of seeds used, etc. These external factors act upon both the variables and cause them to respond together.

*(iv) Chance coincidence*

Sometimes the mathematical calculations may give us the high degree of correlation between

the variables which are not at all related to each other in any way. This high degree of correlation may be due to sheer coincidence. While interpreting the correlation coefficient, it is essential to see whether the variables under study are related to one another. If no relationship exists between them, the statistical measure of correlation coefficient calculated is meaningless. This type of correlation is called 'spurious correlation.' High degree of corelation between coal production and agricultural production is an example of such a 'spurious correlation.'

*(v) High degree of correlation in a small sample*

Due to sampling fluctuations or due to the bias of the statistical investigator high degree of correlation may be obtained for small samples. There may not be any significant relationship between the results obtained for the small samples and that of the variables of universe. For example, the high degree of negative correlation between the heights of fathers and sons might be on account of the study of a small sample of pairs of them.

## 22.4. IMPORTANCE OF CORRELATION

The study of correlatiopn is extremely useful to government, business and consumers for knowing the behaviour of various interrelated and interdependent variables. Some of the specific uses of the study of correlation are explained below.

(i) Correlation analysis helps us to understand the relationship between different economic variables such as demand, price and supply of commodities, savings, investment and profits earned, per capita income and consumption pattern, exports and imports, etc. This type of analytical study helps the government and manufacturers to plan the production in accordance with the circumstances prevailing in the economy.

(ii) Correlation analysis helps the business management in finding out the direction and extent of relationship between the variables which affect the smooth functioning of the organisations. For example, with the help of correlation technique, the business executive can find out the degree of relationship between advertisement expenditure and amout of sales generated. They can also find out the functional relationship between costs, sales prices, etc. This sort of functional relationship reduces the impact of uncertainity in decision making and replaces the guess work with scientific methods. The predictions based on the statistical measure of correlation would be more reliable than the estimate based on intution.

(iii) The accuracy of relationship between the variables derived on the basis of unscientific or crude methods can be verified and tested for significance with the help of correlation technique.

(iv) Correlation coefficient is a pure number with the help of which the relationship between the variables, whose values are expressed in different units, can be studied and analysed without any difficulty and confusion.

(v) Correlation analysis helps us not only to understand the economic behaviour but also

to locate the critically important variables on which other variables depend. Thus, the disturbing forces can be identified and made neutral with the help of certain other stabilising forces.

(vi) Statistical techniques such as regression and ratio of variation depend heavily on the measure of correlation.

## 22.5. TYPES OF CORRELATION

On the basis of the nature relationship between the variables, correlation may be classified into following three categories:

(i) Positive and negative correlation

(ii) Simple, partial and multiple correlation

(iii) Linear and non-linear correlation

*(i) Positive and negative correlation*

Positive and negative correlation between two variables is identified with reference to the degree and direction of change that takes place in the variables. If the values of both the variables change in the same direction, correlation is said to be positive i.e, if the increase or decrease in the values of one variable is accompanied by the increase or decrease in the values of another variable, correlation is said to be positive.

Note the following examples:

(a) Price of commodity : Rs. 20,24,30,36,40

Supply of commodity : Kg. 30,40,44,50,74

(b) Price of a commodity : Rs. 40,35,30,20,15

Supply of a commodity: Kg. 25,23,15,10 ,5

On the other hand, if the decrease in the values of one variable is accompanied by the increase in the values of other variable or the increase in the values of one variable is followed by decrease in the values of other variable, the correlation is said to be negative or inverse.

Note the following examples:

(a) Price of commodity : Rs .10, 15, 20, 30, 40

Demand for a commodity : Kg. 20, 15, 11 , 8, 5

(b) Price of a commodity : Rs. 50, 45, 30, 20, 15

Quantity demanded : Kg. 5, 10, 15, 20, 25

*(ii) Simple, partial and multiple correlation*

Simple, partial and multiple correlation is identified with reference to the number of variables sutdied. If the relationship is calculated between two variables only, it is called 'simple correlation' i.e., correlation between demand and price of a commodity. On the other hand, the relationship

calculated for more than two variables is referred to either multiple or partial correlation. But in case of multiple correlation, the relationsip among all the variables (more than two) is studied simultaneously, i.e., study of relationship between agricultural production, rainfall, fertilizers used etc. On the other hand, in case of partial correlation, the relationship between agricultural production and rainfall is studied and the impact of fertilizers is ignored or kept constant.

*(iii) Linear and non-linear or curvi-linear correlation*

Linear and non-linear correlation is identified with reference to the amount or rate of change in the values of variables. If the amount or rate of change in one variable is accompanied by the same amount or rate of change in the other variable, it is known as linear correlation.

Note the following example:

X : 10, 20, 30, 40, 50 , 60

Y : 25, 50, 75, 100, 125, 150

On the other hand, if the amount or rate of change in one variable is not accompanied by same amount or rate of change in the other variable, correlation is said to be non-linear or curvi-linear.

Note the following example :

X : 20, 24, 44 , 68, 90 , 112

Y : 100, 110, 120, 180, 196, 240

In the case of a linear correlation, data if plotted on a graph paper, gives a straight line. On the other hand, data in respect of non-linear correlation it gives the shape of a curve.

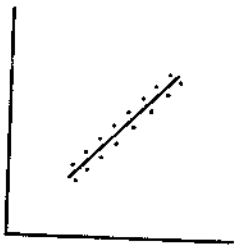


Fig. 22.1 Positive linear relationship

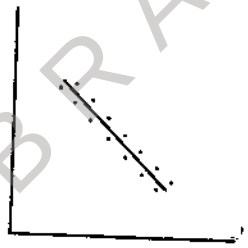


Fig. 22.2 Negative linear relationship

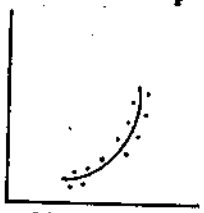


Fig. 22.3 Positive curvilinear relationship

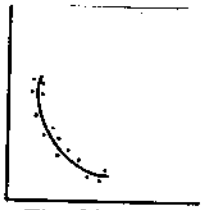


Fig. 22.4 Negative curvilinear relationship

In our practical life, we are concerned with the variables that have non- linear relationship. Since the statistical measures to study the non-linear relations are far more complicated, we

generally assume the presence of linear relationship between the variables.

**Check your progress-2**

Explain linear correlation and give a numerical example

## 22.6. DEGREE OF CORELATION

The extent of relationship between the variables is calculated with the help of a statistical technique known as correlation coefficient. According to the formula given by 'Karl Pearson', correlation coefficient always varies between ±1. While the algebraic sign (+)indicates the positive relationship between the variables, the sign (−) denotes the negative relationship. If no relationship exists between the variables under study, correlation coefficient will be zero. While +1 denotes the perfect positive correlation, −1 denotes the perfect negative correlation. If the correlation coefficient is closer to +1, we may say, there is higher degree of positive correlation. On the other hand, if, correlation coefficient is near to −1, we may say, there is higher degree of negative correlation. Correlation is said to be perfectly positive, if an increase/decrease in one variable is accompanied by the same amount or rate of increase/decrease in the other variable. On the other hand, correlation is said to be perfectly negative, if an increase/decrease in one variable is accompanied by the same amount or rate of change in the other variable in the reverse direction. In the case of positive or negative correlation, the change in one variable need not be accompanied by the change in the other variable in a definite ratio.

The following chart illustrates the approximate degrees of correlation between the variables:

| Degree of correlation | Positive | Negative |
|---|---|---|
| Absence of correlation | 0 | 0 |
| Very low degree of correlation (Insignificant) | 0 to + 0.25 | 0 to −0.25 |
| Low degree of correlation | + 40.25 to + 0.5 | − 0.25 to −0.5 |
| Moderate degree of correlation | + 0.5 to + 0.75 | − 0.5 to − 0.75 |
| High degree of correlation | + 0.75 to + 1.00 | − 0.75 to − 1.00 |
| Perfect correlation | + 1 | −1 |

## 22.7 SUMMING UP

Correlation analysis is an analytical stastical measure which helps to find out the direction and degree of relationship between two or more related variables. However, correlation analysis does not specify the causal relationship between the variables. Thus, the significant degree of correlation coefficient may be due to several other factors that are not included in the study. On the basis of the nature of relationship, correlation analysis may be classified into three categories, viz., (i) positive and negative correlation (ii) simple, partial and multiple correlation, and (iii) linear and non-linear correlation. The study of correlation is extremely useful to government, business and consumers for knowing the behaviour of various inter-related and inter-dependent variables.

## 22.8 CHECK YOUR PROGRESS:MODEL ANSWERS

1. Correlation analysis attempts to determine the degree and direction of relationship between the variables.
2. If the rate or amount of change in one variable is accompanied by the same amount or rate of change in the other variable, it is known as linear correlation. Eg.

| | | | | | |
|---|---|---|---|---|---|
| X : | 5 | 10 | 15 | 20 | 25 |
| Y : | 10 | 20 | 30 | 40 | 50 |

## 22.9 MODEL EXAMINATION QUESTIONS

**A.Short Questions**

1. What is meant by Correlation?
2. What is meant by 'dependent variable'?
3. What is meant by 'independent variable'?
4. What is meant by 'Causation'?
5. Distinguish between positive and negative correlation.
6. Distinguish between linear and non-linear correlation.
7. Distinguish between Simple, Partial and Multiple correlation.
8. Explain the term 'Spurious correlation'.
9. Explain the meaning and significance of the concept of correlation.

**B.Essay Questions**

10. Explain the various types of correlation with suitable examples.
11. "Even a high degree of correlation does not mean that a realtionship of cause and effect exists between the two correlated variables." Explain the statement with suitable examples.

## 22.10 RECOMDED BOOKS


1. Gupta, S.P. : "Statistical Methods" Sultan Chand & Company, New Delhi.
2. Gupta, B.N. : "Statistics", Sahitya Bhavan,Agra.
3. Gupta, S.C : "Fundamentals of statistics", Himalaya Publishing House, Bombay.
4. Simpson and Kafka : "Basic statistics", Oxford and IBH Publishing Company, Calcutta.


## 22.11 GLOSSARY

1. Correlation : It refers to the association or interdependence between two or more realted variables.
2. Covariation : It refers to interdependence or interrelationship between the variables.
3. Dependent variable : The variable whose value is influenced or is to be predicted is called dependent variable.
4. Independent variable : The variable which influences the values or is used for prediction is called independent variable.
5. Linear correlation : If the amount of change or rate of change is equal between the values of variables,it is called linear correlation.
6. Multiple correlation : It is the calculation of association between more than two variables
7. Negative correlation : If the values of two variables change in the opposite direction (i.e., if one increases the other decreases), the correlation between them is said to be negative.
8. Non-linear correlation : If the amount of change or rate of change is unequal between the values of variables, it is called non-linear correlation.
9. Partial correlation : It studies the relationship between just two variables and the remaining variables are ignored.

10. Positive correlation : If the values of two variables change in the same direction, (i.e, either increasing or decreasing) the correlation between them is said to be positive.

11. Simple Correlation : It is the calculation of association between two variables.

12. Spurious Correlation : When correlation is established between two unrelated variables, it is called spurious correlation.

# UNIT-23 : METHODS OF STUDYING CORRELATION-I

**Contents**



## 23.0 AIMS AND OBJECTIVES

This unit aims at presenting the methods of studying correlation.

After going through this unit, you should be able to :

- explain the method of studying correlation by Scatter Diagram Method
- Identify the correlation by graphic method
- compute correlation by applying the formula of Karl pearson.



## 23.1 INTRODUCTION

In the previous unit we explained the meaning and types of correlation. In this unit we discuss the methods of studying correlation. They are as follows:

1. Scatter Diagram Method
2. Graphic method
3. Karlpearson's Coefficient of Correlation
4. Spearman's Rank coefficient of Correlation
5. Coefficient of concurrent Deviation

While the first two are basically graphic methods, other are mathematical methods. In this unit, the first three methods are discussed in detail. The other two are discussed in the subsequent units.

Both scatter diagram method and graphic method would reveal the nature of correlation i.e., whether it is positive or negative. In other words, they provide just the direction of variables

in which they are moving. They do not give the extent of correlation between the variables.

To overcome this drawback mathematical methods have come into existence. They express both the nature and the extent of correlation between the variables. Karl Pearson's coefficient of correlation is one of such measures avialable to us.

## 23.2 SCATTER DIAGRAM METHOD

A scatter diagram is also called dotogram or scattergram. According to this method, the values of each pair of observations are shown on the graph paper with the help of a single dot. Thus we get as many dots as the number of pairs of observations. With the help of the dots plotted on the graph, we get a visual idea of the relationship between the variables. While the closeness of the dots on the diagram indicates the high degree of correlation. Correlation is said to be positive if the points on the diagram rise from the lower left hand corner to the upper right hand corner. On the other hand, correlation is said to be negative if the points show a decreasing tendency from the upper right hand corner to the lower left hand corner. If all the points fall on a straight line rising from the left hand lower corner to the right hand upper corner, correlation is said to be perfectly positive (i.e., $r = +1$). On the other hand, if all the points fall on a straight line decreasing from the right hand upper corner to the left hand lower corner, correlation is said to be perfectly negative (i.e., $r = -1$). Correlation is said to be absent if the plotted dots lie on a straight line parallel to the X-axis or in a haphazard manner (i.e., $r = 0$). The following diagrams would illustrate the difference between the varying degrees of correlation :



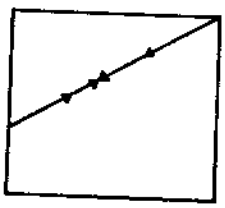


Fig 23.1 Perfect Positive Correlation($r = +1$)

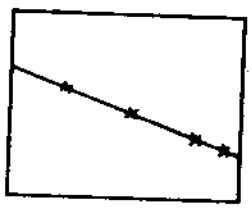


Fig. 23.2 Perfect Negative Correlation ($r = -1$)

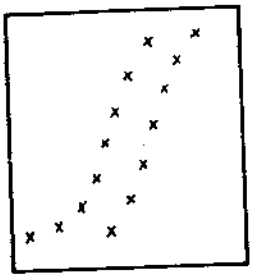


Fig. 23.3 High degree of Positive Correlation

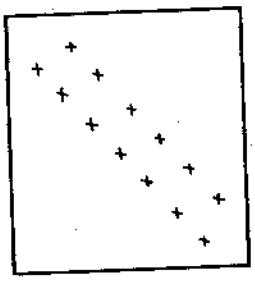


Fig. 23.4 High degree of Negative Correlation

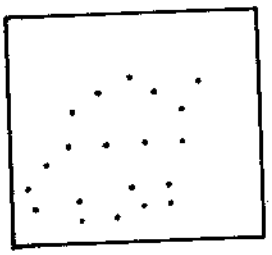


Fig. 23.5 Low degree of Positive Correlation

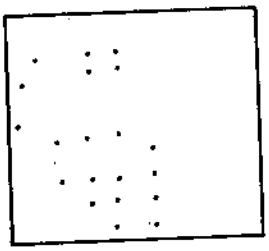


Fig. 23.6 Low degree of Negative Correlation

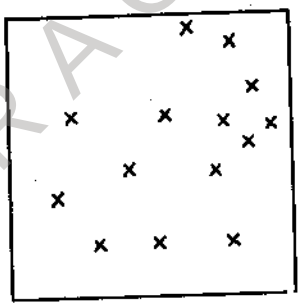


Fig. 23.7 No correlation ($r = 0$)



**Merits and Limitations of Scatter Diagram Method**

The following are the merits of scatter diagaram method:

(i) It is a simple and visual method which does not require any complicated mathematical calculations.

(ii) It tells us at a glance whether there is any relationship between the variables.

(iii) It indicates the direction and extent of correlation between the variables.

(iv) It tells us whether the correlation is linear or non-linear. It also helps to estimate the

values of dependent variables on the basis of the values of independent variables.

(v) It helps in obtaining the line of best fit.

(vi) It helps to locate the abnormal points, in the data. If any point or points fall far away from other points of the data, it indicates abnormal tendency in the values of variables. Such an abnormal tendency necessitates further investigation and detailed analysis.

(vii) Unlike the mathematical methods, it is not influenced by the extreme values of the data.

However, this method suffers from certain limitations. Scatter diagram gives only a rough idea of the relationship between two variables. As such, definite conclusions cannot be drawn on the basis of a scatter diagram. It does not measure the degree of correlation in numerical terms. Further, it is useful to study correlation between only two variables.

## 23.3 GRAPHIC METHOD

According to this method, the values of variables are plotted on a graph paper and separate curves are obtained for each variable. By examining the closeness and direction of curves we can determine the direction and extent of relationship between two or more variables. Correlation betwen two variables said to be positive if both the curves move in the same direction (either upward or downward). On the other hand, if the curves move in the opposite direction, correlation is said to be negative. Erratic fluctuations in the curves indicate either the absence of correlation or a low degree of correlation between the variables.

Like a scatter diagram, graphic method is also simple and visual method which does not require any mathematical calculations. The greatest advantage of this method is that it can be used for determining the relationship between two or more variables. Though this method helps to get a rough idea about the relationship among the variables, it does not measure the degree of correlation in exact numerical terms. Thus, with the help of this method, we cannot draw definite conclusions about the relationship among the variables. The following illustration would further explain the nature and importance of graphic method.



**Illustration :** From the following data, determine the correlation by graphic method.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Year | : | 1978 | 1979 | 1980 | 1981 | 1982 | 1983 | 1984 |
| Average Income(Rs.) | : | 100 | 150 | 200 | 250 | 275 | 300 | 350 |
| Average Expenditure(Rs.) | : | 75 | 100 | 125 | 125 | 150 | 175 | 225 |

**Solution :**

To determine correlation between average income and average expenditure, Time Period i.e., years are shown on "X" axis and the figures relating to average income and expenditure are shown on "Y" axis.

Scale:

On X-axis, 1 cm. = one year

On Y-axis, 1 cm. = Rs. 50

Now two seperate curves are drawn with the help of the given values one for the average income and the other for average expenditure.

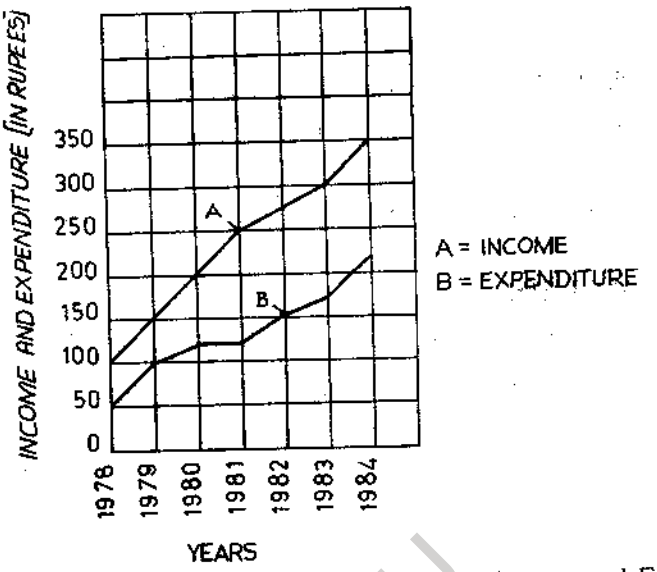



Fig. 20.8 Correlation showing the Average Income and Expenditure

The graph shows that the variable, income and expenditure are closely related.

## 23.4 KARL PEARSON'S COEFFICIENT OF CORRELATION

Karl pearson's coefficient of Correlation, (Pearsonian Coefficient of Correlation) is a popular mathematical method of measuring correlation between the variables. Pearsonian Coefficient of Correlation is denoted by the letter "r" and is computed either by direct method or by short-cut method



***Direct Method***

According to this method,coefficient of correlation is a ratio of the covariance to the standard deviations in the two series.

It can be explained symbolically as,

$$r = \frac{\Sigma xy}{N.\sigma x.\sigma y} \quad \text{.....(i)}$$

Where,

x = Deviations of "X" series taken from its arithmetic mean (i.e., $X - \overline{X}$ )

y = Deviations of "Y" series taken from its arithmetic mean (i.e., $Y - \overline{Y}$ )

$\Sigma xy$ = Sum of the products of deviations of X and Y series taken from their respective arithmetic means

$\sigma x$ = Standard deviation of "X"

$\sigma y$ = Standard deviation of "Y"

N = Number of pairs of observations

This formula is applied only where the deviations of items are taken from actual means of respective series.

**Illustration - 1**

Calculate the coefficient of correlation between X and Y series from the following data:

| | X Series | Y Series |
|---|---|---|
| No. of pairs observed | 100 | 100 |
| Standard deviation | 9 | 4 |
| Sum of products of deviations of X and Y series from their respective arithmetic means | 1340 | |

**Solution : Calculation of Coefficient of Correlattion**

$$r = \frac{\Sigma xy}{N.\sigma x.\sigma y}$$

We are given that N = 100, $\sigma x = 9$, $\sigma y = 4$, $\Sigma xy = 1340$.

Substituting the values in the formula,

$$r = \frac{1340}{100 \times 9 \times 4}$$

$$r = \frac{1340}{3600}$$

$$r = 0.37$$

Thus, the coefficient of correlation between X and Y = 0.37.

**Illustration 2**

If covariance between "X" and "Y" series is 16.2 and the variance of X and Y are respectively 25.2 and 15.5, find the coefficient of correlation between X and Y.

**Solution:**

Covariance between two variables is the sum of products of the deviations of two series taken from their respective arithmetic means divided by the number of pairs observed. Thus, covariance between X and Y, $= \frac{\Sigma xy}{N} = 16.2$.

Calculation of standard deviation of X series ($\sigma x$):

$$\sigma x^2 = \text{variance of X series}$$

We are given that,

$$\sigma y^2 = 25.2$$

$$\sigma x = \sqrt{25.2}$$

$$\sigma x = 5.02$$

Calculation of standard deviation of Y series ($\sigma$ y):

$$\sigma y^2 = \text{variance of Y series}$$

We are given that,

$$\sigma y^2 = 15.5$$

$$\sigma y = \sqrt{15.5}$$

$$\sigma y = 3.94$$

Calculation of coefficient of correlation:

$$r = \frac{\Sigma xy}{N.\sigma x.\sigma y.}$$

Here $\frac{\Sigma XY}{N} = 16.2$, $\sigma x = 5.02$ and $\sigma y = 3.94$.

Substituting the values in the formula,

$$r = \frac{16.2}{5.02 \times 3.94}$$

$$r = \frac{16.2}{19.78}$$

$$r = 0.82$$

**Thus, coefficient of correlation betwen X and Y = 0.82.**

**Illustration 3 :**

**Karl Pearson's coefficient of correlation between X and Y series is 0.82, their covariance is +16.2. If the variance of X is 25.2, find the Standard Deviation of Y series.**

**Solution :**

Calculation of Standard Deviatilon of X series:

$$\sigma x^2 = \text{variance of X series}$$

$$\sigma x = \sqrt{\textit{varianace of X series}}$$

We are given that variance of X sereis = 25.2

$$\sigma x^2 = 25.2$$

$$\sigma x = \sqrt{25.2}$$

$$\sigma x = 5.02$$

Further, we are given that,

Covariance between X and Y = 16.2

Coveriance between X and Y = $\frac{\Sigma xy}{N}$

Hence $\frac{\Sigma xy}{N} = 16.2$

Calculation of Standard Deviation of Y series:

$$r = \frac{\Sigma xy}{N.\sigma x.\sigma y}$$

$r = 0.82$ (as given in the problem)

$\frac{\Sigma xy}{N} = 16.2$ and $\sigma$ x = 5.02

Substituting the values in the formula,

$$0.82 = \frac{16.2}{5.02 \times \sigma y}$$

$$0.82 \times 5.02 \times \sigma \text{ y} = 16.2$$

$$4.12 \times \sigma \text{ y} = 16.2$$

$$\sigma \text{ y} = \frac{16.2}{4.12}$$

$$\sigma \text{ y} = 3.93.$$

Standard Deviation of Y series = 3.93.

Formula (i) can be simplified in the following way, so that we can avoid the calculation of standard deviations of X and Y series.

$$r = \frac{\Sigma xy}{N.\sigma x.\sigma y} \quad \text{.....(i)}$$

Where $\sigma$ x $= \sqrt{\frac{\Sigma x^2}{N}}$

and $\sigma$ y $= \sqrt{\frac{\Sigma y^2}{N}}$

If these values of $\sigma$x and $\sigma$ y are substituted in formula (i), we get,

$$r = \frac{\Sigma xy}{N\sqrt{\frac{\Sigma x^2}{N} \times \frac{\Sigma y^2}{N}}}$$

$$= \frac{\Sigma xy}{N\sqrt{\frac{\Sigma x^2 \times \Sigma y^2}{N^2}}}$$

$$= \frac{\Sigma xy}{\frac{N}{N}\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

$$= \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} \quad \text{....(ii)}$$

The following steps are required to compute the Pearsonian Coefficient of Correlation:

(i) Calculate the arithmetic mean of X series $\overline{X}$.

(ii) Calculate the arithmetic mean of Y series $\overline{Y}$.

(iii) Take the deviations of 'X' series from its arithmetic mean $\overline{X}$ and denote these deviations by 'X'.

(iv) Square the deviations of 'X' series and find the total $\Sigma X^2$.

(v) Take the deviations of 'Y' series from its arithmetic mean $\overline{Y}$ and denote these deviations by 'Y'.

(vi) Square the deviations of 'Y' series and find out the total $\Sigma Y^2$.

(vii) Multiply the deviations of X and Y series and find the total $\Sigma XY$.

(viii) Substitute the values of $\Sigma XY$, $\Sigma X^2$ and $\Sigma Y^2$ in the formula and obtain the coefficient of correlation.

**Illustration - 4**

Calculate Karl Pearson's coefficient of correlation from the following data:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Price(Rs.) : | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 14 | 15 |
| Demand(Kgs.) : | 20 | 18 | 18 | 15 | 12 | 12 | 12 | 10 | 8 | 5 |

**Solution :**

**CALCULATION OF KARL PEARSON'S COEFFICIENT OF CORRELATION**

Let price be denoted by 'X' and demand by 'Y'.

| X | $X - \overline{X}$ | $x^2$ | Y | $Y - \overline{Y}$ | $y^2$ | xy |
|---|---|---|---|---|---|---|
| | x | | | y | | |
| 4 | −5 | 25 | 20 | 7 | 49 | −35 |
| 5 | −4 | 16 | 18 | 5 | 25 | −20 |
| 6 | −3 | 9 | 18 | 5 | 25 | −15 |
| 7 | −2 | 4 | 15 | 2 | 4 | −4 |
| 8 | −1 | 1 | 12 | −1 | 1 | 1 |
| 9 | 0 | 0 | 12 | −1 | 1 | 0 |
| 10 | 1 | 1 | 12 | −1 | 1 | −1 |
| 12 | 3 | 9 | 10 | −3 | 9 | −9 |
| 14 | 5 | 25 | 8 | −5 | 25 | −25 |
| 15 | 6 | 36 | 5 | −8 | 64 | −48 |
| $\Sigma X = 90$ | $\Sigma x = 0$ | $\Sigma x^2 = 126$ | $\Sigma Y = 130$ | $\Sigma y = 0$ | $\Sigma y^2 = 204$ | $\Sigma xy = -156$ |

Calculation of Arithmetic Mean of X series:

$$\overline{X} = \frac{\Sigma X}{N}$$

$$= \frac{90}{10}$$

distribution is said to be negatively skewed.

To find out the absolute measure of skewness, the difference between mean and mode is used because of the fact that in a symmetrical distribution the values of mean, median and mode are alike, but in an asymmetrical distribution mean moves away from mode on either side. Hence, the distance between mean and mode is used to measure skewness. The greater the distance between mean and mode whether it be positive or negative, the more would be the asymmetry in the distribution.

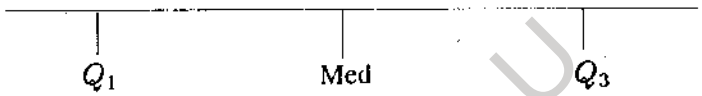
When skewness is calcutated on the basis of quartiles, the following formula is used :

Absolute skewness $= Q_3 + Q_1 - 2$ Med.

This formula is based on the relationship of positional averages of a moderately skewed distribution. In a moderately skewed distribution,

$(Q_3 - Med) = (Med - Q_1)$

Hence, $Q_3 + Q_1 - 2$ Med $= 0$




Absolute measure of skewness is considered to be unsatisfactory on account of the following points :

i) Absolute measure of skwness is expressed in the same unit of value in which the values of distribution are expressed. Hence absolute measure of skewness is not useful to compare two or more distributions whose values are expressed in different units.

ii) Two or more distributions may have similar frequency curves, but in one series the difference between mean and mode in absolute terms may be greater while in others it may be smaller. Thus absolute measures of skewness does not help to compare the frequency distributions.

## 21.3 RELATIVE MEASURES OF SKEWNESS

The absolute measure of skewness which is expressed in relation to some measure of dispersion is called relative measure of skewness or coefficient of skewness. Relative measure of skewness gives us a pure number which is independent of the units in which the values of distribution are expressed. This enables us to compare and interpret the results of two or more distributions whose values are given in different units.

There are four important methods of relative measures of skewness,they are:

i) The Karl pearson's coefficient of skewness

ii) The Bowley's coefficient of skewness

iii) The Kelley's coefficient of skewness

iv) Measures of skewness based on moments

However, the first two methods are discussed in this book.

## 21.4 KARL PEARSON'S COEFFICIENT OF SKEWNESS

According to' Karl Pearson', skewness is calculated with the help of the following formula :

Karl Pearson's skewness $(SK_P) = \frac{Mean - Mode}{Standard\ deviation}$ .....(i)

Characteristics of Karl Person's coefficint of skewness :

i) Skewness is zero for a symmetrical distribution.

ii) In a positively skeweed distribution, the value of mean is greater than mode.

iii) In a negatively skewed distribution the value of mode is greater than mean:

iv) Theoretically this measure varies within the limits of which $\pm$ 3 is rare phenomenon in the real life. Usually it varies in between $\pm 1$

The above formula is not useful to calculate skewness for a bi-modal frequency distribution. Hence Pearson has suggested another formula which is as follows :

$$SK_P = \frac{3(Mean - Mode)}{Standard\ deviation} \quad .....(ii)$$

This formula is based on the relationship of different averages of a moderately asymmetrical distribution. In such a distribution :

$$Mode = 3\ Median - 2\ Mean$$

$$Mode = 3\ Median - 3\ Mean + Mean$$

$$Mode - Mean = 3(Med - Mean)$$

$$Mean - Mode = 3\ (Mean - Med)$$

If the value of (Mean − Mode) is substituted in the formula (i) we get :

$$\frac{3(Mean - Median)}{Standard\ deviation} \quad .....(iii)$$

**Illustration - 1**

Calculate the coefficient of skewness for the following distribution of weekly wages :

| Weekly wages (Rs.) : | 45 | 60 | 75 | 105 | 120 | 135 | 150 |
|---|---|---|---|---|---|---|---|
| No. of workers : | 6 | 50 | 38 | 32 | 8 | 10 | 4 |

**Solution:**

CALCULATION OF COEFFICIENT OF SKEWNESS

| Wages X | No. of workers f | $\frac{(X-90)}{15}$ $d^1$ | $fd^1$ | $fd^{1^2}$ |
|---|---|---|---|---|
| 45 | 6 | −3 | −18 | 54 |
| 60 | 50 | −2 | −100 | 200 |
| 75 | 38 | −1 | −38 | 38 |
| 90 | 32 | 0 | 0 | 0 |
| 105 | 8 | 1 | 8 | 8 |
| 120 | 10 | 2 | 20 | 40 |
| 135 | 12 | 3 | 36 | 108 |
| 150 | 4 | 4 | 16 | 64 |
| | N=160 | | $\Sigma$ fd = −76 | $\Sigma fd^{1^2}$ =512 |

**Calculation of Mean :**

$$\overline{X} = A + \frac{\Sigma fd^1}{N} \times C$$

$A = 90, \Sigma fd^1 = -76, N = 160, C = 15$

Substituting the values in the formula,

$$\overline{X} = 90 + \frac{-76}{160} \times 15$$

$$= 90 - \frac{21}{4}$$

$$= 90 - 7.125$$

$$= 82.87$$

Calculation of Standard Deviation :

$$\sigma = \sqrt{\frac{\Sigma fd^{1^2}}{N} - \left(\frac{\Sigma fd^1}{N}\right)^2} \times C$$

$\Sigma f d^{1^2} = 512, \Sigma fd^1 = -76$ , N = 160, C = 15

Substituting the values in the formula,

$$\sigma = \sqrt{\frac{512}{160} - \left(\frac{-76}{160}\right)^2} \times 15$$

$$= \sqrt{3.2 - (-0.475)^2} \times 15$$

$$= \sqrt{3.2 - 0.226} \times 15$$

$$= \sqrt{2.974} \times 15$$

$$= 1.723 \times 15$$

$$= 25.845$$

Calculation of Mode :

**GROUPING TABLE**

| Wages | Col.I | Col.II | Col.III | Col.IV | Col.V | Col.VI |
|---|---|---|---|---|---|---|
| 45 | 6 | 56 | | | | |
| 60 | 50 | | 88 | 94 | | |
| 75 | 38 | 70 | | | 120 | 78 |
| 90 | 32 | | 40 | | | |
| 105 | 8 | 18 | | 50 | | |
| 120 | 10 | | 22 | | 30 | |
| 135 | 12 | 16 | | | | 26 |
| 150 | 4 | | | | | |

**ANALYSIS TABLE**

| Col. No. | Wages in Rupees | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 45 | 60 | 75 | 90 | 105 | 120 | 135 | 150 |
| I | | 1 | | | | | | |
| II | | | 1 | 1 | | | | |
| III | | 1 | 1 | | | | | |
| IV | 1 | 1 | 1 | | | | | |
| V | | 1 | 1 | 1 | | | | |
| VI | | | 1 | 1 | 1 | | | |
| | 1 | 4 | 5 | 3 | 1 | | | |

Since the value 75 has occurred the maximum number of times, i.e.5, the model wage is Rs. 75.

$$SKp = \frac{Mean - Mode}{Standard\ Deviation}$$

Here, N = 10, $\Sigma XY = 939$, $\Sigma X = 55$, $\Sigma Y = 135$, $\Sigma X^2 = 385$ and $\Sigma Y^2 = 2303$.

Substituting the values in the formula,

$$r = \frac{10\times939-55\times135}{\sqrt{10\times385-(55)^2}\sqrt{10\times2303-(135)^2}}$$

$$= \frac{9390-7425}{\sqrt{(3850-3025)}\sqrt{(23030-18225)}}$$

$$= \frac{1965}{\sqrt{825\times4805}}$$

$$= \frac{1965}{\sqrt{3964125}}$$

$$= \frac{1965}{1991.011}$$

$$r = 0.987$$

Thus, there is a high degree of positive correlation between X and Y series.

**Illustration - 8**

Calculate coefficient of correlation from the following data:

Number of pairs of observations = 10

$\overline{X} = 33.6$, $\overline{Y} = 21.3$, $\sigma x = 12$, $\sigma y = 8$

Assumed mean of 'X' series = 40

Assumed mean of 'Y' series = 25

Sum of products of deviations of X and Y series from their respective assumed means = 65.

**Solution :**

When deviations are taken from assumed means, we can also apply another formula, i.e.,

$$r = \frac{\Sigma dxdy - N(\overline{X}-Ax)(\overline{Y}-Ay)}{N.\sigma x.\sigma y}$$

Where

$\Sigma dxdy$ = Sum of plroducts of deviations taken from assumed means of X and Y series

$Ax$ = Assumed mean of X series

$Ay$ = Assumed mean of Y series

We are given that N = 10, $\overline{X} = 33.6$, $\overline{Y} = 21.3$, $\sigma x = 12$, $\sigma y = 8$, $Ax = 40$, $Ay = 25$ and $\Sigma dxdy = 65$.

Substituting the values in the formula, we get,

$$r = \frac{65-10(33.6-40)(21.3-25)}{10\times12\times8}$$

$$= \frac{65-10(-6.4)(-3.7)}{960}$$

marks secured by 10 students in a class-test in Accountancy and Statistics :

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Roll Numbers | : | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Marks in Accountancy | : | 50 | 60 | 65 | 30 | 40 | 35 | 70 | 75 | 80 |
| Marks in Statistics | : | 45 | 55 | 60 | 40 | 45 | 60 | 58 | 62 | 72 |

(Ans: r = 0.791)

14. Calculate coefficient of correlation from the following data taking deviations from 48 in case of X series and 20 in case of Y series :

| | | | | | | |
|---|---|---|---|---|---|---|
| X | 40 | 42 | 46 | 48 | 50 | 56 |
| Y | 10 | 12 | 15 | 23 | 27 | 30 |

(Ans: r = 0.956)

## 23.8 RECOMMENDED BOOKS


1. Gupta, S.P. : "Statistical Methods", Sultan Chand & Company, New Delhi.
2. Gupta, B.N. : "Statistics", Sahitya Bhavan, Agra.
3. Gupta, S.C. : "Fundamentals of Statistics", Himalaya Pub. House, Bombay.
4. Simpson and Kafka : "Basic Statistics", Oxford and I.B.H. Publishing Company, Calcutta.


## 23.9 GLOSSARY

1. Co-efficient of Correlation : It is a measure of correlation showing the direction and the degree of correlation between the variables.
2. Graphic Method : A method of studying correlation by plotting the values of two variables seperately on a graph paper.
3. Scatter Diagram Method : It is the pictorial presentation of bivariate data to study correlation.

## UNIT-24 : METHODS OF STUDYING CORRELATION-II

**Contents**



## 24.0 AIMS AND OBJECTIVES

In this unit we aim at explaining the remaining two methods of studying correlation. We introduce two more measures relating to correlation, such as coefficient of determination and probable error.

After going through this unit, you should be able to :

- calculate co-efficient of correlation by concurrent deviation method
- calculate co-efficient of correlation by rank correlation method
- calculate coefficient of determination to explain the change in dependent variable
- calculate probable error to judge the significance of correlation between the variables.

## 24.1 INTRODUCTION

In this unit, we deal with concurrent deviation and rank correlation methods. Under concurrent deviation method, deivations are obtained from the preceding item. While doing so only the direction of deviation i.e., positive or negative is taken into account. After ascertaining the sum of signs coefficient of correlation is calculated through the formula given.

Rank correlation is calculated when two sets of ranks are given. Generally, ranks are given to qualitative characteristics, such as honesty, beauty, morality etc. In such a case, rank correlation is applied.

Co-efficient of determination is used to explain the causes for variation in the dependent variable. Probable error is used to judge the significance of correlation between the variables. Let us get into details.

## 24.2 COEFFICIENT OF CONCURRENT DEVIATION

According to this method, correlation between two variables is obtained on the basis of the direction of change of the variables.

The following steps are involved in finding out the coefficient of correlation by this method.

(i) Find out the direction of change of X variable and denote the column by Dx.
(To obtain the direction of change, one value of a series is compared with its preceding value. If a value is greater than its preceding value, the direction of change is denoted by (+) sign. On the other hand, if a value is lesser than its preceding one, the direction of change is denoted by (–) sign. If a value is equal to that of its preceding value, the direction of change is said to be constant and denoted by '0' (zero). This process is repeated for all the values of 'X' series and the outcome is denoted by Dx.)

(ii) Find out the direction of change of 'Y' variable and denote the column by 'Dy'. (The procedure as adopted to obtain Dx, is also followed to obtain Dy).

(iii) Find out the product of Dx and Dy and obtain the value of 'C', i.e., the number of concurrent deviations.

(iv) Find out the value of 'n' by deducting 1 from the total number of pairs of observations i.e., n = N – 1.

(v) compute coefficient of concurrent deviation by using the following formula:

$$r_c = \pm\sqrt{\pm\left(\frac{2c-n}{n}\right)}$$

Where $r_c$ = Coefficient of concurrent deviations

c = Number of concurrent deviations

n = Number of pairs of observations less one

If 2c–n is – positive, the positive sign must be used within the root so that '$r_c$' is positive. On the other hand, if 2c–n is negative, the negative sign under the root must be used so that $r_c$ is negative.

The following illustrations would further explain the coefficient of correlation by concurrent devation method.

**Illustration - 1**

Calculate the coefficient of correlation by concurrent deviations method from the following data :

| Year | : 1976 | 1977 | 1978 | 1979 | 1980 | 1981 | 1982 | 1983 | 1984 |
|---|---|---|---|---|---|---|---|---|---|
| Supply | : 250 | 270 | 260 | 275 | 280 | 290 | 290 | 285 | 292 |
| Price | : 150 | 130 | 140 | 135 | 132 | 130 | 131 | 134 | 127 |

**Limitations**

(i) This method emphasises only the direction of change, but it does not measure the degree of relationship in precise terms.

(ii) It is not concerned with the extent of change in the values. It gives equal weightage to both big and small changes if they vary in the same direction.

(iii) It is not useful to study the long-term changes in the data as it does not take the trend into account.

(iv) The results obtained by this method give only a rough idea of relationship between the variables.

## 24.3. RANK CORRELATION COEFFICIENT

Charles Edward Spearman developed another method of finding out correlation between two variables. This method is used when the data are not normal or when the shape of the distribution is not known. This method is especially useful when the characteristics of the items are not subject to direct measurement. For, example, certain variables like judgement of female beauty, evaluation of leadership ability, etc., cannot be measured directly in quantitative terms. In such cases, the relationship betwen the variables is obtained by ranking the items. Ranks are assigned to all the items in the series in the order of their size, i.e., the highest value in the series is given first rank, the next highest value is given the second rank and so on. An alternative way of assigning ranks is that the lowest value in the series can be given first rank, the next lowest value, the second rank so on. According to this method, all the calculations are based on ranks rather than the original values of the observations.

Spearman's rank correlation coefficient is denoted by $\rho$ (rho) and computed as detailed below:

(i) Assign ranks to the items of first series and denote the column by $R_1$.

(ii) Assign ranks to the items of second series and denote the column by $R_2$.

(iii) Obtain the differences of the two ranks i.e., $(R_1 - R_2)$ and denote these differences by D.

(iv) Apply the following formula and obtain the rank correlation coefficient.

$$\rho = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

where D = Difference between $R_1$ and $R_2$

N = Number of pairs observed.

The value of this coefficient ranges between $\pm 1$. while $+1$ indicates complete agreement in the order of the ranks, $-1$ implies the complete disagreement in the order of the ranks. The complete agreement in the order of ranks indicates that the ranks are in the same direction. On the other hand, the complete disagreement in the order of ranks indicates that the ranks are in opposite direction.

**Illustration - 3**

Calculate Spearman's coefficient of rank correlation from the following data :

Marks in
Accountary : 60 48 52 50 55 62 65 70 68

Marks in
Statistics : 72 62 68 65 70 60 55 52

**Solution :**

**CALCULATION OF RANK CORRELATION COEFFICIENT**

| Marks in Accountancy | $R_1$ | Marks in Statistics | $R_2$ | $(R_1 - R_2)$ D | $(R_1 - R_2)^2$ $D^2$ |
|---|---|---|---|---|---|
| 60 | 5 | 72 | 2 | 3 | 9 |
| 48 | 9 | 62 | 6 | 3 | 9 |
| 52 | 7 | 68 | 4 | 3 | 9 |
| 50 | 8 | 65 | 5 | 3 | 9 |
| 55 | 6 | 70 | 3 | 3 | 9 |
| 62 | 4 | 60 | 7 | −3 | 9 |
| 65 | 3 | 55 | 8 | −5 | 25 |
| 70 | 1 | 52 | 9 | −8 | 64 |
| 68 | 2 | 79 | 1 | 1 | 1 |
| | | | | | $\Sigma D^2 = 144$ |

$$= 1 - \frac{6\Sigma D^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 144}{9^3 - 9}$$

$$= 1 - \frac{864}{729 - 9}$$

$$= 1 - \frac{864}{720}$$

$$= 1 - 1.2$$

$$= -0.2$$

Thus, rank correlation coefficient $= -0.2$.

**Check your progress - 1**

Calculate Rank correlation coefficient from the ranks given below

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $R_1$ | 1 | 2 | 4 | 6 | 7 | 5 | 3 |
| $R_2$ | 2 | 1 | 4 | 5 | 6 | 7 | 3 |

**Equal Ranks**

Sometimes, there may be two or more individual items which may have equal values. In such cases, each individual item is given an average rank. For example, if two individual items are ranked equal at third place, each of them is given the rank 3.5 (i.e., $\frac{3+4}{2} = 3.5$ ) If three items are ranked equal at third place, each of them is given the rank 4 (i.e, $\frac{3+4+5}{3} = 4$). In other words, when two or more items are to be ranked equal, they are given the average of the ranks which these individual items would have got, had they differed from each other. In such cases, rank correlation coefficient is calculted after adding $\frac{1}{12}(m^3 - m)$ to the value of $\Sigma D^2$ in the original formula. Here $m$ stands for the number of items whose ranks are common. If more than one such groups are there with common rank, the adjustment factor will have to be added to the value of $\Sigma D^2$ as many number of items as the number of such groups. Thus, the formula can be written as:

$$= 1 - \frac{6\left\{\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m)\right\}}{N^3 - N}$$

**Illustration - 4**

Compute rank correlation coefficient from the following data :

X : 60 55 60 70 45 35 27 60

Y : 40 35 38 42 35 50 48 60

**solution :**

**COMPUTATION OF RANK CORRELATION COEFFICIENT**

| X | $R_1$ | Y | $R_2$ | $(R_1 - R_2)$ D | $(R_1 - R_2)^2$ $D^2$ |
|---|---|---|---|---|---|
| 60 | 3 | 40 | 5 | −2.0 | 4.00 |
| 55 | 5 | 35 | 7.5 | −2.5 | 6.25 |
| 60 | 3 | 38 | 6 | −3.0 | 9.00 |
| 70 | 1 | 42 | 4 | −3.0 | 9.00 |
| 45 | 6 | 35 | 7.5 | −1.5 | 2.25 |
| 35 | 7 | 50 | 2 | 5.0 | 25.00 |
| 27 | 8 | 48 | 3 | 5.0 | 25.00 |
| 60 | 3 | 60 | 1 | 2.0 | 4.00 |
| N = 8 | | | | | $\Sigma D^2 = 84.50$ |

$\Sigma D^2 = 84.5$, N = 8

The item 60 is repeated three times in series 'X' so $m = 3$. The item 35 is repeated twice in series 'Y' so $m = 2$.

$$= 1 - \frac{6\left\{\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m)\right\}}{N^3 - N}$$

Substituting the values in the formula,

$$= 1 - \frac{6\left\{84.5+\frac{1}{12}(3^3-3)+\frac{1}{12}(2^3-2)\right\}}{8^3-8}$$

$$= 1 - \frac{6\left\{84.5+\frac{1}{12}(24)+\frac{1}{12}(6)\right\}}{504}$$

$$= 1 - \frac{6\left\{84.5+2+0.5\right\}}{504}$$

$$= 1 - \frac{6(87.0)}{504}$$

$$= 1 - \frac{522.0}{504}$$

$$= 1 - 1.036$$

$$= -0.036$$

Thus, rank correlation coefficient $= -0.036$.

**Merits and Limitations of Rank Correlation Coefficient**

**Merits**

(i) It is the simplest method of calculating correlation between two variables.

(ii) When we are not given original data and only the ranks are given, this is the only method of computing correlation co-efficient.

(iii) In the case of qualitative data, which cannot be measured directly, this method is used to calcualte correlation coefficient by simply assigning ranks to the attributes.

(iv) This method is not affected by extreme values of series as correlation coefficient is calcualted not on the basis of orginal values, but on the basis of ranks assigned to them.

**Limitations**

This method suffers from the following limitations:

(i) This method gives only a rough idea of the relationship between the variables as the calculations are not based on actual observations.

(ii) This method is not useful for computing correlation between the variables when the observations are fairly large. Especially when the observations are more than 30, ranking becomes difficult and cumbersome.

(iii) This method is useful to compute corelation in the case of single observations only. It is not useful in the case of grouped frequency distribution.

## 24.4 COEFFICIENT OF DETERMINATION

Coefficient of correlation reveals the direction and degree of relationship between the variables. It does not specify the reasons of change in the variables. The change in the dependent variable may be due to two factors: (i) variation in the independent variable; and (ii) variation unaccounted by the independent variable and due to some other factors. The extent of variance in the dependent variable due to the independent variable is explained by the coefficient of determination. Coefficient of determination is equal to $r^2$ and is the ratio of the explained variance to the total variance. Thus,

$$Coefficient\ of\ determination = \frac{Explained\ variance}{Total\ variance}$$

Coefficient of determination is a convenient and useful statistical technique of interpreting the value of correlation coefficient. If the coefficient of correlation betw en two variables is 0.6, coefficient of determination would be 0.36. This would mean that 36% of the variation in the dependent variable is due to the independent variable and the remaining 64% of the variation is due to other factors. The relationship between $r$ and $r^2$ is given below.

| $r$ | $r^2$ | percentage of explained variance |
|---|---|---|
| 1.00 | 1.00 | 100 |
| 0.9 | 0.81 | 81 |
| 0.8 | 0.64 | 64 |
| 0.7 | 0.49 | 49 |
| 0.707 | 0.50 | 50 |
| 0.6 | 0.36 | 36 |
| 0.5 | 0.25 | 25 |
| 0.4 | 0.16 | 16 |
| 0.3 | 0.09 | 9 |
| 0.2 | 0.04 | 4 |
| 0.1 | 0.01 | 1 |



It is lcear from the above that as the value of r decreases from its maximum value of 1, the value of $r^2$ also decreases, but at a more rapid rate. Usually the value of r is always greater than the value of $r^2$ unless$r = 0$ or 1 in which case $r = r^2$. Further, the value of coefficient of determination is always positive. The value of $r$ may quite often mislead the readers, e.g., correlation between two variables = 0.8 and the correlation between two other variables = 0.4. It does not mean that the correlation between the first two variables is double the correlation between the second set of variables. The relationship between the variables can be better understood by computing the value of their respective coefficient of determination. The coefficients of determination would be 0.64 and 0.16 respectively. This indicates that in the first case, 64% of the total variation in the dependent variable and in the second case, only 16%

of the total variation is explained by the independent variable.

The usefulness of coefficient of determination is clear from the observations of Tuttle who states, "The results of correlation analysis has been grossly over rated and is used entirely too much. Its square, the coefficient of determination is a much more useful measure of the linear covariation of two variables. The reader should develop the habit of squaring every correlation coefficient he finds cited or stated before coming to any conclusion about the extent of the linear relationship between the two correlation variables.

Though the coefficient of determination is a useful measure of interpreting the value of correlation coefficient, it does not establish any sort of causal relationship between the variables. This has to be determined on the basis of the evidence other than the quantitative observations.

Another limitation of coefficient of determination is that it is always a positive (+) number. Thus, it does not tell us the direction of covariation of the variables.

However, correlation coefficient and coefficient of determination used together will serve a better analytical purpose. the combined usage of correlation coefficient and coefficient of determination is very common in economic analysis.

## 24.5 PROBABLE ERROR

The reliability or significance of correlation coefficient is determined

with the help of probable error. The probable error helps in interpreting the value of coefficient of correlation. According to Horace Secrist, "The probable error of $r$ is an amount which if added to and subtracted from the average correlation coefficient produces amounts within which the chances are even that a coefficient of correlation from a series selected at random will fall". According to Wheldon, "Probable error defines the limits above and below the size of the coefficient determined within which there is an equal chance that coefficient of correlation similarly calculated from other samples will fall". The probable error of the coefficient of correlation is calculated with the help of the following formula:

$$P.E_r = 0.6745\frac{1-r^2}{\sqrt{N}}$$

Where 0.6745 is constant number

'r' is pearsonian coefficient of correlation

'N' is number of pairs of observations.

Probable error of coefficient of correlation is interpreted in the following way :

(i) If the value of coefficient of correlation is less than the probable error, it can be said that there is no evidence of correlation between the variables under study.

(ii) If the value of coefficient of correlation is greater than six times the probable error th : coefficient of correlation is said to be practically certain.

(iii) The upper and lower limits within which the value of coefficient of correlation in the population lies are obtained by adding and subtracting the value of probable error from the concerned coefficient of correlation.

Symbolically,

$$\rho = r \pm P.E_r$$

Where $\rho$ (rho) denotes correlation in the population.

**Conditions for the use of Probable Error**

According to Riggleman and Frisbee, the statistical measure of probable error can be properly used only when the following three conditions exist:

(i) The data must approximate normal frequency curve (bell shaped curve).

(ii) The statistical measure for which the probable error is computed must have been calculated from a sample.

(iii) The sample must have been selected in an unbiased manner and the individual items must be independent.

**Illustration - 5**

If $r = 0.4$ and $N = 16$, find out the probable error of the coefficient of correlation and determine the limits for population coefficient of correlation ($\rho$).

**Solution :**

$$P.E._r = 0.6745\frac{1 - r^2}{\sqrt{N}}$$

We are given that,

$$r = 0.4 \text{ and } N = 16$$

Substituting the values in the formula

$$P.E._r = 0.6745\ \frac{1-(0.4)^2}{\sqrt{16}}$$

$$= 0.6745\ \frac{1-0.16}{\sqrt{16}}$$

$$= \frac{0.6745 \times 0.84}{4}$$

$$= \frac{0.567}{4}$$

$$= 0.142$$

Limits of population corelation ($\rho$ ) $= 0.4 \pm 0.142$

$$= 258 - 0.542$$

## 24.6 SUMMING UP

According to concurrent deviation method, the correlation between two variables is obtained on the basis of the direction of change of the variables. Rank correlation is computed by assigning ranks to the given variables in the series in order of their size.

The coefficient of determination is the ratio of the explained variance to the total variance. The reliability or significance of correlation co- efficient is determined with the help of probable error. It helps to interpret the value of coefficient of correlation.

## 24.7 CHECK YOUR PROGRESS : MODEL ANSWERS

1.

| $R_1$ | $R_2$ | $(R_1 - R_2)$ D | $D^2$ |
|---|---|---|---|
| 1 | 2 | −1 | 1 |
| 2 | 1 | 1 | 1 |
| 4 | 4 | 0 | 0 |
| 6 | 5 | 1 | 1 |
| 7 | 6 | 1 | 1 |
| 5 | 7 | −2 | 4 |
| 3 | 3 | 0 | 0 |
| | | | 8 |

$\rho = 1 - \frac{6\Sigma D^2}{N^3 - N}$ $\quad \Sigma D^2 = 8;\ N = 7$

$\rho = 1 - \frac{6 \times 8}{7^3 - 7} = 1 - \frac{48}{343 - 7}$

$\rho = 1 - \frac{48}{336} = 1 - 0.412$

$\rho = 0.858$



## 24.8 MODEL EXAMINATION QUESTIONS

### A. Short Questions

1. Define 'coefficient of Concurrent Deviations'.
2. Define Rank Correlation.
3. Define 'Coefficient of Determination'.
4. What is a probable error?
5. What are the circumstances in which rank correlation coefficient is useful ?

## EXERCISES

6. Calculate the coefficient of concurrent deviations from the following data:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Firms | : A | B | C | D | E | F | G | H | I | J |
| Sales (Rs.) | : 50 | 55 | 55 | 60 | 65 | 65 | 65 | 60 | 70 | 80 |
| Expenses(Rs.) | : 11 | 13 | 14 | 16 | 16 | 15 | 15 | 14 | 18 | 21 |

(Ans : $r_c = -0.33$)

7. Calculate the coefficient of concurrent deviations from the following data:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Year | : '75 | '76 | '77 | '78 | '79 | '80 | '81 | '82 | '83 | '84 |
| Supply (Kg.): | 80 | 82 | 86 | 83 | 85 | 89 | 96 | 93 | 92 | 97 |
| Price (Rs.) : | 146 | 140 | 130 | 137 | 133 | 127 | 115 | 95 | 100 | 97 |

(Ans : $r_c = -0.882$)

8. The ranks secured by a group of 10 students in written selection test (X) and the Aptitude Test (Y) are given below. Calculate coefficient of rank coefficient.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Students group : | I | II | III | IV | V | VI | VII | VIII | IX | X |
| X | : 2 | 5 | 3 | 9 | 6 | 4 | 1 | 7 | 8 | 10 |
| Y | : 4 | 3 | 2 | 1 | 5 | 6 | 7 | 8 | 10 | 9 |

(Ans : $R_k = 0.273$)

9. Find out correlation coefficient for the following data:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Marks in Accountancy : | 24 | 30 | 36 | 31 | 28 | 20 | 19 | 26 | 30 |
| Marks in Statistics : | 31 | 39 | 33 | 35 | 38 | 32 | 30 | 19 | 37 |

(Ans : $R_K = -0.583$)

10. Find out rank correlation coefficient for the following data:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| X : | 40 | 55 | 72 | 35 | 40 | 48 | 65 |
| Y : | 60 | 65 | 45 | 50 | 51 | 45 | 45 |

(Ans : $R_K = -0.411$)

## 24.9 RECOMMENDED BOOKS


1. Gupta, S.P. : "Statistical Methods ", Sultan Chand & Company, New Delhi.
2. Gupta, B.N. : "Statistics", Sahitya Bhavan, Agra.
3. Gupta, S.C : "Fundamentals of Statistics", Himalaya Pub.House, bombay.
4. Simpson and Kafka : "Basic Statistics ", Oxford and I.B.H. Publishing Company, Calcutta.

## 24.10 GLOSSARY

| | | |
|---|---|---|
| 1. Coefficient of concurrent deviation | : | A method of calculating correlation coefficient by taking into account the direction of deviation of variables. |
| 2. Co-efficient of determination | : | It is the measure which explains the extent of variation in dependent variable due to variation in independent variable. |
| 3. Probable Error | : | It is the measure of judging the significance of correlation between the variables. |
| 4. Rank correlation | : | It is a method of calculating correlation between two sets of ranks. |

## UNIT-25 : REGRESSION ANALYSIS

**Contents**





## 25.0 AIMS AND OBJECTIVES

The aims of this unit are to explain the meaning, importance and computational process of regression.

After going through this unit, you should be able to :

- define the term 'Regression analysis'
- distinguish between correlation and regression
- indentify the types of regression
- list out the uses of regression analysis
- draw the regression lines
- explain the regression equations
- indentify the properties of regression coefficeints
- compute the problems on regression.

## 25.1 INTRODUCTION

The Dictionary meaning of the word 'regression'is 'to revert' or 'return back to normal'. This term was frist introduced by Sir Francis Galton in 1877 in his study of relationship between the heights of fathers and sons. The results of his research work relating to about 1,000 fathers and sons, revealed an interesting relationship that tall fathers were likely to have tall sons, but the average height of the sons would be less than the average height of fathers. On the other hand, short fathers were likely to have short sons, but the average height of sons would be more than the average height of the fathers. Thus, Galton proved that there is a tendency of human race to 'regress' or 'return to a normal height'. The deviations of heights of fathers are likely to be corrected by the heights of their sons. In the words of F.C Mills, 'Sons deviate less on the average from the mean height of the race than their fathers, whether the fathers were above or below the average, sons tended to go back or regress towards mean". If this regressive tendency does not exist, the human race would have broken up into two extremes of giants and pygmies. the term regression, originated in this context, is widely used now in various fields of study despite the absence of regressive tendency in the variables studied.

While correlation analysis determines the presence or absence of relationship and the extent of covariance between the variables, regression analysis helps to estimate the value of one variable for a given value of another variable. For example, if we know that price and demand are closely related, we may find out the expected amount of demand for a given price level and vice-versa. Regression analysis reveals the average relationship between two variables and this makes possible to estimate or predict the values of variables.

## 25.2 DEFINITION OF REGRESSION ANALYSIS

Some of the important definitions of the term regression are given below :

(i) According to Blair, "Regression is the measure of the average relationship between two or more variables in terms of the original units of the data".

(ii) According to 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".

(iii) Morris Hamburg defined the term 'regression analysis' as " the methods by which estimates are made of the values of a variable from a knowledge of the values of other variables and to the measurement of the errors involved in this estimation process".

A close observation of above definitions reveals the following :

(i) Regression analysis is a statistical technique with the help of which unknown values of one variable are estimated or predicted on the basis of known values of another variable.

(ii) It is a measure of the average relationship between two or more variables.

(iii) It attempts to establish cause-and-effect relationship between two or more variables.

(iv) It reveals the relationship between the variables in terms of the original units of measure in which the data under study are expressed.

(v) It attempts to establish functional relationship between two or more variables.

## 25.3 DISTINCTION BETWEEN CORRELATION AND REGRESSION

Both the correlation and regression techniques are concerned with the common objective of establishing the degree and direction of relationship between two or more variables. But there are certain basic differences regarding the nature, scope and importance of correlation and regression analysis. The following are some of the points of distinction between correlation and regression :

| | CORRELATION | REGRESSION |
|---|---|---|
| (i) | Correlation refers to the movements in two or more related variables which may be either in the same direction or in the reverse direction. | Regression is a statistical technique which measures the average relationship between two or more variables. |
| (ii) | Correlation analysis determines the presence or absence of relationship between the variables. It does not establish any cause and effect relationship between the variables. | Regression studies not only the relationship between the variables but also establishes 'cause and effect' relationship between the variables. |
| (iii) | In the case of correlation analysis, there is an interdependence between the variables. Such interdependence is mutual and is immaterial whether $Y = f(X)$ or $X = f(Y)$. | It establishes a functional relationship between the variables. Such relationship is mathematical and shows the dependence of one variable on the other. the functional relationship may be either $Y = f(X)$ or $X = f(Y)$ and both need not be true. |
| (iv) | If the correlation between the variables studied is influenced by external factors. i.e., other than the variables studied, it may be a case for non-sense or spurious correlation. | There is no such non-sense or spurious regression.The relationship is mathematical and explained quantitatively. |

| | | |
|---|---|---|
| (v) | Correlation coefficient is a relative measure and obtained in terms of a pure number which is independent of the units in which the original data are expressed. | It is an abosolute measure and the relationship between the variables is expressed in terms of the units of measure in which the original data are expressed. |
| (vi) | Correlation coefficient tells the degree and direction of relationship between the variables. The range of relationship lies in between ± 1. | With the help of regression analysis the values of dependent variable can be ascertained on the basis of corresponding values of independent variable. |
| (vii) | Correlation technique is a simple statistical device and it is used for testing and verifying the relationship between the variables. | Regression technique is not only used for testing and verifying the relationship between the varibales, but it can also be used for predicting the values of variables. |

## 25.4 TYPES OF REGRESSION

Regression analysis may be classified as under :

(a) Simple and multiple regression analysis

(b) Linear and non-linear regression analysis

(c) Total and partial regression analysis

**(a) Simple and multiple regression analysis**

Simple regression analysis is concerned with the study of two variables, only, i. e., the impact of price on demand for a commodity. In this case price is taken as an independent variable (X) and demand for the commodity is taken as a dependent variable (Y). With this, the functional relationship between price and demand is expressed as Y = f(X). This is a general form of expression and it does not tell us whether the relationship is linear or non-linear. On the other hand, multiple regression is concerned with the study of more than two variables. In this case, one variable is a dependent variable and the remaining are independent variables. For example, sales (Y) may depend on the amount spent on advertisement (X) and income levels of people (i). This type of functional relationship is expressed as Y = f(X & i). Like simple regression analysis, multiple regression analysis also does not tell us whether the relationship among the variables is linear or non-linear.

**(b) Linear and non-linear regression analysis**

Linear relationship is based on straight line trend. It can be both simple and multiple. The linear relationship between the variables is simple to calculate and easy to understand. It helps in predicting the future values. On the other hand, non-linear relationship gives curved trend lines which are called parabolic. In the case of non-linear relationships, non-linear regression coefficients are used to predict the values. Unlike linear relationship, it is difficult to calculate and understand the non-linear relationship. This is the reason why in most of the studies linear relationship is assumed despite its absence in the data.

**(c) Total and partial regression analysis**

All the related variables are studied in case of total regression analysis. It takes the form of a multiple relationship. Usually economic and business phenomena are affected by multiplicity of inter-dependent and interrelated operating forces. In the case of partial regression analysis, the study is confined to the study of some of the variables, but not all, i.e., in the case of sales (Y), advertisement expenditure (X), income level of people (i) and price of the commodity (P), the functional relationship will be;

In the case of total relationship : $Y = f$ (X, i and P)

In the case of partial relationship : $Y = f$ (X but not i and P)

$Y = f$ (i but not X and P)

$Y = f$ (P but not X and I)

## 25.5 USES OF REGRESSION ANALYSIS

The following are some of the practical uses of regression analysis :

(i) Regression analysis helps us to find out functional relationship between two or more variables. With the help of such relationship, the values of dependent variables can be estimated on the basis of the values of independent variables.

(ii) regression analysis also helps us to obtain cause and effect relationship between two or more variables. With the help of such a relationship the inter-dependency and interrelationship among the variables can be ascertained. This type of study enables us to find out the nature and type of relationship among the variables.

(iii) Regression analysis is widely used in the study and analysis of economic and business problems to find out the 'cause and effect' relationships. Regression technique is widely used in the statistical estimation of demand and supply curves, production functions, cost functions and consumption functions.

(iv) With the help of regression coefficients, correlation coefficient can be calculated which in turn helps to obtain the coefficient of determination. With the help of coefficient of determination, we can interpret the degree and direction of the relationship between the variables.

(v) Regression analysis helps us to obtain the error involved in the estimation of values of the dependent variable on the basis of the values of independent variable. This is done with the help of the standard error of estimate.

## 25.6 REGRESSION LINES

To quote J.R. Stockton, the line of regression is "the device used for estimating the values of one variable form the value of the other, consists of a line, through the points drawn, in such a manner as to represent the average relationship between the two variables." If two variables X and Y are chosen for regression anlysis, there will be two regression lines as the regression line of X on Y and regression line of Y on X. While the regression line of X on Y gives the most probable values of X for given values of Y, regression line of Y on X gives the most probable values of Y for given values of X. Both the regression lines cut each other at the point of their respective means, i.e., if a perpendicular line is drawn on X-axis from the point of their intersection, it touches the X-axis at the mean value of X series. Similarly, if a perpendicular line is drawn on Y-axis from the point of their intersection, it touches Y-axis at the mean value of Y series. With the help of regression lines, we can obtain the following average relationships :

(i) If the correlation between two variables is perfect (either positive or negative), the two regression lines will coincide and we will have only one line.

(ii) If the correlation between two variables is zero (i.e., two variables are independent), the regression lines fall at right angles.

(iii) Nearer the two regression lines, higher would be the correlation between the variables.

(vi) Farther the regression lines (from each other), lesser would be the correlation between the variables.




Fig. 25.1 If r = +1

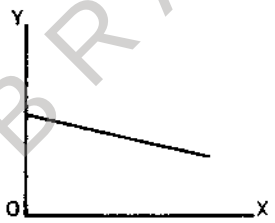



Fig. 25.2 If r = −1

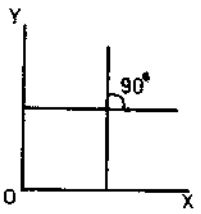



Fig. 25.3
If r = 0

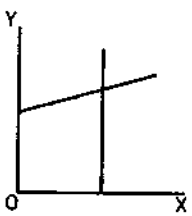



Fig. 25.4
Low degree of 'r'

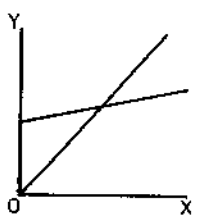



Fig. 25.5
High degree of 'r'

Regression lines are drawn on the assumption of 'least squares'. According to this assumption the sum of squares of the deviations of the observed values of Y from the fitted line would be the minimum possible i.e., $\Sigma(Y - Y_c)^2$ is minimum. Further, the sum of deviations above the line is equal to the sum of deviations below the line, i.e., $\Sigma(Y - Y_c) = 0$.

## 25.7 REGRESSION EQUATIONS

The algebraic expressions of regression lines are called regression equations. Since there are two regression lines in the case of two variables, there are two regression equations (i) the regression equation of X on Y and (ii) the regression equation of Y on X. While the regression equation of X on Y describes the variation in the values of X for the given changes in the values of Y, the regression equation of Y on X describes the variation in values of Y for the given changes in the values of X.

**Regression equation of X on Y :**

Symbolically, the regression equaion of X on Y is expressed as $X_c = a+by$.

Where $X_c$ is the most probable value of X (computed) and 'a' and 'b' are constants, while 'a' denotes the level of the fitted line (i.e., the distance of the line directly above or below the origin), 'b' denotes the slope of the line (i.e., the change in X variable per unit change in Y variable). The values of 'a' and 'b' are obtained through the following two normal equations:

$$\Sigma X = Na + b\Sigma Y \qquad \text{...(i)}$$

$$\Sigma XY = a\Sigma Y + b\Sigma Y^2 \qquad \text{.....(ii)}$$

While $\Sigma X$, $\Sigma Y$, $\Sigma XY$ and $\Sigma Y^2$ indicate the totals that are obtained from the original values of X and Y series, N denotes the number of pairs observed for the purpose of regression analysis.

**Regression Equation of Y on X**

Symbolically, regression equation of Y on X is expressed as $Y_c = a + bX$, where $Y_c$ is the most probable value of Y (computed) and 'a' and 'b' are constants. while 'a' denotes the level of the fitted lines, 'b' denotes the slope of the line (i.e., the change in 'Y' variable per unit change in 'X' variable). The values of 'a' and 'b' are obtained through the following two normal equations:

$$\Sigma Y = Na + b\Sigma X \qquad \text{....(i)}$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 \qquad \text{....(ii)}$$

While $\Sigma Y$, $\Sigma X$, $\Sigma XY$ and $\Sigma X^2$ indicate the totals that are obtained from the original values of X and Y series, N denotes the number of pairs observed for the purpose of regression analysis.

## 25.8 REGRESSION COEFFICIENTS

Regression coefficient indicates the degree and the direction of change in the dependent variable in resposnse to a unit change in the independent variable. Regression coefficients are denoted by 'b' in the regression equations. Since we have two regression equations for two variables, we will also have two regression coefficients; one for the regression equation of X on

Y and the other for regression equation of Y on X.

**(A) Regression Coefficient of X on Y**

Regression coefficient of X on Y indicates the degree and the direction of change in 'X' variable in response to unit change in 'Y' variable. It is denoted by 'b' or 'bxy' and calculated with the help of the following methods :

(i) By solving the Normal Equations

Regression equation of X on Y,

$X_c$ = a+by and the two normal equations are :

$$\Sigma X = Na + b\Sigma Y \quad \text{.....(i)}$$

$$\Sigma XY = a\Sigma Y + b\Sigma Y^2 \quad \text{....(ii)}$$

Where, ΣX and ΣY denote the totals of values of X and Y series respectively.

$\Sigma Y^2$ = Sum of the squares of values of 'Y' series

ΣXY = Sum of the products of the values of X and Y series

N = Number of pairs observed

While'b' denote the regression coefficient of X on Y, 'a' denotes the level of the fitted line. The value of 'b' can be computed by solving the two normal equations given above.

(ii) when deviations are taken from arithmetic means of X and Y

$$\text{bxy} = r\frac{\sigma x}{\sigma y}$$

Where, bxy = regression coefficient of X on Y

r = Coefficient of correlation between X and Y

$\sigma$x = Standard deviation of 'X' series

$\sigma$y = standard deviation of 'Y' series

The simplified version of the above formula is

$$\text{bxy} = \frac{\Sigma xy}{\Sigma y^2}$$

Where, Σxy = Sum of the products of deviations of X and Y variables taken from their respective arithmetic means

$\Sigma y^2$ = sum of the squares of deviations of Y series taken from its arithmetic mean.

This is derived as below :

$$\text{bxy} = r\frac{\sigma x}{\sigma y} = \frac{\Sigma xy}{N.\sigma x.\sigma y} \times \frac{\sigma x}{\sigma y}$$

$$\text{bxy} = \frac{\Sigma xy}{N.\sigma y^2} \text{ where } \quad \sigma y^2 = \text{variance of Y series}$$

and $\sigma y^2 = \frac{\Sigma y^2}{N}$

$$\text{bxy} = \frac{\Sigma xy}{N\frac{\Sigma y^2}{N}}$$

$$bxy = \frac{\Sigma xy}{\Sigma y^2}$$

(iii) when deviations are taken from assumed means

$$bxy = \frac{\Sigma dxdy - \frac{\Sigma dx . \Sigma dy}{N}}{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}$$

Where,

$\Sigma dx$ = sum of deviations of X series taken from its assumed mean

$\Sigma dy$ = Sum of deviations of Y series taken from its assumed mean.

$\Sigma dxdy$ = Sum of the products of deviations of X and Y variables taken from their respective assumed means

$\Sigma dy^2$ = Sum of squares of deviations of 'Y' series taken from its assumed mean

$\Sigma dx^2$ = Sum of squares of deviations of 'X' series taken from its assumed mean

N = Number of pairs observed

(iv) when figures are given in original values

$$bxy = \frac{\Sigma XY - N\overline{X}\overline{Y}}{Y^2 - N(\overline{Y})^2}$$

Where bxy = Regression coefficient of X on Y

$\overline{X}$ and $\overline{Y}$ = denote the arithmetic means of X and Y series respectively

$\Sigma Y^2$ = Sum of squares of values of 'Y' series

$\Sigma XY$ = sum of products of values of X and Y

N = Number of pairs observed

**(B) Regression coefficient of Y on X**

Regression coefficient of Y on X indicates the degree and the direction of change in 'Y' variable in reponse to unit change in 'X' variable. It is denoted by 'b' or byx and calculated with the help of the following methods :

(i) By solving the normal equations

Regression equation of Y on X, $Y_c = a + bx$ and the two normal equations are :

$$\Sigma Y = Na + b\Sigma X \qquad \text{....(i)}$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 \qquad \text{....(ii)}$$

Where, $\Sigma$ X and $\Sigma$ Y denote the totals of values of X and Y series.

$\Sigma X^2$ = Sum of the squares of the values of 'X' and Y series.

$\Sigma XY$ = sum of the products of values of 'X' and 'Y' series

N = Number of pairs observed

While 'b' denotes the regression coefficient of Y on X, 'a' denotes the level of the fitted line. The value of 'b' can be computed by solving the two normal equations given above.

(ii) When deviations are taken from arithmetic means of X and Y

$$byx = r\frac{\sigma y}{\sigma x}$$

The simplified version of the above formula is

$$byx = \frac{\Sigma xy}{\Sigma X^2}$$

Where, $\Sigma x^2$ denote the sum of squares of deviations of 'X' series taken from its arithmetic mean.

The simplified version of the formula is derived through the folowing steps :

$$byx = r\frac{\sigma y}{\sigma x}$$

$$= \frac{\Sigma xy}{N.\sigma x.\sigma y} \times \frac{\sigma y}{\sigma x}$$

$$= \frac{\Sigma xy}{N.\sigma x^2} \text{ where } \quad \sigma x^2 = \text{variance of X series}$$

$$\text{and } \sigma x^2 = \frac{\Sigma X^2}{N}$$

$$byx = \frac{\Sigma xy}{N\frac{\sigma x^2}{N}}$$

$$\text{Thus, } byx = \frac{\Sigma xy}{\Sigma x^2}$$

(iii) when deviations are taken from the assumed means

$$byx = \frac{\Sigma dxdy - \frac{\Sigma dx.\Sigma dy}{N}}{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}}$$

Where dx and dy denote the deviations of X and Y series taken from their respective assumed means.

$\Sigma dxdy$ = sum of the products of deviations of X and Y variables taken from their respective assumed means

$\Sigma dx^2$ = sum of squares of deviations of X series taken from its assumed mean

$\Sigma dy^2$ = Sum of squares of deviations of Y series taken from its assumed mean

N = Number of pairs observed

(iv) When figures are given in original values

$$byx = \frac{\Sigma XY - N\bar{X}\bar{Y}}{\Sigma X^2 - N(\bar{X})^2}$$

where, byx = Regression coefficient of Y on X

$\bar{X}$ and $\bar{Y}$ denote the arithmetic means of X and Y series respectively.

$\Sigma X^2$ = Sum of squares of values of 'X' series

$\Sigma XY$ = Sum of products of values of X and Y series

N = Number of pairs observed

## 25.9 PROPERTIES OF REGRESSION COEFFICIENTS

The follwoing are some of the important properties of regression coefficeints :

(i) Both the regression coefficients will have the same sign i.e., both of them will have either (+) sign or (−) sign. There is no chance of obtaining (+) sign for one regression coefficients and (−) sign for another regression coefficient.

(ii) The square root of the product of both the regression coefficients is equal to the coefficient of correlation between the variables.

Symbolically,

$$r = \sqrt{bxy \times byx}$$

(iii) The coefficient of correlation will have the same sign as that of regression coefficients. If bxy and byx are negative, 'r' will also be negative. On the other hand, if bxy and byx are positive, 'r' will also be positive.

(iv) Since 'r' is always either equal to or lesser than 1, the square root of the product of bxy must be less than 1. In any case, it will not exceed. 1. To be more clear, the values of both the regression coefficients cannot be greater than 1. If bxy is grater than 1, byx should be lesser than 1 and vice versa.

(v) The arithmetic mean of bxy and byx is either equal to or greater than the coefficient of correlation.

symbolically,

$$\frac{bxy+byx}{2} \geq r$$

(vi) The regression coefficients bxy and byx are not symmetric i.e., bxy$\neq$ byx. Hence, a clear cut identification is necessary to differentiate dependent and independent variables.

(vii) If the value of a regression coefficient is positive, the slope of the regression line will be from left to right, upwards. On the other hand, if the value of a regression coefficient is negative, the slope will also be negative i.e., it will be downward from left to right.

## 25.10 COMPUTATION OF REGRESSION COEFFICIENTS

**(i) By solving the normal equations**

The following illustrations would explain the computation of regression coefficients :

**Illustration - 1**

From the following data, find the two regression coefficients.

X = 4 2 6 8 10 5 7
Y = 7 10 8 9 5 6 4

**Solution :**

COMPUTATION OF REGRESSION COEFFICIENTS

| X | $X^2$ | Y | $Y^2$ | XY |
|---|---|---|---|---|
| 4 | 16 | 7 | 49 | 28 |
| 2 | 4 | 10 | 100 | 20 |
| 6 | 36 | 8 | 64 | 48 |
| 8 | 64 | 9 | 81 | 72 |
| 10 | 100 | 5 | 25 | 50 |
| 5 | 25 | 6 | 36 | 30 |
| 7 | 49 | 4 | 16 | 28 |
| $\Sigma X = 42$ | $\Sigma X^2 = 294$ | $\Sigma Y = 49$ | $\Sigma Y^2 = 371$ | $\Sigma XY = 276$ |

$\Sigma X = 42$, $\Sigma Y = 49$, $\Sigma X^2 = 294$, $\Sigma Y^2 = 371$, $\Sigma XY = 276$ and N = 7.

Regression equation of X on Y : $X_c = a+by$. the two normal equations are :

$$\Sigma X = Na + b\Sigma Y$$

$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

Substituting the values, we get,

$$42 = 7a+49b \quad \text{....(i)}$$

$$276 = 49a+371b \quad \text{....(ii)}$$

Multiplying equation (i) by 7 :

$$294 = 49a+343b \quad \text{.....(iii)}$$

$$276 = 49a+371b \quad \text{.....(iv)}$$

Subtracting equation (iv) from equation (iii),

$$18 = -28b$$

Hence $b = -\frac{18}{28}$

$$= -0.64$$

Thus, the regression coefficient of X on Y is $-0.64$.

Regression equation of Y on X : $Y_c = a+bx$ and the two normal equations are :

$$\Sigma Y = Na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

Substituting the values, we get,

$$49 = 7a+42b \quad \text{.....(i)}$$

$$276 = 42a + 294b \quad \text{....(ii)}$$

Multiplying equation (i) by 6,

$$294 = 42a + 252b \quad \text{....(iii)}$$

$$276 = 42a + 294b \quad \text{....(iv)}$$

Subtracting equation (iv) from equation (iii),

$$18 = -42b$$

Hence $b = -\frac{18}{42}$

$$= -0.43$$

Thus, the regression coefficients of Y on X is $-$ 0.43.

(ii) Computation of regression coefficients if deviations are taken from arithmetic means of X and Y :

**Illustration - 2**

From the data of illustration $-$ 1, claculate the regression coefficients, taking the deviations of items from the means of X and Y series.

**Solution :**

| | $(X - \overline{X})$ | | | $(Y - \overline{Y})$ | | |
|---|---|---|---|---|---|---|
| X | x | $x^2$ | Y | y | $y^2$ | xy |
| 4 | $-2$ | 4 | 7 | 0 | 0 | 0 |
| 2 | $-4$ | 16 | 10 | 3 | 9 | $-12$ |
| 6 | 0 | 0 | 8 | 1 | 1 | 0 |
| 8 | 2 | 4 | 9 | 2 | 4 | 4 |
| 10 | 4 | 16 | 5 | $-2$ | 4 | $-8$ |
| 5 | $-1$ | 1 | 6 | $-1$ | 1 | 1 |
| 7 | 1 | 1 | 4 | $-3$ | 9 | $-3$ |
| $\Sigma X = 42$ | | $\Sigma x^2 = 42$ | $\Sigma Y = 49$ | | $\Sigma y^2 = 28$ | $\Sigma xy = -18$ |

Calculation of arithmetic mean of X series :

$$\overline{X} = \frac{\Sigma X}{N} \; ; \Sigma X = 42 \quad \text{and} \quad N = 7$$

$$\therefore \quad \overline{X} = \frac{42}{7}$$

$$\overline{X} = 6$$

Calculation of arithmetic mean of Y series :

$$\overline{Y} = \frac{\Sigma Y}{N} \; ; \Sigma Y = 49 \quad \text{and} \quad N = 7$$

$$\therefore \quad \overline{Y} = \frac{49}{7}$$

$$\overline{Y} = 7$$

Now $\Sigma xy = -18$, $\Sigma x^2 = 42$ and $\Sigma y^2 = 28$.

Calculation of regression coefficients :

Regression coefficient of X on Y

$$byx = \frac{\Sigma xy}{\Sigma y^2}$$

$$= \frac{-18}{42}$$

$$= -0.43$$

Thus, regression coefficient of X on Y $= -0.64$ and regression coefficient of Y on X $= -0.43$.

**(iii) If deviations are taken from assumed means :**

**Illustration - 3**

The following data relate to the experience of (10) machine operators and their performance ratings as given by the number of good parts turned out per 100 pieces.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Operator | : 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Experience | : 8 | 9 | 5 | 6 | 7 | 10 | 12 | 3 | 14 | 4 |
| Performance ratings | : 85 | 65 | 75 | 72 | 70 | 65 | 80 | 50 | 90 | 60 |

Calculate the regression coefficients of X on Y and Y on X.

**Solution :**

Let us take 10 and 70 as assumed means of X and Y series respectively.

**COMPUTATION OF REGRESSION COEFFICIENTS**

| Experience X | (X−10) dx | $dx^2$ | performance Ratings Y | (Y−70) dy | $dy^2$ | dxdy |
|---|---|---|---|---|---|---|
| 8 | −2 | 4 | 85 | 15 | 225 | −30 |
| 9 | −1 | 1 | 65 | −5 | 25 | 5 |
| 5 | −5 | 25 | 75 | 5 | 25 | 25 |
| 6 | −4 | 16 | 72 | 2 | 4 | −8 |
| 7 | −3 | 9 | 70 | 0 | 0 | 0 |
| 10 | 0 | 0 | 65 | −5 | 25 | 0 |
| 12 | 2 | 4 | 80 | 10 | 100 | 20 |
| 3 | −7 | 49 | 50 | 20 | 400 | 140 |
| 14 | 4 | 16 | 90 | 20 | 400 | 80 |
| 4 | −6 | 36 | 60 | 10 | 100 | 60 |
| $\Sigma$ dx = −22 | | $\Sigma dx^2 = 160$ | | $\Sigma$dy = 12 | $\Sigma dy^2 = 1304$ | $\Sigma$dxdy = 242 |

Regression coefficient of X on Y :

$$bxy = \frac{\Sigma dxdy - \frac{\Sigma dx.\Sigma dy}{N}}{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}$$

Here, $\Sigma dxdy = 242$, $\Sigma dx = -22$, $\Sigma dy = 12$ $\Sigma dy^2 = 1304$ and $N = 10$.

Substituting the values in the formula,

$$bxy = \frac{242 - \left(\frac{-22 \times 12}{10}\right)}{1304 - \frac{(12)^2}{10}}$$

$$= \frac{242 + \frac{264}{10}}{1304 - \frac{144}{10}}$$

$$= \frac{242 + 26.4}{1304 - 14.4}$$

$$bxy = 0.208$$

Regression coefficient of Y on X :

$$byx = \frac{\Sigma dxdy - \frac{\Sigma dx.\Sigma dy}{N}}{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}}$$

Here, $\Sigma dxdy = 242$, $\Sigma dx = -22$, $\Sigma dy = 12$, $\Sigma dx^2 = 160$ and $N = 10$.

Subatltuting the values in the formula,

$$byx = \frac{242 - \left(\frac{-22 \times 12}{10}\right)}{160 - \frac{(-22)^2}{10}}$$

$$= \frac{242 + \frac{264}{10}}{160 - \frac{484}{10}}$$

$$= \frac{242 + 26.4}{160 - 48.4}$$

$$= \frac{268.4}{111.6}$$

$$byx = 2.405$$

Thus, regression coefficient of X on Y = 0.208 and regression coefficient of Y on X = 2.405.

**(iv) If figures are given in original values**

**Illustration - 4**

The following calculations have been made for prices of ten stocks (X) on the Bombay Stock Exchange on a certain day along with the volume of sales in thousands of shares (Y). From the calculations, find the regression coefficient of prices of stock on the volume of sales of shares and regression coefficient of volume of sales of shares on prices of stocks.

$\Sigma X = 380$, $\Sigma Y = 170$, $\Sigma XY = 10150$,

$\Sigma X^2 = 31200$ $\Sigma Y^2 = 14200$

**Solution :**

Calculation of arithmetic mean of X series

$$\overline{X} = \frac{\Sigma X}{N}$$

$$= \frac{380}{10}$$

$$\overline{X} = 38$$

Calculation of arithmetic mean of Y series

$$\overline{Y} = \frac{\Sigma Y}{N}$$

$$= \frac{170}{10}$$

$$\overline{Y} = 17$$

Calculation of regression coefficient of X on Y :

$$bxy = \frac{\Sigma XY - N\overline{X}\overline{Y}}{\Sigma Y^2 - N(\overline{Y})^2}$$

Here, $\Sigma XY = 10{,}150$, $\overline{X} = 38$, $\overline{Y} = 17$, $\Sigma Y^2 = 14{,}200$ and N = 10.

Substituting the values in the formula,

$$bxy = \frac{10150 - 10 \times 38 \times 17}{142000 - 10(17)^2}$$

$$= \frac{10150 - 6460}{14200 - 2890}$$

$$= \frac{3690}{11310}$$

$$= 0.326$$

Calculation of regression coefficient of Y on X :

Stocks (X) :

$$bxy = \frac{\Sigma XY - N\overline{X}\overline{Y}}{\Sigma X^2 - N(\overline{X})^2}$$

Here, $\Sigma XY = 10150$, $\overline{X} = 38$, $\overline{Y} = 17$, $\Sigma X^2 = 31200$ and N = 10.

Substituting the values in the formula,

$$bxy = \frac{10150 - 10 \times 38 \times 17}{31200 - 10(38)^2}$$

$$= \frac{10150 - 6460}{31200 - 14440}$$

$$= \frac{3690}{16760}$$

$$= 0.22$$

Hence, regression coefficient of prices of stocks (X) on volume of sales of shares (Y) = 0.326 and regression coefficient of volume of sales of shares (Y) on prices of stocks (X) = 0.22.

## 25.11 COMPUTATION OF COEFFICIENT OF CORRELATION WITH THE HELP OF REGRESSION COEFFICIENTS

**Illustration - 5**

From some bivariate data, the following information is available :

Regression coefficient of X on Y = 0.4

Regression coefficient of Y on X = 0.9

Calculate the coefficient of correlation between X and Y.

**Solution :**

Coefficient of correlation (r) = $\sqrt{bxy.byx}$

We are given that bxy = 0.4 and byx = 0.9.

Hence = $\sqrt{0.4 \times 0.9}$

= $\sqrt{0.36}$

= 0.6

**Check Your Progress - 1**

In a distribution bxy = 0.45 and byx = 1.44.

Find out coefficient of determination.



**Illustration - 6**

Calculate regression coefficients and correlation coefficient from the following data :

X : 1 2 3 4 5 6 7 8 9 10

Y : 3 5 2 10 8 7 6 4 6 1

**Solution :**

Let us take 5 and 8 as assumed means of X and Y series respectively.

## CALCULATION OF REGRESSION COEFFICIENTS AND CORRELATION COEFFICIENT

| X | (X−5) dx | $dx^2$ | Y | (Y−8) dy | $dy^2$ | dxdy |
|---|---|---|---|---|---|---|
| 1 | −4 | 16 | 3 | −5 | 25 | 20 |
| 2 | −3 | 9 | 5 | −3 | 9 | 9 |
| 3 | −2 | 4 | 2 | −6 | 36 | 12 |
| 4 | −1 | 1 | 10 | 2 | 4 | −2 |
| 5 | 0 | 0 | 8 | 0 | 0 | 0 |
| 6 | 1 | 1 | 7 | −1 | 1 | −1 |
| 7 | 2 | 4 | 6 | −2 | 4 | −4 |
| 8 | 3 | 9 | 4 | −4 | 16 | −12 |
| 9 | 4 | 16 | 5 | −3 | 9 | −12 |
| 10 | 5 | 25 | 1 | −7 | 49 | −35 |
| | $\Sigma dx = 5$ | $\Sigma dx^2 = 85$ | | $\Sigma dy = -29$ | $\Sigma dy^2 = 153$ | $\Sigma dxdy = -25$ |

Regression coefficient of X on Y :

$$bxy = \frac{\Sigma dxdy - \frac{\Sigma dx.\Sigma dy}{N}}{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}$$

Here $\Sigma dxdy = -25$, $\Sigma dx = 5$, $\Sigma dy = -29$, $\Sigma dy^2 = 153$, and $N = 10$.

Substituting the values in the formula,

$$bxy = \frac{-25 - \frac{5 \times -29}{10}}{153 - \frac{(-29)^2}{10}}$$

$$= \frac{-25 + \frac{145}{10}}{153 - \frac{841}{10}}$$

$$= \frac{-25 + 14.5}{153 - 84.1}$$

$$= \frac{-10.5}{68.9}$$

$$= -0.152$$

Regression coefficient of X on Y $= -0.152$.

Regression coefficient of Y on X :

$$byx = \frac{\Sigma dxdy - \frac{\Sigma dx.\Sigma dy}{N}}{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}}$$

Here, $\Sigma dxdy = -25$, $\Sigma dx = 5$, $\Sigma dy = -29$, $\Sigma dx^2 = 85$, and $N = 10$.

Substituting the values in the formula,

$$= \frac{-25 - \frac{5 \times -29}{10}}{85 - \frac{(-5)^2}{10}}$$

$$= \frac{-25+\frac{129}{10}}{85-\frac{25}{10}}$$

$$= \frac{-25+14.5}{85-2.5}$$

$$= \frac{-10.5}{82.5}$$

$$= -0.127$$

$\therefore$ Regression coefficient of Y on X $= -0.127$.

Coefficient of correlation (r) $= \sqrt{bxy.byx}$

Here bxy $= -0.152$ and byx $= -0.127$.

Substituting the values in the formula,

$$= \sqrt{-0.152 \times -0.127}$$

$$= -\sqrt{0.019}$$

$$= -0.14$$

Coefficient of correlation $= -0.14$.

## 25.12 SUMMING UP

Regression analysis is a statistical technique with the help of which the values of an unknown variable are estimated on the basis of the known values of another variable. While correlation analysis helps to find out the mere presence or absence and the degree and direction of relationship between the variables, regression analysis helps to establish the functional relationship. This can be done with the help of regression lines. If the correlation coefficient between two variables is perfect, there will be only one regression line. The algebraic expression of regression lines is known as regression equations. Both the correlation and regression analysis are extremely useful to businessmen, government and consumers.

Regression co-efficient indicates the degree and the direction of change in the dependent variable in response to a unit change in the independent variable. Since there are two regression equations for two variables, there will be two regression co-efficients - one for the regression equation of X on Y and the other for regression equation of Y on X. While the regression co-efficient of X on Y indicates the degree and direction of change in 'X' variable in response to a unit change in 'Y' variable, the regression co-efficient of Y on X indicates the degree and direction of change in 'Y' variable in repsonse to a unit change in 'X' variable. Further, the square-root of the product of both the regression co-efficients is equal to the co-efficient of correlation between the variables.

## 25.13 CHECK YOUR PROGRESS : MODEL ANSWERS

1. $r^2 = \text{bxy} \times \text{byx}$

$= 0.45 \times 1.44$

$= 0.648$

## 25.14 MODEL EXAMINATION QUESTIONS

**A. Short Questions**

1. What is meant by 'regresssion'?
2. What are regression lines?
3. Why should there be two regression lines?
4. What are the regression equations?
5. Distinguish between 'Correlation' and 'Regression'.
6. Distinguish between simple and multiple regression.
7. What is the difference between linear and non-linear regression analysis?
8. Distinguish between total and partial regression.
9. What are the regression coefficients?
10. Write the normal equations used to compute the regression coefficients.
11. What are the properties of regression coefficients?

**B.Essay Questions**

12. What would be the lines of regression if (i) $r = +1$,

(ii) $r = -1$ and (iii) $r = 0$ ? Give your interpretation in each case.

13.Define 'regression' and discuss its utility.

EXERCISES

14. Calculate the two regression coefficients from the following data.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| X : | 30 | 40 | 75 | 60 | 50 | 42 | 70 | 72 |
| Y : | 40 | 25 | 35 | 40 | 65 | 52 | 60 | 35 |

(Ans : bxy = 0.048.. byx = 0.032)

15. Calculate the two regression coefficients and corelation coefficient from the following data.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| X : | 6.9 | 8.5 | 5.8 | 8.6 | 9.6 | 8.0 | 9.7 |
| Y : | 2.9 | 3.8 | 6.5 | 2.3 | 5.5 | 3.5 | 3.2 |

(Ans : bxy = 0.31., byx = 0.35; r = 0.33)

16. You are given that $\Sigma X = 190$, $\Sigma Y = 85$, $\Sigma XY = 575$, $\Sigma X^2 = 15600$ and $\Sigma Y^2 = 7100$.

Calculate the two regression equations. Also compute the coefficient of correlation.

(Ans : X on Y = X = 0.613 Y+17.16; Y on X = Y = 0.0867X+83.35)

## 25.15 RECOMMENDED BOOKS


1. Gupta, S.P. : "Statistical Methods", sultan chand & company, New Delhi.
2. Gupta, B.N. : "Statistics", Sahitya Bhavan, Agra.
3. Gupta, S.C. : "Fundamentals of statistics", Himalaya Pub. House, Bombay.
4. Simpson and Kafka : " Basic Statistics", Oxford and I.B.H. publishing Company, Calcutta.


## 25.16 GLOSSARY

| | | |
|---|---|---|
| 1 | Partial Regression | : partial Regression is a study of regression between two variables keeping the other variables constant |
| 2 | Regression Analysis | : It is a mathematical measure of the average relationship between two or more variables in terms of the original units of data. |
| 3 | Regression Coefficient | : It shows the degree and direction of change in the dependent variable in response to a unit change in the independent variable. |
| 4 | Regression Line | : It is a device used for estimating the value of one variable from the value of the other consists of a line through the points drawn in such a manner as to represent the average relationship between the two variables. |
| 5 | Simple Regression | : This studies regression between two variables only. |
| 6 | Total Regression | : It is a study of regression of all variables that affect the problems under consideration. |

## BLOCK V : INDEX NUMBERS

## UNIT-26 : INDEX NUMBERS

**Contents**





## 26.0 AIMS AND OBJECTIVES

This unit aims at explaining the meaning, characteristics, utilities and kinds of Index Numbers.

At the end of the unit, you should be able to :

- Define the term 'Index Numbers'.
- list the characteristics of Index Numbers
- explain the utility of Index Numbers
- identify the kinds of Index Numbers

## 26.1 INTRODUCTION

In our daily life, we often make judegement by summarising and comparing changes in an economic variable with time or place. We also see the headlines of newpapers with regard to increase or decrease of prices, rise and fall of indsutrial production, increase and decrease of imports and exports and rising of crime in a particular period as compared to its earlier period. All these changes in different variables over a period of time or place are indicated by the index numbers only.

Index numbers are specialised devices for measuring variations in magnitude of a group of related variables. Index numbers are generally used to measure the changes in prices of commodities, volume of production, national income, wages, imports and exports etc. By using index numbers, one can compare and analyse the cost of living at different times or in different countries or locations, the physical volume of production in different years, efficiency of different institutions, etc. For example, to compare the price level in India during 1983, with what it was in 1980, we shall have to take the prices of certain commodities such as wheat, rice, clothing, oil, house rent, etc., for the years 1983 on the basis of 1980 , we can draw useful inferences in price level changes. Index numbers are used to feel the pulse of the economy and they have come to be used as indicators of inflationary and deflationary tendencies. These are also described as "Barometers of Economic Activity". The purpose of index numbers is to show the magnitude of variations in one figure which are not susceptible to direct measurement or observations.

## 26.2 DEFINITION OF INDEX NUMBERS

Index numbers are statistical devices designed to measure the relative change in the level of variables with repsect to time, geographical locations or other characteristics such as income, profits, production, sales etc. The following are some of the important definitions of index numbers.

In the words of Morris Hamburg, "In its simplest form an index number is nothing more than a relative number or a 'relative' which expresses the relationship between two figures, where one of the figures is used as a base."

Index numbers have been defined by Croxton and Cowden as "Devices for measuring differences in the magnitude of a group of related variables".

According to Horace Secrist, "Index numbers are a series of numbers by which changes in the magnitude of a phenomenon are measured from time to time, place to place".

In the words of Spiegel, "An index number is a statistical measure designed to show changes in variables or group of related variables with respect to time, geographic location or other characteristics".

According to Patterson, "An index number is a statistical measure designed to show changes in one variable or in a group of related variables over time or with respect to geographic loation, or other characteristics".

John I.Griffin writes, "An index number is a quantity which by reference to a base period, shows by its variations the changes in the magnitude over a period of time. In general, Index numbers are used to measure changes over time in magnitude which are not capable of direct measurement".

It is evident from the above definitions that an index number is a specialised average desgined to measure the changes in a group of related variables over a period of time. For example, if the price index is 150 for the year 1983 as compared to 1980 , it indicates that the net increase in the prices of commodities is to the extent of 50 per cent in 1983 as compared to 1980.

## 26.3 CHARACTERISTICS OF INDEX NUMBERS

The following are the characteristics of index numbers:

**(I) Index numbers are specialised averages**

Index numbers measure the relative changes in group of related items with reference to some base period. According to L.R.Connor, "In its simplest form, it (index number) represents a special case of an average, generally a weighted average compiled from a sample of items judged to be representative of the whole". Hence index numbers are specialised averages capable of averaging different units of measurement. For example, to construct consumer price index, the various items, such as food, clothing, fuel and lighting, house rent and others are measured in different types of units, such as kilograms or quintals, metres, litres, number of rooms, etc. Since all these items are not directly measurable and comparable,index numbers can be used to measure the different variations of commodities in a single figure by averaging and comparing different units of measurement in a common relative variable.

**(ii) Index numbers measure net change in a group of related variables**

Index numbers as a measuring technique are capable of measuring changes that occur in a group of related variables under reivew. These groups of variables may be prices of a set of commodities, the volume of production, imports and exports, etc. For example, the consumer price index of a working class of Hyderabad has increased to 140 in the year 1983, when compared to 1980. This reflects the net increase of 40 per cent in the commodities of 1983 over 1980. Similarly, index numbers also measure net changes in industrial production, sales, profits, etc.

**(iii) Index numbers measure the effect of changes over a period of time**

The techniques of index numbers are most widely used for measuring changes over a period of time. For instance, one can find out through index numbers, the changes in the agriculture production from the begining of the Third Plan period to the end of the Fourth Plan period, i.e., 1961 to 1974. Similarly, we can compare wholesale prices of commodites, exports, imports, wages, industrial production,etc, over a long period of time. At this point, the index numbers are not only applicable to measure the net changes over a long period of time, but also useful to compare the economic conditions of different locations, different industries, different cities or different countries.

**(iv) Index numbers are expressed in percentages**

Index numbers are devices to measure the net changes occurred in a related variable not in absolute terms, but in relative terms. The values of different commodities are expressed in absolute terms which are not directly comparable, as they are in different units of measurement. Therefore, the values of all these variables are brought into a common comparable relative term, viz., percentages for enabling us to show the extent of change. Though the values of index numbers are expressed in percentages, the same is not actually shown in any index number value.

**(v) Index numbers measure changes not capable of direct measurement**

Index numbers have some special characteristics to measure the changes in magnitude which are not capable of direct measurement. For example, magnitude of price level, cost of living, business or economic activity, etc., are not directly measurable in the original amount of change. Therefore index numbers as specialised averages are capable of measuring the magnitude of change in related variables in a meaningful manner without any bias.

**Check your progress - 1**

List out the charactersitics of Index Numbers.

## 26.4. UTILITY OF INDEX NUMBERS

Index numbers are useful in the study and analysis of economic activity of any economy irrespective of political and social structure engaged in production, distribution and consumption of goods and services. In order to understand the progress in the economy, all economic activities are aggregated averaged and approximated with a convenient device. The index numbers are proved to be very useful in this process. Economists generally make use of various indices to appraise the prerformance of the economy and to analyse its structure and behaviour. For example to know the state of economic activity in a country, the indices of industrial and agricultural production, stock market prices, wholesale prices, consumer prices, imports and exports, incomes of various types and so on, can be used. Index numbers are also used in connection with decision making and analysis in business and Government . For instance, the consumer price indices are used in the determination of wage negotiations and dearness allowances.

According to Simpson and Kafka, "Index numbers are today one of the most widely used statistical devices. They are used to feel the pulse of the economy and they have come to be used as indicators of inflationary and deflationary tendencies".

Index numbers are significantly used to measure and compare the changes of prices and purchasing power of money, and they are also useful to study the changes in other variables such as business activity,employment, industrial and agricultural production. The following are the important uses of index numbers.

**(i) index numbers measure the change in values**

The primary utility of index numbers is to measure changes in related variables or a set of variables. This facilitates to measure the changes in variables or a set of variables. This facilitates to measure the changes in variables from time to time when the variables are expressed in different physical units such as kilograms, metres, litres, etc. One can make use the techniques of index numbers to measure the changes in common comparable terms, because the absolute amount of magnitude of change in different variables not directly measurable and comparable.

Therefore, the index numbers are useful to express the changes in the variables in common neausre i.e., percentages.

**(ii) index numbers measure purchasing power of money**

Index numbers are useful to measure the purchasing power of money. They are helpful :o adjust the original data for price level changes to get real values. In order to understand :he purchasing ,power of rupee earnings, the nomina! earning are to be converted by applying leflating technique into real earnings as the purchasing power of rupee is constantly changing. For example, the value of rupee in 1983 is only 18 paise as compared to the value of rupee in 1960. All the rupee earnings are in current values, but their real values may be significantly ower due to changes in the purchasing power of rupee. The money incomes are to be converted nto real incomes with the help of the following deflating formula.

$$Realwages = \frac{Money\ wages}{Price\ index} \times 100$$

The study of real wages as compared to money wages is more meaningful and helpful for naking adjustments in wages and salaries of employees.

For example, a worker's wage in 1983 is Rs. 300/- whereas the price index is 150. Then the eal wage of the worker will be :

$$\frac{300}{150} \times 100 = 200$$

Though the salary of the worker is Rs. 300/- , yet the purchasing power is only Rs. 200/-

**iii) index numbers measure and compare changes**

One of the main purposes of index numbers is to measure changes in related variables ind compare this with some base figure to draw inferences with regard to the percentage of :hange over a period of time. This also facilitates comparison of changes from time to time among different places, which are expressed in different absolute units. The changes in price evel, cost of living, etc., are not directly measured and compared without taking the help of index numbers. For example, if a construction contractor wants to know the extent of increase in the cost of construction in the current year compared to the previous year, it can be found out by measuring the price changes in construction materials like steel, cement, labour, wood, brick, etc. But the prices of these items are not directly measurable and comparable, owing to their different units of measurement. If he uses the technique of index numbers he can easily measure and compare the changes in the cost of construction. The usefulness of index numbers n this regard is significant due to the following reasons:

(a) Index numbers enable us to process complex and mass data into simple numbers to reflect the relative change through time or space.

(b) The relative changes measured at various points in time and space can be easily compared for drawing useful inferences.

(c) Index numbers enable the comparison of dissimilar units by bringing them on to a single comparable values.

**(iv) Index numbers study the trends and tendencies**

Index numbers are used to study the price level changes at different periods of time. If a series of index numbers are calculated for a variable over a period of time, say for 8 to 12 years, they reveal a pattern of increasing or decreasing tendencies. Indices relating to output, volume of trade, imports and exports, etc., are useful in studying the changes in the phenomenon due to the influecnce of the components of time series, such as, trend, seasonal, cyclical and irregular variations. This reflects a general trend of production and business activity. By observing the general trend of the pattern of indices, one can draw the conclusions regarding the amount of change that is taking place due to various components of time series.

**(v) Index numbers help in the formulation of policies**

The policies relating to economic and business matters are guided by index numbers, because they measure relative changes over a time and place. They are guides to business and economic policy. Index numbers of data relating to prices, production, profits, imports and exports, personnel and financial matters are significant for any organisation for efficient planning and formulation of policies and executive decisions. Apart from this, index numbers are very widely used for studying general economic and business conditions. They are also useful to sociologists in studying population changes; for psychologist to measure "intelligence quotients" ; for health authorities to show adequacy of hospital facilities and educational research organisations to study the effectiveness of school system through appropriate indices.

The cost of living index numbers are used by industrial and business concerns and government for determining the dearness allowances of their employees to meet the rise in the cost of living from time to time. The composition of excise duty by the government is being adjusted from time to time on the basis of the index number of various commodities. Besides this, the indices of different commodities are significantly helpful to understand the past behaviour of the variables and to plan for the future production. Quantity indices like indices of industrial production give a measure of relative change in physical output. Different types of index numbers are prepared by the Government to know the changes in the economy in general and national income in particular to have basic guidance for the future economic policy.

**(vi) Index numbers are economic barometers**

Index numbers are rightly called economic barometers because they measure the pressure of economic and business behaviour like barometers used in physics to measure the pressure of atmosphere. Index numbers are useful to measure the general economic behaviour of a country. Indicies such as prices -wholesale and retail - output ,volume of trade, imports and exports, agricultural and industrial production, bank deposits and foreign exchange reserves throw light on the nature and variations of general economic and business activity of the country. A careful study of series of index numbers pertaining to differnt aspects of economy reflect the general

$= 9$

Calculation of Arithemetic Mean of Y series:

$$\overline{Y} = \frac{\Sigma Y}{N}$$

$$= \frac{130}{10}$$

$$= 13$$

Calculation of Karl Pearson's coefficient of correlation:

$$= \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}}$$

Here, $\Sigma xy = -156$, $\Sigma x^2 = 126$, $\Sigma y^2 = 204$

Substituting the values in the formula,

$$r = \frac{-156}{\sqrt{126 \times 204}}$$

$$= \frac{-156}{\sqrt{25704}}$$

$$= \frac{-156}{160.33}$$

$$r = -0.97$$

Thus, there is a high degree of negative correlation between price demand for the commodity.

**Illustration - 5**

From the following data calculate Pearsonian coefficient of correlation between X and Y:

| | 'X' Series | 'Y' Series |
|---|---|---|
| No. of pairs observed | 45 | 45 |
| Arithmetic mean | 75 | 54 |
| Sum of squares of deviations taken form arithemetic mean | 408 | 414 |
| Sum of products of deviations of 'X' and 'Y' series from their respective arithmetic means | 122 | |

**Solution :**

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}}$$

We are given that

N = 45, $\overline{X} = 75$, $\overline{Y} = 54$, $\Sigma x^2 = 408$, $\Sigma y^2 = 414$, $\Sigma xy = 122$.

Substituting the values in the formula,

$$r = \frac{122}{\sqrt{408 \times 414}}$$

$$= \frac{122}{\sqrt{168912}}$$

$$= \frac{122}{410.99}$$

$$r = 0.297$$

Thus, Karl Pearson's coefficient of correlation between X and Y = 0.297.

***Short-cut Method***

When actual means are in fractions the computation of correlation by direct method involves tedious calculations and would take a lot of time. In such cases, deviations taken from assumed means would minimise the calculations and thus save time. This method of finding out correlation is called short-cut method and is computed with the help of the following steps:

(i) Take the deviations of X series from an assumed mean, denote these deviations by $dx$ and find the total $\Sigma dx$.

(ii) Take the deviations of Y series from an assumed mean, denote these deviations by $dy$ and fond the total $\Sigma dy$.

(iii) Square $dx$ and find the total $(\Sigma dx^2)$.

(iv) Square $dy$ and find the total $(\Sigma dy^2)$.

(v) Multiply $dx$ and $dy$ and find the total $\Sigma dx\,dy$

(vi) Substitute the values of $\Sigma dx\,dy$, $\Sigma dx$, $\Sigma dy$, $\Sigma dx^2$ and $\Sigma dy^2$ in the following formula and find the correlation coefficinet:

$$r = \frac{N.\Sigma dx\,dy - (\Sigma dx)(\Sigma dy)}{\sqrt{N.\Sigma dx^2 - (\Sigma dx)^2}\sqrt{N.\Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{\Sigma dx\,dy - \frac{\Sigma dx.\Sigma dy}{N}}{\sqrt{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}}\sqrt{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}}$$

Where $\Sigma dx\,dy$ = Sum of the product of the deviations of X and Y series taken form their assumed means.

$\Sigma dx^2$ = Sum of the squares of deviations of 'X' series from its assumed mean.

$\Sigma dy^2$ = Sum of the squares of deviations of 'Y' series from its assumed mean.

$\Sigma dx$ and $\Sigma dy$ refer to sum of the deviations of X and Y series from their respective assumed means.

Illustration - 6

Calculate Karl Pearson's coefficient of correlation from the following data :

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| X | 25 | 30 | 52 | 40 | 20 | 22 | 32 | 35 | 38 | 42 |
| Y | 12 | 22 | 15 | 20 | 18 | 18 | 25 | 23 | 30 | 30 |

**Solution:**

**CALCULATION OF KARL PEARSON'S COEFFICIENT OF CORRELATION**

Let us assume that 35 and 25 are the means of X and Y series respectively.

| X | X−35 $dx$ | $dx^2$ | Y | Y−25 $dy$ | $dy^2$ | $dxdy$ |
|---|---|---|---|---|---|---|
| 25 | −10 | 100 | 12 | −13 | 169 | 130 |
| 30 | − 5 | 25 | 22 | − 3 | 9 | 15 |
| 52 | 17 | 289 | 15 | −10 | 100 | −170 |
| 40 | 5 | 25 | 20 | − 5 | 25 | − 25 |
| 20 | −15 | 225 | 18 | − 7 | 49 | 105 |
| 22 | −13 | 169 | 18 | − 7 | 49 | 91 |
| 32 | − 3 | 9 | 25 | 0 | 0 | 0 |
| 35 | 0 | 0 | 23 | − 2 | 4 | 0 |
| 38 | 3 | 9 | 30 | 5 | 25 | 15 |
| 42 | 7 | 49 | 30 | 5 | 25 | 35 |
| $\Sigma dx = -14$ | | $\Sigma dx^2 = 900$ | | $\Sigma dy = -37$ | $\Sigma dy^2 = 455$ | $\Sigma dxdy = 196$ |

$$r = \frac{\Sigma dxdy - \frac{\Sigma dx.\Sigma dy}{N}}{\sqrt{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}}\sqrt{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}}$$

Here, $\Sigma dxdy = 196$, $\Sigma dx = -14$, $\Sigma dy = -37$, $\Sigma dx^2 = 900$, $\Sigma dy^2 = 455$ and N = 10.

Substituting the values in the formula,

$$r = \frac{196 - \frac{(-14 \times -37)}{10}}{\sqrt{900 - \frac{(-14)^2}{10} \times 455 - \frac{(-37)^2}{10}}}$$

$$r = \frac{196 - \frac{518}{10}}{\sqrt{900 - \frac{196}{10} \times 455 - \frac{1369}{10}}}$$

$$= \frac{196 - 51\ 8}{\sqrt{900 - 19.6 \times 455 - 136.9}}$$

$$= \frac{144.2}{\sqrt{880.4 \times 318.1}}$$

$$= \frac{144.2}{\sqrt{280055.24}}$$

$$= \frac{144.2}{529.20}$$

$$r = 0.27$$

**Check your progress - 1**

In a distribution N = 10, $\Sigma dx^2 = 3775$. $\Sigma dx = 85$, $\Sigma dy^2 = 3125$, $\Sigma dy = 35$, and $\Sigma dxdy = 2900$.

Find out coefficient of correlation

*Calculation of coefficient of correlation directly from original values*

According to this method, we need not take the deviations of both the series either from their actual means or from their respective assumed means.

In this case, the following formula is applied.

$$r = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2 - (\Sigma X)^2}\sqrt{N.\Sigma Y^2 - (\Sigma Y)^2}}$$

Where,

$\Sigma X$ = Total of X series

$\Sigma Y$ = Total of Y series

$\Sigma X^2$ = Sum of the squares of original values of X series

$\Sigma Y^2$ = Sum of the squares of original values of Y sereis

$\Sigma XY$ = Sum of the products of values of X and Y series

N = Number of pairs observed

**Illustration - 7**

Calculate correlation from the following data:

X : 10 9 8 7 6 5 4 3 2 1

Y : 25 23 20 16 12 12 10 8 5 4

**Solution :**

CALCULATION OF COEFFICIENT OF CORRELATION

| X | $X^2$ | Y | $Y^2$ | XY |
|---|---|---|---|---|
| 10 | 100 | 25 | 625 | 250 |
| 9 | 81 | 23 | 529 | 207 |
| 8 | 64 | 20 | 400 | 160 |
| 7 | 49 | 16 | 256 | 112 |
| 6 | 36 | 12 | 144 | 72 |
| 5 | 25 | 12 | 144 | 60 |
| 4 | 16 | 10 | 100 | 40 |
| 3 | 9 | 8 | 64 | 24 |
| 2 | 4 | 5 | 25 | 10 |
| 1 | 1 | 4 | 16 | 4 |
| $\Sigma X = 55$ | $\Sigma X^2 = 385$ | $\Sigma Y = 135$ | $\Sigma Y^2 = 2303$ | $\Sigma XY = 939$ |

$$r = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2 - (\Sigma X)^2}\sqrt{N.\Sigma Y^2 - (\Sigma Y)^2}}$$

Here, $N = 10$, $\Sigma XY = 939$, $\Sigma X = 55$, $\Sigma Y = 135$, $\Sigma X^2 = 385$ and $\Sigma Y^2 = 2303$.

Substituting the values in the formula,

$$r = \frac{10\times939-55\times135}{\sqrt{10\times385-(55)^2}\sqrt{10\times2303-(135)^2}}$$

$$= \frac{9390-7425}{\sqrt{(3850-3025)}\sqrt{(23030-18225)}}$$

$$= \frac{1965}{\sqrt{825\times4805}}$$

$$= \frac{1965}{\sqrt{3964125}}$$

$$= \frac{1965}{1991.011}$$

$$r = 0.987$$

Thus, there is a high degree of positive correlation between X and Y series.

**Illustration - 8**

Calculate coefficient of correlation from the following data:

Number of pairs of observations = 10

$\overline{X} = 33.6$, $\overline{Y} = 21.3$, $\sigma x = 12$, $\sigma y = 8$

Assumed mean of 'X' series = 40

Assumed mean of 'Y' series = 25

Sum of products of deviations of X and Y series from their respective assumed means = 65.

**Solution :**

When deviations are taken from assumed means, we can also apply another formula, i.e.,

$$r = \frac{\Sigma dxdy - N(\overline{X}-Ax)(\overline{Y}-Ay)}{N.\sigma x.\sigma y}$$

Where

$\Sigma dxdy$ = Sum of plroducts of deviations taken from assumed means of X and Y series

$Ax$ = Assumed mean of X series

$Ay$ = Assumed mean of Y series

We are given that N = 10, $\overline{X} = 33.6$, $\overline{Y} = 21.3$, $\sigma x = 12$, $\sigma y = 8$, $Ax = 40$, $Ay = 25$ and $\Sigma dxdy = 65$.

Substituting the values in the formula, we get,

$$r = \frac{65-10(33.6-40)(21.3-25)}{10\times12\times8}$$

$$= \frac{65-10(-6.4)(-3.7)}{960}$$

$$= \frac{65-236.8}{960}$$

$$= \frac{-171.8}{960}$$

$$r = -0.179$$

Thus, Karl Pearson's coefficient of correlation $= -0.179$.

**Assumptions of the Pearsonian Coefficient of Correlation**

Karl Pearson's coefficient of correlation is based on the following assumptions :

(i) There is a linear relationship between the variables studied.

(ii) The variables under study are affectd by a large number of independent causes.

(iii) There is a 'cause and effect' relationship between the variables under study.

**Merits and Limitations of the Pearsonian Coefficient of Correlation**

Karl Pearson's Correlation Coeffcient is a mathematical method of finding out the nature and extent of relationship in exact numerical terms. The coefficient obtained is a pure number which is not affected by different units of measure in which the values of the variables are expressed. Thus, Karl Pearson's method measures the relationship in a single figure which also reveals the direction and degree of correlation. However, it suffers from the following limitaions.

(i) It assumes that the data contain linear relationship, but linearity is a rare phenomenon in our practical life.

(ii) Interpretation of correlation coefficient needs greater amount of care and experience.

(iii) The value of the correlation coefficient is subject to the influence of extreme value of the items of data.

(iv) It involves tedious mathematical calculations.

## 23.5 SUMMING UP

Correlation is measured with the help of coefficient of correlation which always varies in between $\pm$ 1. While $+1$ indicates the perfect positive correlation, $-1$ indicates the perefect negative correlation. On the other hand, 'O' (zero) indicates the absence of correlation between two variables. Correlation is studied with the help of the following methods:

(i) Scatter diagram method

(ii) Graphic method

(iii) Karl Pearson's coefficient of correlation

(iv) Spearman's rank coefficient of correlation

(v) Co-efficient of concurrent deviations

## 23.6 CHECK YOUR PROGRESS : MODEL ANSWERS

1. The following formula is used

$$r = \frac{\Sigma dx dy - \frac{\Sigma dx . \Sigma dy}{N}}{\sqrt{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}} \sqrt{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}}$$

The answer i 478.

## 23.7 MODEL EXAMINATION QUESTIONS.

### A. SHORT QUESTIONS

1. What is a scatter diagram?
2. What are the limits of the value of $r$ ?
3. How is correlation measured with the help of graphic method?
4. How do you interpret the Karl Pearson's Coefficient of Correlation?

### EXERCISES

5. From the following data ascertain whether the income and expenditure of 100 workers of a factory are correlated with the help of graph.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Year | : 1978 | 1979 | 1980 | 1981 | 1982 | 1983 | 1984 |
| Average Income(Rs.) | : 400 | 450 | 500 | 525 | 510 | 475 | 550 |
| Average Expenditure(Rs) | : 150 | 125 | 200 | 175 | 250 | 300 | 400 |

(Ans: r = 0.703)

6. Calculate Karl Pearson's Coefficient of correlation for the following series :

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Price (X) Rs. | : 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Demand (Y) Tonnes | : 30 | 29 | 29 | 25 | 24 | 24 | 24 | 21 | 18 | 16 |

(Ans: r = 0.967)

7. Calculate Pearsonian coefficient of correlation from the following data:

| Age | Blind persons per lakh |
|---|---|
| 0-5 | 5 |
| 5-10 | 6 |
| 10-15 | 9 |
| 15-20 | 12 |
| 20-25 | 10 |
| 25-30 | 4 |
| 30-35 | 15 |

(Ans: r = 0.522)

8. From the following information compute coefficient of correlation:

| | X Series | Y Series |
|---|---|---|
| Arithmetic Mean | 37.25 | 62.75 |
| Assumed Mean | 34.50 | 56.00 |
| Standard Deviation | 16.53 | 17.93 |

Sum of products of deviations of X and Y series taken from their respective assumed means = 1038.

Number of pairs of observations of X and Y series = 10.

(Ans: r = 0.287)

9. From the following information calculate coefficient of correlation:

| | |
|---|---|
| Total of the deviations of X series | = −85 |
| Total of the deviations of Y series | = −10 |
| Total of the product of deviations of X and Y series | = 1522 |
| Total of the squares of deviations of X series | = 4144 |
| Total of the squares of deviations of Y series | = 1132 |
| Number of pairs of observations | = 10 |

Assumed means of X and Y series are 41 and 35 respectively

(Ans: r = 0.733)

10. Karl Pearson's coefficient of correlation between X and Y is 0.4, their covariance is +8. If the variance of X is 12, find the standard deviation of Y series.

(Ans: $\sigma Y = 5.77$)

11. Calculate the coefficient of correlation between X and Y series from the following data:

| | X Series | Y Series |
|---|---|---|
| Number of pairs observed | 10 | 10 |
| Standard Deviation | 6 | 4 |
| Sum of products of deviation of X and Y series taken from their respective arithmetic means | 135 | |

(Ans: r = 0.562)

12. If covariance between 'X' and 'Y' series is 4.2 and the variance of X and Y are 15.1 and 13.2 respectively, find the coefficient of correlation between X and Y.

(Ans: r = 0.297)

13. Find Karl Pearson's coefficient of correlation from the following data in respect of

tendencies of economic development..

## 26.5 KINDS OF INDEX NUMBERS

On the basis of the study of relative changes in different variables, index numbers are divided into the following three categories:

26.5.1 Price Index Numbers

26.5.2 Quantity index Numbers and

26.5.3 Value Index Numbers.

## 26.5.1 PRICE INDEX NUMBERS

Among all the Index numbers, price index numbers are commonly used in economic and business fields to measure the relative price level changes of commodities at some time or at certain place with reference to some base period. These are also useful to study the price level changes of shares, debentures, etc. Price index numbers are further divided into two types.

(a) Wholesale Price Index Numbers and

(b) Retail Price Index Numbers

**(a) Whole sale Price Index Numbers :** The general price level change in a country studied by using wholesale price index numbers. The first wholesale general price index number was calculated in India for the year 1947, based on 1939 prices. The new series of index numbers were computed in India on the basis of 1961 - 62 prices with the recommendations of "Wholesale Price Index Revision Committee". It covered 139 commoditeis, 255 markets and 774 quotations. The latest wholesale price index numbers in India are constructed with 1970-71 as base year.

**(b) Retail Price Index Numbers :** Retail Price Index numbers are useful to measure general changes in retail prices of various commodities such as consumption goods, shares, bank deposits, bonds etc.

Consumer price index or cost of living index is the specialised kind of retail price index. It enables us to study the price level changes of a basket of goods or purchasing power of rupee or cost of living of a particular section of people, like labourers, agricultural workers, etc. In India, the cost of living indices are calculated by studying the cost of living of (i) Central and State Government employees, (ii) Middle class people and (iii) Working classes.

## 26.5.2 QUANTITY INDEX NUMBERS

These index numbers help to study the changes in the volume of goods produced, purchased or sold, consumed and distributed during a particular period of time as compared to its base period. In these indices quantities are considered prominent along with the prices of different commodities. Indices of agricultural and industrial production, imports and exports, etc, are examples of quantity index numbers. They are widely used to study the level of physical output in an economy.

## 26.5.3 VALUE INDEX NUMBERS

These are meant to measure the changes in the total value of commodities. The product of price and quantity is known as value of a commodity. The value indices such as retail sales, profits and inventories are pr... nent among value indices. However, these are not commonly used like price and quantity index numbers.

## 26.6 SUMMING UP

Index numbers are specialised averages to measure the changes in the prices commodities, volume of production, national income, wages, imports, exports, etc. They are used for analysis and prediction of future values. Thus, index numbers are useful in the study and analysis of economic activity of any economy irrespective of political and social structure. Index numbers are useful to sociologists in studying population changes. The index numbers are classified into price index numbers, quantity index numbers, value index numbers, etc.,

## 26.7 CHECK YOUR PROGRESS:MODEL ANSWERS

1.

- They are specialised averages
- They measure the net change in a group of related variables.
- They measure the effect of changes over a period of time.
- They are expressed in percentages
- They measure the changes which are not capable of direct measurement.

## 26.8 MODEL EXAMINATION QUESTIONS

**A.Short Questions**

1.What is an index number?

2. What is a Price Index Number?

3. What is a value Index Number?

4. What is a Quantity Index Number?

**B. Essay Questions**

5. 'Index numbers are devices for measuring differences in the magnitude of a related variable'. Discuss the statement and point out the uses of index numbers.

6. What is the importance of index numbers in economic and commercial studies?

## 26.9 RECOMMENDED BOOKS


1. Gupta, S.P. : "Statistical Methods", Sultan Chand & Company, New Delhi.
2. Gupta,B.N. : "Statistics", Sahitya Bhavan, Agra.
3. Gupta, S.C. : "Fundamentals of Statistics", Himalaya Publishing House, Bombay.
4. Simpson and Kafka : "Basic Statistics", Oxford and I.B.H. Publishing company, Calcutta.


## 26.10 GLOSSARY

| | | |
|---|---|---|
| 1 | Base Year | : The year selected for making comparison. |
| 2 | Current Year | : The year which requires comparison. |
| 3 | Index Numbers | : Index numbers are specialised averages used to measure the change in a group of related variables over a period of time. |
| 4 | Price Index numbers | : These are used to measure the relative price level changes of commodities at two different time periods. |
| 5 | Quantity Index Numbers | : They study the changes in the quantity of goods produced, consumed or distributed at two different time periods. |
| 6 | Value Index Numbers | : They study the changes in the total value of production of two different time periods. |

# UNIT - 27 CONSTRUCTION OF INDEX NUMBERS

**Contents**



## 27.0 AIMS AND OBJECTIVES

This unit aims at explaining the problems faced in the construction of index Numbers.

On completion of this unit, you should be able to :

– identify the various problems involved in the construction of index Numbers.

## 27.1 INTRODUCTION

Index numbers are powerful statistical devices of measuring and comparing the changes in the value of different types of commodities over two different periods. In order to fulfil this aim, it is necessary to take utmost care and precaution while constructing indices so as to get more accurate values to facilitate further analysis and predictions. If index numbers are computed improperly, they give rise to fallacious conclusions and prove to be improper devices.



## 27.2 PROBLEMS IN THE CONSTRUCTIN OF INDEX NUMBERS

The following are some of the problems which are generally faced while constructing the index numbers.

**1. Defining the purpose**

The first and foremost problem faced while constructing index number is to define the objective and purpose in clear and correct terms. Any single index number is not used for all purposes. Every index number has got its own uses and purposes. Hence, the determination of purpose is the basic and deciding factor. In the construction of an index number, the other related aspects such as the nature of data to be collected , the statistical techniques to be used, the selection of commodities, the selection of base period, the selection of the average and so on are determined on the basis of the purpose.

According to Croxton and Cowden, "An index number properly designed for the purpose

in hand is a most useful and powerful tool; if not properly compiled and constructed, it can be dangerous one". For example, if we wish to study the cost of living of a particulart class of people, we have to decide the class of people whose the index is to be computed, because the consumption pattern of the commodities differs from people to people. If the cost of living index relates to poor class of people, then the rise in the prices of luxury items will not effect the cost of living of poor people, but will definitely affect the total expenditure of rich class of people. The objective of index number also influences the number and types of commodities to be taken. If the objective of constructing an index number is to study the general changes in the price level in the country, the price quotations are to be obtained from the wholesale market and a large number of commodities should be included. On the other hand, if the objective of an index number is to study the cost of living of a particular class of people, the price quotations are to be obtained from the retailed shops and the number of commodities to be included may be relatively smaller. Unless the objective of the index number is clearly defined, the data obtained may not be appropriate to compute the desired index number.

**2.Selection of base period**

The base period is the period of comparison for the relative change in the level of commodities from time to time. While computing index numbers for various years, the base period index number is generally taken as 100. The base period may be a year a month or a day.

The selection of a base period primarily depends on the purpose of the index number. However, the following points are significant to select an appropriate base period.

*(i) Base period chosen should be a period of normal and stable economic conditions*

Generally, the period chosen for the base should be normal and it should not be effected by abnormalities and irregular forces such as, earth-quakes, floods, famines, strikes, lock-outs, trade-cycles, etc. If we select a particular year as base and it is a period of economic boom, the prices of various commodities will be abnormally high or on the other hand , if the year is a period of depression or economic instability, the prices of commodities will be abnormally low. Thus, the index computed by taking such an abnormal year as a base may be over-stated or under-stated. However, the selection of a normal year as a base is a difficult job. If we select a year by assuming it as a normal year it may be normal in one respect and may be abnormal in some other respect. In order to overcome this problem, it is always better to take the average of 3 to 4 years as a base period, since the process of averaging will balance the effect of abnormalities in that period.

*(ii) The base period should not be too distant from the current period*

The base period should not be too distant from the current period, as it makes the short-term comparisons irrelevant on account of changes in the habits, customs, tastes and consumption pattern of the customers. If the time gap is longer between the current year and base year, it may influence the values of various commodities to a great extent. Apart from this the commmodities

used in the base year may not be continued in the current year due to obsolete or out-dated or non-availability of commodities and they may be replaced by new commodites with better quality.

*(iii) Choice of Base period*

While selecting the base period, one has to decide whether to have a fixed base or a chain base. In the fixed base indices the base or reference period will be fixed and the indices for different periods are computed on the basis of the prices of single base period . For instance, if the indices for 1981, 1982 and 1983 are computed with 1980 as base year, the resulting indices are knwon as fixed base indices.

On the other hand, the chain base indices are computed by relating each year's price to that of the immediately preceding year. For example, the index for 1981 is computed with 1980 as base, the index for 1982 with 1981 as base, for 1983 with 1982 as base and so on. As compared to the fixed base indices, the chain base indices reflect a better pattern of the recent fluctuations in the prices of commodities. But the indices computed on fixed base facilitate the comparison of prices of any given year with the earlier prices.

**3.Selection of items**

The next problem of computing index number is the selection of items. The selection of items depends mainly on the purpose of the index number. While selecting commodities one has to see that the seleted items should be representative of the tastes, habits and customs of the people for whom the index is computed. Further , it is necessary that the items selected should be relevant for the purpose . For instance to calculate the cost of living index number of a low income group or poor people, care must be taken in selecting the items which are generally used by the people belonging to that group. It is necessary to avoid those commodities which are used by middle or higher income group people. In this instances, selection of high quality commodities and luxury items like scooters, television sets, refregirators etc., are to be avoided. A decision must be taken on the number of commodities to be included and their quantities.If a larger number of items are included in a sample, the index number will be more representative, but at the same time, it involves more time and cost. Thus, the purpose of the index number will play a significant role in selecting the number of items. While selecting the items, the following important points must be kept in mind.

(a) Classification of items into relatively homogeneous sub-groups, such as:

(i) Food, cereals, rice, wheat, pulses, grams, etc.

(ii) clothing

(iii) Fuel and lghting

(iv) House rent

(v) Miscellaneous

(b) Selection of adequate number of representative items on the basis of stratified sampling

technique.

(c) Uniformity in the quality and grades of the commodities to facilitate easy comparison.

**4. Collection of data**

The prices of selected commodities along with their quantities consumed in different periods form the data for the construction of an index number. Reliable data can be obtained for the purpose of constructing index numbers from regular published quotations and periodic special reports from the merchants, producers, exporters or others who possess the basic information needed for constructing an index number. Apart from this, the data can also be obtained from standard trade journals, newspapers and reliable and unbiased filed agencies which have conducted enquiry.

**5. Selection of an Average**

Since indices are specialised averages, the next problem is the selection of an appropriate average. It is necessary to decide a particular average which can be used for constructing an index number. There are several averages such as arithmetic mean, median, mode, geometric mean and harmonic mean which can be used to construct an index number. Though theoretically all such averages can be used, yet in practice, a choice has to be made between arithmetic mean and geometric mean. Median and mode are not used in the construction of index numbers, as they are affected by the few middle items in the series. The arithmetic mean is popularly used due to its simplicity and understandability but the value of the index is affected by the extreme items. it gives greater weights to bigger items and if there is a substantial rise in one commodity, the value of mean will shoot-up very significantly. Since the arithmetic mean measures only the absolute change , it is not suitable average for the computation of index numbers.

From the theoretical point of view, geometric mean is considered to be the best average to compute the index numbers because of the following reasons :

i) Geometric mean gives equal wieghts to equal ratio of change.

ii) Geometric mean is less susceptible to major fluctuations that arise in the values of individual items due to violent fluctuations.

iii) Index numbers constructed with the help of geometric mean are reversible and as such base shifting is easily possible.

**6. Seleletion of weights**

the next problem in the construction of index numbers is selection of appropriate weights. The weights should be relevant, timely and free from bias. According to John I Griffen, "Weighting is designed to give component series an importance in proper relation to their real significance". In other words , wieghts refer to the relative importance of the different items in the construction of index number. In this regard, the weights may be average quantities or prices of base year or current year, or the average quantities or prices of several years and hypothetical quantities or prices, etc. While selecting the wieghts, the following two points are to be kept in

mind:

(i) The method of slecting the wieghts must be on the basis of the purpose of index number as the changes in weights will also change the value of index number.

(ii) Selection of appropriate weights which involves less computational work and permits precise interpretation.

The weighting system can be classified into two categories, viz., (i) implicit weights and (ii) Explicit weights.

i ***implicit weights :*** By implict weights, we mean the weights that are not explicitly assigned to any commodity but the commodity to which greater importance is attached will represent a number of times . In this method, weights are not apparent but items are implicitly weighted. For instance, if in an index number, rice is to recieve 4 weights and wheat two weights, then four itmes of rice and two times of wheat are to be included. In India, the Bombay wholesale prices are computed by adopting implicit weights only.

ii ***Explicit Weights*** :In the case of explicit weigting, some outward evidence of importance of the various items in the index is given. The weights are explicity assigned to commodities. Only one variety of commodity is included in the construction of index number by getting the product of price relatives and the assigned weights. These weights are decided on some logical basis. For example, if wheat and rice are weighted in accordance with their value of the net output 8 : 2 then wheat would get a weight of 8 and rice 2.

The explicit weights include the following:

(1) The weights can be in terms of production figures, consumption figures, distribution figures which are used to identify the economic importance of the commodities invloved in the construction of an index number.

(2) These can be quantity weights and value weights. Quantity weights (q) imply the quantity of commodities produced, distributed or consumed in some period . Value weight is the product of price and quantity produced, distributed or consumed (Pq).

Generally, the weights may be fixed weights or fluctuating weights. In the case of fixed weights, the weights are fixed and they do not reflect the fluctuations in the relative importance of commodities. On the other hand, the fluctuating weights will take into account the changes in the relative importance of the commodities from time to time.

**7. Selection of an appropriate formula**

The final problem of constructing an index number is tne selection of an appropriate formula. There are more than a hundred formulae which have been suggested to compute index number. These different formula usually produce different results when applied to the same data. Therefore, there is a necessity to select an appropriate formula. A formula is said to be appropriate, if it satisfies the mathematical tests suggested by some theoretical statisticians. Among these tests, the most important are Time Reverseal Test, Factor Revearsal Test and

Circular Test.

The choice of an appropriate index number formuale would depend not only on the purpose of the index number, but also on the availability of the data. Fisher suggested that an appropriate index formula is that which satisfies both Time Reversal and Factor Reversal Tests. Of all the index number formulae none is regarded as the best under all circumstances. Therefore, on the basis of the knowledge of the different formulae, an investigator has to choose the appropriate formula for the construction of an index number.

**Check your progress - 1**

List out the problems involved in the construction of index numbers

## 27.3 SUMMING UP

While constructing index numbers, it is necessary to take utmost care and precaution to avoid misinterpretation. While constructing index number, proper consideration must be given to the problems such as the purpose of an index number, selection of base period, selection of items, selection of suitable average, selection of an appropriate formula, sources of data and system of weighting.

## 27.4 CHECK YOUR PROGRESS : MODEL ANSWERS

1.

- Define the purpose
- Selection of base period
- Selection of items
- Collection of data
- Selection of an average
- Selection of Weights
- Selection of an appropriate formula.

## 27.5 MODEL EXAMINATION QUESTIONS

**A. Short Questions**

1. Selection of base period.
2. Selection of weights.
3. Selection of an average

**B. Essay Questions**

4. Index numbers are economic barometers'. Explain this statement and mention what precautions should be taken in making use of index numbers.

5. Explain the problems that are involved in the construction of index numbers.

## 27.6 RECOMMENDED BOOKS


1. Gupta, S.P : "Statistical Methods", Sultan Chand & Company, New Delhi.

2. Gupta, B.N. : "Statistics", Sahitya bhavan, Agra.

3. Gupta S.C. : "Fundamentals of Statistics", Himalaya Pub. House, Bombay.

4. Simpson and Kafka : "Basic statistics", Oxford and I.B.H. Publishing Company, Calcutta.


## 27.7 GLOSSARY

**Explicit Weight** : Assigning a weightage to a commodity depending on some outward evidence.

**Implicit weight** : Assigning a weightage to a commodity depending on its importance.

# UNIT-28 : UNWEIGHTED INDEX NUMBERS

**Contents**



## 28.0 AIMS AND OBJECTIVES

The aim of this unit is to expose you to the construction of unweighted Index Numbers.

At the end of this unit, you should be able to :

- explain the meaning of unweighted Index Numbers
- work out the problems based on simple aggregative method
- workout the problems on average of relatives method.

## 28.1 INTRODUCTION

The methods of constructing an index number can be classified into two broad categories viz.,

i) Unweighted indices and

ii) Weighted indices.

While constructing unweighted indices, weights are not assigned to the commidities, but the prices of the commodities alone are taken into account. On the other hand, to construct the weighted index numbers, not only the prices of commodities are taken into account, but their quantities i.e., weights are also considered.

**Check your progress - 1**

Distinguish between weighted and unweighted index numbers

## 28.2 CONSTRUCTION OF UNWEIGHTED INDICES

The unweighted indices can be computed with the help of the following two methods viz.,

28.2.1 Simple aggregative method and

28.2.2 Simple average of relatives method.

### 28.2.1 SIMPLE AGGREGATIVE METHOD

This is one of the simplest methods of constructing price index. According to this method, the current year's aggregate prices are expressed as percentage of the aggregate prices in the base year.

*Procedure*

In order to construct simple aggregative price index the following procedure is adopted :

i) Add the prices of various commodities in the current year;

ii) Add the prices of various commodities in the base year;

iii) Divide the total of current year prices by the total of base year prices and multiply the quotient by 100. The resulting value is the simple aggregative price index.

The price index can be obtained with the help of the following formula.

$$P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100$$

Where, $P_{01}$ = Price index for current year on the basis of base year prices.

$\Sigma p_1$ = Total of current year prices of various commodities.

$\Sigma p_0$ = Total of base year prices of various commodities.

**Illustration - 1**

From the information given below, construct price index for 1983 taking 1980 as base year.

| Commodities | : | A | B | C | D | E |
|---|---|---|---|---|---|---|
| Price in 1980(Rs.) | : | 60 | 40 | 90 | 100 | 30 |
| Price in 1983(Rs.) | : | 80 | 50 | 100 | 110 | 40 |

**Solution :**

CONSTRUCTION OF PRICE INDEX FOR 1983

| Commodities | Prices in 1980 (Rs.) $p_0$ | Prices in 1983 (Rs.) $p_1$ |
|---|---|---|
| A | 60 | 80 |
| B | 40 | 50 |
| C | 90 | 100 |
| D | 100 | 110 |
| E | 30 | 40 |
| | $\Sigma p_0 = 320$ | $\Sigma p_1 = 380$ |

$$P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100$$

Here, $\Sigma p_0 = 320$ and $\Sigma p_1 = 380$.

Substituting the values in the formula

$$P_{01} = \frac{380}{320} \times 100$$

$$= 118.75$$

$\therefore$ Price index for 1983 = 118.75.

**Illustration - 2**

Compute price indices for 1983 and 1980 taking 1976 as base year from the data given below:

| Year | Commodities | | | | | |
|---|---|---|---|---|---|---|
| | Rice | Wheat | Sugar | Dal | Oil | Others |
| 1983 | 55 | 70 | 25 | 70 | 90 | 130 |
| 1980 | 40 | 60 | 20 | 50 | 80 | 120 |
| 1976 | 30 | 40 | 10 | 40 | 60 | 80 |

**Solution :**

**Computation of price indices for 1983 and 1980**

| Commodities | Prices in 1976 (Rs.) $p_0$ | Prices in 1980 (Rs.)$p_1$ | Prices in 1983 (Rs.)$p_2$ |
|---|---|---|---|
| Rice | 30 | 40 | 55 |
| Wheat | 40 | 60 | 70 |
| Sugar | 10 | 20 | 25 |
| Dal | 40 | 50 | 70 |
| Oil | 60 | 80 | 90 |
| Others | 80 | 120 | 130 |
| | $\Sigma p_0 = 260$ | $\Sigma p_1 = 370$ | $\Sigma p_2 = 440$ |

$$\text{Price index for 1980} = \frac{\Sigma p_1}{\Sigma p_0} \times 100$$

Here, $\Sigma p_0 = 260$ and $\Sigma p_1 = 370$.

Substituting the values in the formula,

$$P_{01} = \frac{370}{260} \times 100$$

$$= 142.3$$

$\therefore$ Price index for 1980 = 142.3.

Price index for 1983 :

$$P_{02} = \frac{\Sigma p_2}{\Sigma p_0} \times 100$$

Here, $\Sigma P_0 = 260$ and $\Sigma P_2 = 440$.

Substituting the values in the formula,

$$P_{02} = \frac{440}{260} \times 100$$
$$= 169.2$$

$\therefore$ Price index for 1983 = 169.2.

Simple aggregative method is easy to follow and simple to compute the price index. However, it has some limitations. This method does not take into account the relative importance of various commodities. High-priced commodities may show their influence on the low-priced commodities, as prices of commodities alone are considered to compute price index.

## 28.2.2 SIMPLE AVERAGE OF RELATIVE METHOD

This is another method of computing unweighted price index. This method is based on averaging the price relatives of individual commodities. The price relatives are obtained by expressing the price of a commodity in a current year as a percentage of its price in the base year i.e.,

$$\text{Price relative (P)} = \frac{\textit{Price in the current year}(p_1)}{\textit{Price in the base year}(p_0)} \times 100$$

The average of the price relatives is obtained by using any one of the measures of central tendency such as arithmetic mean, median, mode, geometric mean, harmonic mean, etc. However, arithmetic mean and geometric mean are generally used to calculate the index number.

**Calculation of Index Number**

(a) When arithmetic mean is used to average the price relatives, the following procedure is adopted :

(i) Calculate the price relative (P) for various commodities by using the formula $\frac{p_1}{p_0} \times 100$.

(ii) Obtain the total of price relatives ($\Sigma P$) .

(iii) Apply the formula

$$P_{01} = \frac{\Sigma P}{N}$$

Where, $P_{01}$ = Price index for current year on the basis of base year price.

$\Sigma P$ = Sum of price relatives of all the commodities.

N = Number of items or commodities.

**Illustration - 3**

Construct price index for 1983 on the basis of 1980, with the help of average of price relatives method by using arithmetic mean.

Commodities : A B C D E

Prices in 1980(Rs.) : 20 30 40 50 10

Prices in 1983(Rs.) : 30 60 60 80 30

**Solution :**

**Construction of price index using arithmetic mean of price relatives.**

| Commodities | Prices in 1980 (Rs.) $p_0$ | Prices in 1983 (Rs.) $p_1$ | Price relatives $\left(P = \frac{p_1}{p_0} \times 100\right)$ |
|---|---|---|---|
| A | 20 | 30 | $\frac{30}{20} \times 100 = 150$ |
| B | 30 | 60 | $\frac{60}{30} \times 100 = 200$ |
| C | 40 | 60 | $\frac{60}{40} \times 100 = 150$ |
| D | 50 | 80 | $\frac{80}{50} \times 100 = 150$ |
| E | 10 | 30 | $\frac{30}{10} \times 100 = 300$ |
| | | | $\Sigma P = 900$ |

$$P_{01} = \frac{\Sigma P}{N}$$

Here, $\Sigma P = 960$ and N = 5.

**Substituting the values in the formula,**

$$P_{01} = \frac{960}{5}$$

$$= 192$$

$\therefore$ **Price index for 1983 = 192.**

**Illustration - 4**

**Calculate price index for the data given below.**

| Items | Units | Prices in 1978 (Rs.) | Prices in 1983 (Rs.) |
|---|---|---|---|
| Edible Oil | Kg. | 4 | 10 |
| Butter | Kg. | 15 | 30 |
| Buscuits | Kg. | 8 | 24 |
| Bread | 400 gms. | 1 | 3 |

**Solution :**

**CONSTRUCTION OF PRICE INDEX BY USING ARTIHMETIC MEAN OF PRICE RELATIVES.**

| Items | Units | Prices in 1978 (Rs.) $p_0$ | Prices in 1983 (Rs.) $p_1$ | Price relatives P |
|---|---|---|---|---|
| Edible Oil | Kg. | 4 | 10 | $\frac{10}{4} \times 100 = 250$ |
| Butter | Kg. | 15 | 30 | $\frac{30}{15} \times 100 = 200$ |
| Buscuits | Kg. | 8 | 24 | $\frac{24}{8} \times 100 = 300$ |
| Bread | 400 gms. | 1 | 3 | $\frac{3}{1} \times 100 = 300$ |
| | | | | $\Sigma P = 1050$ |

$$P_{01} = \frac{\Sigma P}{N}$$

Here, $\Sigma$ P = 1050 and N = 4,

Substituting the values in the formula,

$$P_{01} = \frac{1050}{4}$$

$$= 262.5$$

$\therefore$ Price index for 1983 = 262.5.

**Calculation of Index Numbers**

(b) When geometric mean is used :

When geometric mean is used to average the price relatives of various commodities, the following procedure is adopted to compute price index.

(i) Calculate price relatives (P) for various commodities

i.e., $\frac{p_1}{p_0} \times 100$

(ii) Find out log values for various price relatives.

(iii) Obtain the total of log values of price relatives.

(iv) Divide the sum of log values of price relatives by the number of items.

(v) Find the antilog value for the quotient obtained and the resulting value is price index.

Simbolically,

$$P_{01} = \text{Antilog } \frac{\Sigma log P}{N}$$

Where, $P_{01}$ = Price index for current year on the basis of base year price.

$\Sigma log P$ = Sum of log values of price relatives.

N = Number of items.

**Illustration - 5**

Calculate price index by using geometric mean of price relatives method.

| Commodities | Units | (Price in Rs.) 1980 | 1983 |
|---|---|---|---|
| Gold | 10 gms. | 300 | 1500 |
| Silver | 100 gms. | 800 | 2400 |
| Copper | 1000 gms. | 80 | 100 |
| Iron | 1000 gms. | 40 | 80 |

**Solution :**

Construction of price index by using geometric mean of price relatives method:

| Commodities | Units | Price in (Rs.) 1980 $p_0$ | 1983 $p_1$ | $\left(\frac{p_1}{p_0} \times 100\right)$ P | log P |
|---|---|---|---|---|---|
| Gold | 10 gms. | 300 | 1500 | 500 | 2.6990 |
| Silver | 100 gms. | 800 | 2400 | 300 | 2.4771 |
| Copper | 1000 gms. | 80 | 100 | 125 | 2.0969 |
| Iron | 1000 gms. | 40 | 80 | 200 | 2.3010 |
| | | | | | $\Sigma$ log P = 9.5740 |

$$P_{01} = \text{Antilog } \frac{\Sigma log P}{N}$$

Here, $\Sigma$ log P = 9.5740 and N = 4.

Substituting the values in the formula,

$$P_{01} = \text{Antilog } \frac{9.574}{4}$$

$$= \text{Antilog } 2.3925$$

$$= 246.9$$

$\therefore$ Price index for 1983 = 246.9.

**Illustration - 6**

The following are the prices of commodities in 1980 and 1983. Construct price index by using geometric mean of price relatives method.

| Commodities | Prices in 1980 (Rs.) | Prices in 1983 (Rs.) |
|---|---|---|
| Rice | 25 | 45 |
| Wheat | 50 | 60 |
| Sugar | 10 | 20 |
| Dal | 40 | 80 |
| Oil | 80 | 100 |
| Fuel | 120 | 150 |

**Solution :**

**CONSTRUCTION OF PRICE INDEX BY USING GEOMETRIC MEAN OF PRICE RELATIVES METHOD.**

| Commodities | Prices in 1980(Rs.) $p_0$ | Prices in 1983(Rs.) $p_1$ | $\left(P = \frac{p_1}{p_0} \times 100\right)$ P | log P |
|---|---|---|---|---|
| Rice | 25 | 45 | 180 | 2.2553 |
| Wheat | 50 | 60 | 120 | 2.0792 |
| Sugar | 10 | 20 | 200 | 2.3010 |
| Dal | 40 | 80 | 200 | 2.3010 |
| Oil | 80 | 100 | 125 | 2.0969 |
| Fuel | 120 | 150 | 125 | 2.0969 |
| | | | | Σlog P = 13.1303 |

$$P_{01} = \text{Antilog } \frac{\Sigma log P}{N}$$

Here, Σ log P = 13.1303 and N = 6.

Substituting the values in the formula,

$$P_{01} = \text{Antilog } \frac{13.1303}{6}$$

$$= \text{Antilog } 2.1883$$

$$= 154.3$$

∴ Price index for 1983 = 154.3.

The average price relative method eliminates the impact of extreme items, since all the values are given equal importance. The price relatives are pure numbers and they can also remove the effect of different units of measurement. However, this method has certain limitations. Out of the several averages, selection of an appropriate average is a difficult task. This method does not give relative importance to various commodities, as all the items are treated by assigning equal weights.

## 28.3 SUMMING UP

The formulae used to construct index numbers can be grouped under two heads, viz., (i) unweighted indices and (ii) weighted indices.

Each of these types may be further divided under two heads, viz., (i) Aggregative and (ii) Average Relative Indices.

In the unweighted indices, weights are not assigned whereas in the weighted indices, weights are assigned to various items.

The unweighted indices can be computed with the help of two methods, viz., (i) simple aggregative method and (ii) simple average of relatives method.

While simple aggregative unweighted index number is computed by expressing current year aggregate prices as percentage of the aggregate prices in the base year, index number by the method of simple average of relatives is computed by averaging the price relatives of individual commodities. The price relatives are obtained by expressing the price of a commodity in a current year as a percentage of its price in the base year.

## 28.4 CHECK YOUR PROGRESS: MODEL ANSWERS

1. In the case of weighted index numbers both the price and quantity of the commodities are taken into account. But in the case of unweighted index numbers only the prices of the commodities are taken into consideration.

## 28.5 MODEL EXAMINATION QUESTIONS

**A. Short Questions**

1. What is a price relative ?
2. What are the merits of simple aggregative price index ?

**B. Essay Questions**

3. What do you understand by price relatives ? Discuss the methods of constructing index numbers based on price relatives.
4. 'In the construction of index number, the advantages of geometric mean are greater than those of the arithmetic mean'. Discuss.

**EXERCISES**

5. The following are the prices of six different commodities for 1980 and 1983. Compute a price index by

   (a) Simple aggregate method and

   (b) Average of price relatives method.

| Commodities | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| Price in 1980(Rs.) | 40 | 60 | 70 | 20 | 50 | 100 |
| Price in 1983(Rs.) | 50 | 70 | 80 | 30 | 80 | 110 |

(Ans: a) = 123.53 b) = 129.33)

6. Calculate index number for 1989 on the basis of the prices for 1985 from the following.

| Items | Bricks | Timber | Plaster | Board | Sand | Cement |
|---|---|---|---|---|---|---|
| Prices (1985) | 10 | 20 | 4 | 5 | | 10 |
| Prices (1989) | 15 | 25 | 6 | 8 | | 15 |

(Ans: 140.82)

7. Calculate index numbers from the following data by simple aggregative method taking prices of 1980 as base.

| | Price per unit (in Rupees) | | | |
|---|---|---|---|---|
| Commodity | [illegible] | 1982 | 1985 | 1988 |
| A | 0.20 | 0.30 | 0.40 | 0.75 |
| B | 0.30 | 0.35 | 0.35 | 0.85 |
| C | 0.25 | 0.31 | 0.30 | 0.50 |
| D | 1.25 | 2.00 | 2.25 | 2.75 |

(Ans: 148, 165, 242.5)

8. From the following details, construct an Index for 1989 taking 1985 as the base by the price relative method using (a) arithmetic mean, (b) geometric mean for averaging relatives.

| Commodities | A | B | C | D |
|---|---|---|---|---|
| Prices (1985) | 10 | 20 | 25 | 50 |
| Prices (1989) | 12 | 18 | 40 | 75 |

(Ans : a. 130, b = 126.9)

## 28.6 RECOMMENDED BOOKS


1. Gupta, S.P : "Statistical Methods", Sultan Chand & Company, New Delhi.
2. Gupta, B.N : "Statistics", Sahitya Bhavan, Agra.
3. Gupta, S.C : "Fundamentals of Statistics", Himalaya Pub. House, Bombay.
4. Simpson and Kafka : "Basic Statistics", Oxford and I.B.H.Publishing Company,Calcutta.




## 28.7 GLOSSARY

1. Simple Aggregative method : It is the method of expressing the current year's aggregate prices as a percentage of aggregate prices in the base year.
2. Simple Average of Price Relatives method : It involves the ascertaining the price relatives of all the commodities by dividing the sum of price relatives by the number of commodities.

# UNIT - 29 WEIGHTED INDEX NUMBERS

**Contents**





## 29.0 AIMS AND OBJECTIVES

The aim of this unit is to explain both the theoretical and practical aspects of various wieghted index numbers.

At the end of the unit, you should be able to :

- explain and work out the various categories of weighted index numbers
- explain and work out quantity Index Numbers
- explain and workout value Index Numbers.

## 29.1 INTRODUCTION

For the purpose of constructing the weighted index numbers, prices of commodities as well as their quantities are taken into account. To attribute the appropriate importance to each of the items included in an aggregate index, weights are assigned on the basis of their quantities.

The weights in this regard may be production, consumption or distribution figures. The methods of constructing weighted index numbers can be classified as weighted aggregative index numbers and weighted average of relatives index numbers.

## 29.2 WEIGHTED AGGREGATIVE INDEX NUMBERS

Weighted aggregative index numbers are of the simple aggregative type with the fundamental difference that the weights are assigned to the various items included in the index.In this method, the prices and quantities of the various cor modities are taken into account to construct the index numbers. On the basis of methods of assigning the weights, the weighted aggregative index numbers are classified into the following categories.

29.2.1 Laspeyres method

29.2.2 Paasche's method

29.2.3 Dorbish and bowley's method

29.2.4 Marshall-Edgeworth method

29.2.5 Kelly's method

29.2.6 Fisher's Ideal method

29.2.7 Walsch Price Index.

### 29.2.1 LASPEYRES METHOD

It is one of the weighted aggrevative price indices where the quantities of the base year are taken as weights.

**Procedure**

In order to construct Laspeyres index the following procedure is followed:

(i) Multiply the current year prices ($p_1$) of various commodities by the base year quantities ($q_o$) and find out the total ($\Sigma p_1 q_0$)

(ii) Multiply the base year prices ($p_0$) of various commodities by the base year quantities ($q_0$) and find out the total ($\Sigma p_0 q_0$)

(iii) Divide the total of the product of current year prices and base year quantities ($\Sigma p_1 q_0$) by the product of base year prices and base year quantities ($\Sigma p_0 q_0$) and multiply the quotient by 100. The resulting value is price index. The formula for the construction of Laspeyres index is :

$$P_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$

Where $P_{01}$ = Price index for current year on the basis of base year.

$\Sigma p_1 q_0$ = Sum of the products of current year prices and base year quantities.

$\Sigma p_0 q_0$ = Sum of the products of base year prices and base year quantities.

**Illustration - 1**

Construct index number for 1983 on the basis of 1980 prices using Laspeyres method for the following data:

| Commodity | 1980 | | 1983 | |
|---|---|---|---|---|
| | Price (Rs.) | Quantity | Price (Rs.) | Quantity |
| A | 2 | 10 | 2 | 15 |
| B | 3 | 15 | 3 | 20 |
| C | 4 | 20 | 4 | 25 |
| D | 3 | 15 | 3 | 30 |

**Solution:**

CONSTRUCTION OF LASPEYRES PRICE INDEX

| Commodity | 1980 | | 1983 | | | |
|---|---|---|---|---|---|---|
| | Price (Rs.) | Quantity | Price (Rs.) | Quantity | | |
| | $p_0$ | $q_0$ | $p_1$ | $q_1$ | $p_0q_0$ | $p_1q_1$ |
| A | 2 | 10 | 2 | 15 | 20 | 30 |
| B | 3 | 15 | 3 | 20 | 45 | 60 |
| C | 4 | 20 | 4 | 25 | 80 | 100 |
| D | 3 | 15 | 3 | 30 | 45 | 90 |

$\Sigma p_0q_0 = 190 \quad \Sigma p_1q_0 = 280$

$$P_{01} = \frac{\Sigma p_1q_1}{\Sigma p_0q_0} \times 100$$

Here, $\Sigma p_0q_0 = 190$ and $\Sigma p_1q_0 = 280$

Substituting the values in the formula,

$$P_{01} = \frac{280}{190} \times 100$$

$$= 147.4$$

$\therefore$ Price index for 1983 = 147.4

Laspeyres index enables us to find out the change in the aggregate value of the base year quantities in terms of current year prices. This index number is widely used in practical computation of index numbers. However, this formula for computation of index number is frequently criticised for its upward bias in the value of the index number. The assumption of

base year quantities as wieghts and that these will remain same in the current year also is sledom correct, because the customers shift their buying and consumption pattern from time to time and in this process, the consumption of highly price items may be decreased. When the prices are declining, the consumers shift their purchases from the costly items to those items whose prices decline significantly. If we use the base year quantities as weights, such type of fluctuations may not be reflected effectively.

## 29.2.2 PAASCHE'S METHOD

In this method, the quantities of the current year are taken as weights. This method was named after German Statistician Paasche who formulated it in 1874.

**Procedure**

The following procedure is followed to construct the Paasche's index.

(a) Multiply the current year prices of various commodities by their respective current year weights.

(b) Multiply the base year prices of various commodities by current year weights.

(c) Divide the product of current year prices and current year weights by the product of base year price and current year weights and multiply the quotient by 100. The resulting value is the Paasche's price index.

The formula for the construction of Paasche's price index is

$$P_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$$

Where, $p_{01}$ = Price index for current year on the basis of base year values.

$\Sigma p_1 q_1$ = Sum of products of current year prices of various commodities and current year quantities.

$\Sigma p_0 q_1$ = Sum of products of base year prices of various commodities and current year quantities.

**Illustration - 2**

Calculate Paasche's price index from the following data.

| | | Base year 1978 | | Current year 1983 | |
|---|---|---|---|---|---|
| Items | Units | Price (Rs.) | Quantity | Price (Rs.) | Quantity |
| Eggs | Dozen | 4 | 20 | 5 | 20 |
| Milk | Litre | 2 | 30 | 3 | 30 |
| Meat | Kg. | 10 | 50 | 15 | 50 |
| Rice | Kg. | 3 | 70 | 6 | 70 |

**Solution:**

## CALCULATION OF PAASCHE'S INDEX

| | | Base year 1980 | | Current year 1983 | | | |
|---|---|---|---|---|---|---|---|
| Items | Units | $p_0$ | $q_0$ | $p_1$ | $q_1$ | $p_1q_1$ | $p_0q_1$ |
| Eggs | Dozen | 4 | 20 | 5 | 20 | 100 | 80 |
| Milk | Litre | 2 | 30 | 3 | 30 | 90 | 60 |
| Meat | Kg. | 10 | 50 | 15 | 50 | 750 | 500 |
| Rice | Kg. | 3 | 70 | 6 | 70 | 420 | 210 |
| | | | | | | $\Sigma p_1q_1$ =1360 | $\Sigma p_0q_1$=850 |

$$P_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$$

Here $\Sigma p_1 q_1 = 1360$ and $\Sigma p_0 q_1 = 850$.

Substituting the values in the formula,

$\therefore$ Price index for 1983 = 160.

The main difficulty under this method is that since current year quantities are taken as weights, it involves cumbersome task for obtaining current year quantities every time. Further, it also involves additional expenditure for the collection of necessary information. Owing to this reason, the Paasche's index is not frequently used when the number of commodities are more.

**Distinction between Laspeyres Index and Paasche's Index**

The following are the differences between the Laspeyres and Paasche's index numbers:

(i) The Laspeyres index shows the upward bias in the value of index number, whereas Paasche's index shows the downward bias.

(ii) While base year quantities are used as weights in Laspeyres index, current year w eghts are taken as weights in Paasche's index.

(iii) The Laspeyres index number does not take into account the discontinuation of cc :tly items by the customer by substituting them with relatively cheaper items over a perio· of time. But the Paasche's formula takes this fact into account.

## 29.2.3. DORBISH AND BOWLEY'S METHOD

In Dorbish and Bowley's method, the index is obtained by computing the a :rage of Laspeyres and Paasche's indices . It takes into account the current year prices and qu :tities as well as base year prices and quantities . The formula for construction the index numb· r is

$$P_{01} = \frac{L+P}{2}$$

Where L = Laspreyres index

P = Paasche's index

$$P_{01} = \frac{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}{2} \times 100$$

**Illustration - 3**

Calculate the price index from the following data with the help of Dorbish and Bowley's method.

| Items | Price (in Rs.) | | Quantity | |
|---|---|---|---|---|
| | 1975 | 1983 | 1975 | 1983 |
| A | 2 | 3 | 10 | 15 |
| B | 3 | 4 | 20 | 10 |
| C | 4 | 5 | 10 | 10 |
| D | 3 | 4 | 15 | 20 |

**Solution :**

## CALCULATION OF DROBISH AND BOWLEY'S INDEX

| Items | Base year | | Current year | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1975 | | 1983 | | | | | |
| | $p_0$ | $q_0$ | $p_1$ | $q_1$ | $p_1q_0$ | $p_1q_1$ | $p_0q_0$ | $p_0q_1$ |
| A | 2 | 10 | 3 | 15 | 30 | 45 | 20 | 30 |
| B | 3 | 20 | 4 | 10 | 80 | 40 | 60 | 30 |
| C | 4 | 10 | 5 | 10 | 50 | 50 | 40 | 40 |
| D | 3 | 15 | 4 | 20 | 60 | 80 | 45 | 60 |
| | | | | | $\Sigma p_1q_0$ = 220 | $\Sigma p_1q_1$ = 215 | $\Sigma p_0q_0$ = 165 | $\Sigma p_0q_1$ = 160 |

$$P_{01} = \frac{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}{2} \times 100$$

Here $\Sigma p_1q_0 = 220$, $\Sigma p_1q_1 = 215$, $\Sigma p_0q_0 = 165$ and $\Sigma p_0q_1 = 160$

Substituting the values in the formula,

$$P_{01} = \frac{\frac{220}{165} + \frac{215}{160}}{2} \times 100$$

$$= \frac{1.33+1.34}{2} \times 100$$

$$= 1.335 \times 100$$

$$= 133.5$$

∴ Price index for 1983 = 133.5

Dorbish and Bowley's method balances the upward and down ward bias present in Laspeyres and Passche's indices repsectively, since the average of two methods are used in the construction of index number.

## 29.2.4 MARSHALL-EDGEWORTH METHOD

In this method, we have to take into account both prices and quantities of current year and base year to construct the weighted aggregative price index.

Symbolically,

$$P_{01} = \frac{\Sigma p_1 q_0 + \Sigma p_1 q_1}{\Sigma p_0 q_0 + \Sigma p_0 q_1} \times 100$$

**Illustration - 4**

Construct price index by using Marshall-Edge worth method.

| | Base year 1981 | | Current year 1983 | |
|---|---|---|---|---|
| Items | Price(Rs.) | Quantity | Price(Rs.) | Quantity |
| A | 3 | 70 | 4 | 80 |
| B | 6 | 130 | 5 | 150 |
| C | 8 | 50 | 10 | 40 |
| D | 5 | 40 | 8 | 50 |
| E | 4 | 30 | 7 | 60 |
| F | 7 | 100 | 10 | 120 |

**Solution :**

CALCULATION OF MARSHALL- EDGEWORTH PRICE INDEX

| Items | Base year 1981 | | Current year 1983 | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $p_0$ | $q_0$ | $p_1$ | $q_1$ | $p_0q_0$ | $p_0q_1$ | $p_1q_0$ | $p_1q_1$ |
| A | 3 | 70 | 4 | 80 | 210 | 240 | 280 | 320 |
| B | 6 | 130 | 5 | 150 | 780 | 900 | 650 | 750 |
| C | 8 | 50 | 10 | 40 | 400 | 320 | 500 | 400 |
| D | 5 | 40 | 8 | 50 | 200 | 250 | 320 | 400 |
| E | 4 | 30 | 7 | 60 | 120 | 240 | 210 | 420 |
| F | 7 | 100 | 10 | 120 | 700 | 840 | 1000 | 1200 |
| | | | | | $\Sigma p_0q_0$ = 2410 | $\Sigma p_0q_1$ = 2790 | $\Sigma p_1q_0$ = 29607 | $\Sigma p_1q_1$ = 3490 |

$$P_{01} = \frac{\Sigma p_1 q_0 + \Sigma p_1 q_1}{\Sigma p_0 q_0 + \Sigma p_0 q_1} \times 100$$

Here, $\Sigma p_0 q_0 = 2410$, $\Sigma p_0 q_1 = 2790$, $\Sigma p_1 q_0 = 2960$ and $\Sigma p_1 q_1 = 3490$.

Substituting the values in the formula,

$$P_{01} = \frac{2960+3490}{2410+2790} \times 100$$

$$= \frac{6450}{5200} \times 100$$

$$= 124.04$$

$\therefore$ Price index for 1983 = 124.04

Marshall-Edgeworth method is very easy for constructing index number and very simple to understand. This method gives a very close approximation to the results that are obtained from the fisher's Ideal index.

## 29.2.5 KELLY'S METHOD

Kelly's method is also known as fixed wieght aggregative method. In this method, the weights are the quantities which may refer to some period, but not necessarily the quantities of base year or current year. Therefore the average quantities of two or more years are taken as weights.

Symbolically,

$$P_{01} = \frac{\Sigma p_1 q}{\Sigma p_0 q} \times 100$$

$$q = \frac{q_0 + q_1}{2}$$

Where, q = Average weights

$q_0$ = Base year Weights

$q_1$ = Current year weights .

**Illustration - 5**

Following are the price and quantities of certain items. Construct Kelly's index.

| Items | Quantities | | Prices (in Rs.) | |
|---|---|---|---|---|
| | 1981 | 1983 | 1981 | 1983 |
| A | 20 | 30 | 10 | 20 |
| B | 40 | 60 | 8 | 10 |
| C | 30 | 80 | 20 | 30 |
| D | 10 | 30 | 10 | 20 |

**Solution:**

CONSTRUCTION OF KELLY'S INDEX

| Item | Quantities | | Prices (in Rs.) | | | | |
|---|---|---|---|---|---|---|---|
| | 1981 | 1983 | 1981 | 1983 | | | |
| | $q_0$ | $q_1$ | $p_0$ | $p_1$ | $q = \frac{q_0+q_1}{2}$ | $p_0q$ | $p_1q$ |
| A | 20 | 30 | 10 | 20 | 25 | 250 | 500 |
| B | 40 | 60 | 8 | 10 | 50 | 400 | 500 |
| C | 30 | 80 | 20 | 30 | 55 | 1100 | 1650 |
| D | 10 | 30 | 10 | 20 | 20 | 200 | 400 |
| | | | | | | $\Sigma p_0q = 1950$ | $\Sigma p_1q = 3050$ |

$$p_{01} = \frac{\Sigma p_1 q}{\Sigma p_0 q} \times 100$$

Here $\Sigma p_0 q = 1950$ and $\Sigma p_1 q = 3050$

Substituting the values in the formula,

$$P_{01} = \frac{3050}{1950} \times 100$$

$$= 156.4$$

$\therefore$ Price index for 1983 = 156.4

On account of using fixed quantities as weighs to compute index number, Kelly's method is currently a hot favourite with the calculators of index numbers. The utility of Kelly's index is that the change in the base period does not necessitate a corresponding change in the weights and can be kept constant until new data becomes available for revising the index. The important advantage of this formula is that there are no yearly changes in the weights. Further, the base period can be changed without changing the weights.

## 29.2.6 FISHER'S IDEAL METHOD

Irving Fisher has given a formula to construct index, which is known as Ideal index. On account of the following reasons, the Fisher's index is called Ideal index.

(a) It is based on the geometric mean, which is considred to be the best average.

(b) It takes into account both current year and base year prices and quantities.

(c) It satisfies the tests of adequacy of an index number.

(d) It is free from bias since it balances the upward bias of Laspeyre's index and downward bias of Paasche's index.

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

**Illustration - 6**

Construct fisher's Ideal index from the following data.

| Commodity | Unit | Base year 1975 | | Current year 1983 | |
|---|---|---|---|---|---|
| | | Price | Quantity | Price | Quantity |
| Rice | Kg. | 3 | 60 | 4 | 80 |
| Wheat | Kg. | 6 | 100 | 8 | 120 |
| Sugar | Kg. | 8 | 60 | 10 | 80 |
| Edible oil | Litre | 4 | 40 | 6 | 60 |
| Miscellaneous | — | 5 | 30 | 8 | 50 |

**Solution:**

CONSTRUCTION OF FISHER'S IDEAL INDEX.

| Commodity | Base year 1975 | | Current year 1983 | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $p_1$ | $q_0$ | $p_1$ | $q_1$ | $p_0q_0$ | $p_0q_1$ | $p_1q_1$ | $p_1q_0$ |
| Rice | 3 | 60 | 4 | 80 | 180 | 240 | 320 | 240 |
| Wheat | 6 | 100 | 8 | 120 | 600 | 720 | 960 | 800 |
| Sugar | 8 | 60 | 10 | 80 | 480 | 640 | 800 | 600 |
| Edible oil | 4 | 40 | 6 | 60 | 160 | 240 | 360 | 240 |
| Miscellaneous | 5 | 30 | 8 | 50 | 150 | 250 | 400 | 240 |
| | | | | | $\Sigma p_0q_0$ = 1570 | $\Sigma p_0q_1$ = 2090 | $\Sigma p_1q_1$ = 2840 | $\Sigma p_1q_0$ = 2120 |

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

Here, $\Sigma p_0q_0 = 1570$, $\Sigma p_0q_1 = 2090$, $\Sigma p_1q_1 = 2840$ and $\Sigma p_1q_0 = 2120$.

Substituting the values in the formula,

$$P_{01} = \sqrt{\frac{2120}{1570} \times \frac{2840}{2090}} \times 100$$

$$= \sqrt{1.35 \times 1.35} \times 100$$

$$= 1.35 \times 100 = 135.$$

$\therefore$ Price index for 1983 = 135.

**Check your progress - 1**

Calculate fisher's Ideal Index for the following data.

| Commodities | Base year | | Current Year | |
|---|---|---|---|---|
| | $p_0$ | $q_0$ | $p_1$ | $q_1$ |
| A | 3 | 30 | 6 | 40 |
| B | 4 | 60 | 8 | 60 |
| C | 8 | 40 | 10 | 20 |
| D | 10 | 20 | 12 | 30 |
| E | 12 | 40 | 15 | 35 |

**Marshall -Edgeworth and Fisher's Ideal Indices**

The formulae of these two indices will balance the upward bias of Laspeyre's and downward bias Paasche's indices. These two methods provide a better estimation of true price indices. However, these formula require both current year prices and quantities and base year prices and quantities for the computation of index number. Hence, obtaining current year data involves expenditures and certain other difficulties. Further, these two formulae require more calculations.

## 29.2.7 WALSCH PRICE INDEX

In this method, the geometric mean of current year and base year quantities are taken as weights to compute the index number.

**Symbolically ,**

$$P_{01} = \frac{\Sigma p_1 \sqrt{q_0 q_1}}{\Sigma p_0 \sqrt{q_0 q_1}} \times 100$$

**Illustration - 7**

Construct index number by using Walsch method.

| commodities | A | B | C | D | E |
|---|---|---|---|---|---|
| 1980 Price | 25 | 30 | 20 | 45 | 40 |
| Quantity | 30 | 35 | 15 | 40 | 30 |
| 1983 Price | 40 | 45 | 25 | 50 | 45 |
| Quantity | 20 | 40 | 20 | 40 | 35 |

**Solution :**

CONSTRUCTION OF WALSCH PRICE INDEX

| Commodities | 1980 | | 1983 | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $q_0$ | $p_0$ | $q_1$ | $p_1$ | $q_0q_1$ | $\sqrt{q_0q_1}$ | $p_1\sqrt{q_0q_1}$ | $p_0\sqrt{q_0q_1}$ |
| A | 30 | 25 | 20 | 40 | 600 | 24.50 | 980.0 | 612.5 |
| B | 35 | 30 | 40 | 45 | 1400 | 37.42 | 1683.9 | 1122.6 |
| C | 15 | 20 | 20 | 25 | 300 | 17.32 | 433.0 | 346.4 |
| D | 40 | 45 | 40 | 50 | 1600 | 40.00 | 2000.0 | 1800.0 |
| E | 30 | 40 | 35 | 45 | 1050 | 32.40 | 1458.0 | 1296.0 |
| | | | | | | | $\Sigma p_1\sqrt{q_0q_1}$ = 6554.9 | $\Sigma p_0\sqrt{q_0q_1}$ = 5177.5 |

$$P_{01} = \frac{\Sigma p_1 \sqrt{q_0 q_1}}{\Sigma p_0 \sqrt{q_0 q_1}} \times 100$$

Here, $\Sigma p_1 \sqrt{q_0 q_1} = 6554.9$ and $\Sigma p_0 \sqrt{q_0 q_1} = 5177.5$

Subsituting the values in the formula,

$$P_{01} = \frac{6554.9}{5177.5} \times 100$$

$$= 126.6$$

$\therefore$ Price index for 1983 = 126.6.

**Illustration - 8**

compute weighted price index from the following data by using,

(i) Laspeyre's method

(ii) Paasche's method

(iii) Dorbish Bowley's method

(iv) Marshall-Edgeworth method

(v) Fisher's Ideal method

| Commodity | Base year 1981 | | Current year 1983 | |
|---|---|---|---|---|
| | Price | Quantity | Price | Quantity |
| Rice | 3 | 10 | 5 | 20 |
| Wheat | 6 | 12 | 8 | 15 |
| Oil | 5 | 15 | 6 | 20 |
| Ghee | 4 | 20 | 3 | 30 |

**Solution :**

CONSTRUCTION OF WEIGHTED PRICE INDICES

| Commodities | 1981 | | 1983 | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $p_0$ | $q_0$ | $p_1$ | $q_1$ | $p_1q_0$ | $p_1q_1$ | $p_0q_0$ | $p_0q_1$ |
| Rice | 3 | 10 | 5 | 20 | 50 | 100 | 30 | 60 |
| Wheat | 6 | 12 | 8 | 15 | 96 | 120 | 72 | 90 |
| Oil | 5 | 15 | 6 | 20 | 90 | 120 | 75 | 100 |
| Ghee | 4 | 20 | 3 | 30 | 60 | 90 | 80 | 120 |
| | | | | | $\Sigma p_1q_0$ = 296 | $\Sigma p_1q_1$ = 430 | $\Sigma p_0q_0$ = 257 | $\Sigma p_0q_1$ = 370 |

Here, $\Sigma p_1q_0 = 296$, $\Sigma p_1q_1 = 430$, $\Sigma p_0q_0 = 257$ and $\Sigma p_0q_1 = 370$.

(i) Laspeyre's method

$$P_{01} = \frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times 100$$

Substituting the values in the formula,

$$P_{01} = \frac{296}{257} \times 100$$

$$= 115.17$$

(ii) Paasche's method

$$P_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$$

Substituting the values in the formula,

$$P_{01} = \frac{430}{370} \times 100$$

$$= 116.2$$

(iii) Dorbish and Bowley's method

$$P_{01} = \frac{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}{2} \times 100$$

Substituting the values in the formula,

$$P_{01} = \frac{\frac{296}{257} + \frac{430}{370}}{2} \times 100$$

$$= \frac{2.31}{2} \times 100$$

$$= 115.52$$

(iv) Marshall-Edgeworth method

$$P_{01} = \frac{\Sigma p_1 q_0 + \Sigma p_1 q_1}{\Sigma p_0 q_0 + \Sigma p_0 q_1} \times 100$$

Substituing the values in the formula,

$$P_{01} = \frac{296+430}{257+370} \times 100$$

$$= \frac{726}{627} \times 100$$

$$= 115.78$$

(v) Fisher's Ideal method

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

Substituting the values in the formula,

$$P_{01} = \sqrt{\frac{296}{257} \times \frac{430}{370}} \times 100$$

$$= \sqrt{1.15 \times 1.16} \times 100$$

$$= \sqrt{1.334} \times 100$$

$$= 1.155 \times 100$$

$$= 115.5$$

## 29.3 WEIGHTED AVERAGE OF RELATIVE INDEX

The weighted average of relatives index is obtained by multiplying price relatives of different commodities by their respective assigned weights and averaging the total product by using either arithmetic mean or geometric mean.

(a) *When arithmethic mean is used* : The following procedure is adopted when arithmetic mean is used to compute weighted average of relatives index.

(i) Find out relatives for various commodities by expressing current year price as a percentage of base year price.

(ii) Calculate wieghts for various commodities by multiplying base year price with base year quantity.

(iii) Multiply the price relatives with weights.

(iv) Obtain the total of value weights.

(v) Divide the product of price relatives and value weights by total of value weights. The resulting value is the weighted average of relatives index.

The formula is

$$P_{01} = \frac{\Sigma PV}{\Sigma V}$$

Where,

P = Price relatives i.e., $\frac{p_1}{p_0} \times 100$

V = Value weights i.e., $p_0 q_0$.

$\Sigma PV$ = Sum of the products of price relatives and their respective value weights.

$\Sigma V$ = Sum of value weights.

**Illustration - 9**

Calculate weighted average of price relatives index by using arithmetic mean.

| Commodity | Base year 1975 | | Current year 1983 | |
|---|---|---|---|---|
| | Price | Quantity | Price | Quantity |
| A | 8 | 90 | 10 | 130 |
| B | 10 | 100 | 12 | 120 |
| C | 20 | 150 | 25 | 300 |
| D | 12 | 60 | 15 | 80 |

**Solution :**

CONSTRUCTION OF WEIGHTED AVERAGE OF PRICE RELATIVES INDEX BY USING ARTIHMETIC MEAN.

| Commodity | Base year 1975 | | Current year 1983 | | | V | |
|---|---|---|---|---|---|---|---|
| | $p_0$ | $q_0$ | $p_1$ | $q_1$ | $P = \left(\frac{p_1}{p_0} \times 100\right)$ | $(p_0q_0)$ | PV |
| A | 8 | 90 | 10 | 130 | 125 | 720 | 90,000 |
| B | 10 | 100 | 12 | 120 | 120 | 1000 | 1,20,000 |
| C | 20 | 150 | 25 | 300 | 125 | 3000 | 3,75,000 |
| D | 12 | 60 | 15 | 80 | 125 | 720 | 90,000 |
| | | | | | | $\Sigma V = 5440$ | $\Sigma PV = 6,75,000$ |

$$P_{01} = \frac{\Sigma PV}{\Sigma V}$$

Here, $\Sigma PV = 6,75,000$ and $\Sigma V = 5440$.

Subsituting the values in the formula,

$$P_{01} = \frac{6,75,000}{5440}$$

$$= 124.08$$

$\therefore$ Price index for 1983 = 124.08

**Illustration - 10**

An enquiry into the budget of middle class families in a village gave the following information.

| Expenses on | Food | Rent | Clothing | Fuel | Others |
|---|---|---|---|---|---|
| | 40% | 10% | 20% | 15% | 15% |
| Prices in 1981 | 80 | 20 | 80 | 10 | 50 |
| Prices in 1983 | 100 | 30 | 90 | 20 | 60 |

Compute price index number by' using weighted arithmetic mean of price relatives method.

**Solution**

COMPUTATION OF PRICE INDEX NUMBER

| Expenses on | Weights (V) | Prices in 1981 $p_0$ | Prices in 1983 $p_1$ | $P = \left(\frac{p_1}{p_0} \times 100\right)$ | PV |
|---|---|---|---|---|---|
| Food | 40 | 80 | 100 | 125 | 5,000 |
| Rent | 10 | 20 | 30 | 150 | 1,500 |
| Clothing | 20 | 80 | 90 | 112.5 | 2,250 |
| Fuel | 15 | 10 | 20 | 200 | 3,000 |
| Others | 15 | 50 | 60 | 120 | 1,800 |
| $\Sigma V = 100$ | | | | | $\Sigma PV = 13,550$ |

$$P_{01} = \frac{\Sigma PV}{\Sigma V}$$

Here, $\Sigma V = 100$ and $\Sigma PV = 13,550$

Substituting the values in the formula,

$$P_{01} = \frac{13,550}{100}$$

$$= 135.5$$

$\therefore$ Price index for 1983 = 135.5

**Illustration - 11**

From the data given below compute price index.

| Group | Index Number | Weights |
|---|---|---|
| Food | 350 | 50 |
| Fuel | 200 | 20 |
| Clothing | 240 | 10 |
| House Rent | 180 | 15 |
| Others | 200 | 20 |

Solution :

COMPUTATION OF PRICE INDEX

| Group | Index Number | Weights | |
|---|---|---|---|
| | I | V | IV |
| Food | 350 | 50 | 17,500 |
| Fuel | 200 | 20 | 4,000 |
| Clothing | 240 | 10 | 2,400 |
| House Rent | 180 | 15 | 2,700 |
| Others | 200 | 20 | 4,000 |
| | | $\Sigma V = 115$ | $\Sigma IV = 30{,}600$ |

$$P_{01} = \frac{\Sigma IV}{\Sigma V}$$

Here, $\Sigma V = 155$, $\Sigma IV = 30{,}600$

Substituting the values in the formula,

$$P_{01} = \frac{30{,}600}{115}$$

$$= 266.1$$

$\therefore$ Price index = 266.1.

*(b) When geometric mean is used :* Instead of arithmetic mean, if geometric mean is used to average the sum of the product of price relatives and weights, the following procedure is adopted.

(a) Find out price relatives.

(b) Obtain logarithmic values for various price relatives.

(c) Multiply log values of price relatives by their respective weights.

(d) Add various weights.

(e) Divide the product of log values of price relatives and weights by total weights.

(f) Find out antilogarithm value for the quotient obtained in step 'c'. The resultant one is the weighted geometric mean of price relatives index.

The formula is

$$p_{01} = Antilog \frac{\Sigma V log P}{\Sigma V}$$

Where,

P = Price relatives i.e., $\frac{p_1}{p_0} \times 100$

V = Value weights i.e., $p_0 q_0$

**Illustration - 12**

From the following data compute price index by weighted average of price relatives method using geometric mean.

| items | Base year 1980 Price (Rs.) | Quantity | Current year 1983 Price (Rs.) |
|---|---|---|---|
| Sugar | 3 | 30 | 4 |
| Rice | 4 | 50 | 5 |
| Wheat | 3 | 40 | 4 |
| Johar | 2 | 20 | 3 |
| Dal | 5 | 10 | 8 |

**Solution:**

CONSTRUCTION OF PRICE INDEX BY USING GEOMETRIC MEAN OF WEIGHTED PRICE RELATIVES.

| Items | Base year 1980 | | Current year 1983 | $(p_0q_0)$ | | | |
|---|---|---|---|---|---|---|---|
| | $p_0$ | $q_0$ | $p_1$ | V | P | logP | V logP |
| Sugar | 3 | 30 | 4 | 90 | 133.33 | 2.1249 | 191.241 |
| Rice | 4 | 50 | 5 | 200 | 125.00 | 2.0969 | 419.380 |
| Wheat | 3 | 40 | 4 | 120 | 133.33 | 2.1249 | 254.988 |
| Johar | 2 | 20 | 3 | 40 | 150.00 | 2.1761 | 87.044 |
| Dal | 5 | 10 | 8 | 50 | 160.08 | 2.2041 | 110.205 |
| | | | | $\Sigma V = 500$ | | | $\Sigma V log P = 1062.858$ |

$$P_{01} = Antilog \frac{\Sigma V log P}{\Sigma V}$$

Here, $\Sigma V log P = 1062.858$ and $\Sigma V = 500$.

Substituting the values in the formula,

$$P_{01} = \text{Antilog } \frac{1062.858}{500}$$

$$= \text{Antilog } 2.1257$$

$$= 133.6$$

$\therefore$ Price Index for 1983 = 133.6

## 29.4. QUANTITY INDEX NUMBERS

Quantity indices are used to measure average changes in quantities. These are useful to measure and compare the physical volume of goods produced or marketed or distributed in the given year with reference to any base year. Unlike price indices quantity indices are not widely used. Among quantity indices, the production indices are highly effective indicators of the level of production in a country. The quantity indices are constructed by using various formulae of price indices by interchanging q to p and p to q.

Thus,

(i) Laspeyres quantity index

$$Q_{01} = \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times 100$$

(ii) Paasche's quantity index

$$Q_{01} = \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1} \times 100$$

(iii) Fisher's quantity index

$$Q_{01} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}} \times 100$$



**Illustration - 13**

From the following information calculate quantity indices by using (a) Laspeyres method, (b) Paasche's method and (c) Fisher's Ideal method

| | Base Year 1981 | | Current Year 1983 | |
|---|---|---|---|---|
| Item | Price | Quantity | Price | Quantity |
| A | 4 | 40 | 5 | 50 |
| B | 3 | 80 | 4 | 100 |
| C | 5 | 50 | 8 | 50 |
| D | 8 | 40 | 10 | 50 |
| E | 10 | 30 | 12 | 40 |

**Solution :**

COMPUTATION OF QUANTITY INDICES FOR 1983

| | 1980 | | 1983 | | | | | |
|---|---|---|---|---|---|---|---|---|
| Item | $p_0$ | $q_0$ | $p_1$ | $q_1$ | $q_1p_1$ | $q_1p_1$ | $q_0p_1$ | $q_0p_0$ |
| A | 4 | 40 | 5 | 50 | 200 | 250 | 200 | 160 |
| B | 3 | 80 | 4 | 100 | 300 | 400 | 320 | 240 |
| C | 5 | 50 | 8 | 80 | 400 | 640 | 400 | 250 |
| D | 8 | 40 | 10 | 50 | 400 | 500 | 400 | 320 |
| E | 10 | 30 | 12 | 40 | 400 | 480 | 360 | 300 |
| | | | | | $\Sigma q_1p_0$ = 1700 | $\Sigma q_1p_1$ = 2270 | $\Sigma q_0p_1$ = 1680 | $\Sigma q_0p_0$ = 1270 |

Here, $\Sigma q_1p_0 = 1700$; $\Sigma q_1p_1 = 2270$;

$\Sigma q_0p_1 = 1680$; $\Sigma q_0p_0 = 1270$

(i) Laspeyres quantity index number for 1983

$$Q_{01} = \frac{\Sigma q_1p_0}{\Sigma q_0p_0} \times 100$$

Substituting the values in the formula ,

$$Q_{01} = \frac{1700}{1270} \times 100$$

$$= 133.8$$

(ii) Paasche's quantity index number for 1983

$$Q_{01} = \frac{\Sigma q_1p_1}{\Sigma q_0p_1} \times 100$$

Substituting the values in the formula,

$$Q_{01} = \frac{2270}{1680} \times 100$$

$$= 135.1$$

(iii) Fisher's quantity index for 1983

$$Q_{01} = \sqrt{\frac{\Sigma q_1p_0}{\Sigma q_0p_0} \times \frac{\Sigma q_1p_1}{\Sigma q_0p_1}} \times 100$$

Substituting the values in the formula,

$$Q_{01} = \sqrt{\frac{1700}{1270} \times \frac{2270}{1680}} \times 100$$

$$= \sqrt{1.338 \times 1.351} \times 100$$

$$= \sqrt{1.807638} \times 100$$

$$Q_{01} = 134.45$$

## 29.5 VALUE INDEX NUMBERS

The value of commodity is the product of its price and quantity, The value index can be obtained by expressing the total value of current year as percentage of the total value in the base year. It measures the changes in the actual values between two periods of comparison i.e., current year and base year.

Symbolically,

$$V = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} \times 100$$

Where, V = Value Index

$\Sigma p_1 q_1$ = Total values of all commodities in the current year

$\Sigma p_0 q_0$ = Total values of all commodities in the base year

When the values of current year and base year are given, the value index can be calculated by using the formula given below.

$$V = \frac{\Sigma V_1}{\Sigma V_0} \times 100$$

Where

V = Value index

$\Sigma V_1$ = Total values in current year

$\Sigma V_0$ = Total values in base year

**Illustration - 14**

From the following data, compute value index number:

| | Commodity I | | Commodity II | | Commodity III | | Commodity IV | |
|---|---|---|---|---|---|---|---|---|
| Year | Price | Quantity | Price | Quantity | Price | Quantity | Price | Quantity |
| 1975 | 4 | 10 | 10 | 10 | 8 | 4 | 5 | 15 |
| 1983 | 6 | 15 | 8 | 15 | 6 | 10 | 8 | 20 |

**Solution :**

CONSTRUCTION OF VALUE INDEX

| | 1975 | | 1983 | | | |
|---|---|---|---|---|---|---|
| Commodity | $p_0$ | $q_0$ | $p_1$ | $q_1$ | $p_1q_1$ | $p_0q_0$ |
| I | 4 | 10 | 6 | 15 | 90 | 40 |
| II | 10 | 10 | 8 | 15 | 120 | 100 |
| III | 8 | 4 | 6 | 10 | 60 | 32 |
| IV | 5 | 15 | 8 | 20 | 160 | 75 |
| | | | | | $\Sigma p_1q_1$ = 430 | $\Sigma p_0q_0$ = 247 |

$$V = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} \times 100$$

Here, $\Sigma p_1 q_1 = 430$ and $\Sigma p_0 q_0 = 247$

Substituting the values in the formula,

$$V = \frac{430}{247} \times 100$$

$$= 174.08$$

$\therefore$ Value index for 1983 = 174.08.

## 29.6 SUMMING UP

The weighted indices are computed by taking both prices and weights of commodities. The weights are assigned on the basis of their quantities. The weights may be production, consumption or distribution figures. The weighted index numbers can be constructed either by using aggregative method or average of relatives method. The weighted aggregative index numbers are computed with the help of methods enunciated by Laspeyres, Pasche's, Dorbish and Bowley Marshall-Edgeworth, Kelly's, Fisher, and Walsch.

## 29.7 CHECK YOUR PROGRESS : MODEL ANSWERS

1. Ascertain the following values:

$\Sigma p_1 q_0$, $\Sigma p_0 q_0$ , $\Sigma p_1 q_1$, and $\Sigma p_0 q_1$

These values are 1900, 1330, 1805 and 1240 respectively. Substitute these values in the following formula.

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

The asnwer is 144.2

## 29.8 MODEL EXAMINATION QUESTIONS

**A. Short Questions**

1. What is meant by quantity and value index numbers?
2. Distinguish between
   (a) Laspeyres and Paasche's index numbers
   (b) Marshal Edgworth and Fisher's Ideal Index

**B. Essay Questions**

3. What do you mean by weighting in the process of index number construction? Why is it necessary ? What are the commonly used weighting schemes?
4. Explain Fisher's Ideal Method of constructing Index numbers and comment on its utility.
5. 'Laspeyre's formula has an upward bias and Paasche's formula has a downward bias'. Explain.
6. What is Fisher's Ideal Index Number? Why is it called Ideal? Show how it satisfies time reversal and factor reversal test.?

EXERCISES

7. Calculate Fisher's Ideal index for the following data :

| | Base year 1980 | | Current year 1984 | |
|---|---|---|---|---|
| Commodities | Price | Qauntity | Price | Quantity |
| A | 6 | 50 | 10 | 60 |
| B | 2 | 100 | 2 | 120 |
| C | 4 | 60 | 2 | 60 |
| D | 10 | 30 | 12 | 30 |
| E | 15 | 60 | 20 | 50 |

(Ans : 122.71)

8. Annual production of four commodities is given below, compute quantity index numbers and comment on the result.

| | Production | | | |
|---|---|---|---|---|
| Commodities | 1978 | 1980 | 1983 | Weights |
| A | 160 | 200 | 250 | 20 |
| B | 24 | 40 | 50 | 30 |
| C | 50 | 75 | 70 | 13 |
| D | 250 | 170 | 160 | 17 |

9. You are given below the information relating to the four commodities. Construct (a) Laspeyres index, (b) Paasche's index, (c) Dorbish Bowley's index, (d) Marshall Edgeworth index and (e) Fisher's ideal index.

| | Base year 1982 | | Current year 1984 | |
|---|---|---|---|---|
| Group | Price | Quantity | Price | Quantity |
| A | 12 | 50 | 20 | 50 |
| B | 10 | 100 | 15 | 120 |
| C | 14 | 60 | 20 | 70 |
| D | 15 | 40 | 18 | 60 |

(Ans: a)145.39, b)143.48, c)144.44, d)144.35 e)144.33)

10. Construct the index numbers by using (i) Laspeyres method, (ii) Paasche's method (iii) Bowley's method. (iv) Fisher's Ideal index and (v) Marshall Edgeworth method from the following data.

| | Base year | | Current year | |
|---|---|---|---|---|
| Items | Price | Quantity | Price | Quantity |
| A | 2 | 20 | 3 | 40 |
| B | 5 | 40 | 6 | 80 |
| C | 4 | 80 | 8 | 100 |
| D | 8 | 90 | 10 | 120 |
| E | 10 | 20 | 12 | 40 |

(Ans : i. 140.54 ii. 137.5 iii. 139.02 iv. 139.01 v. 138.71)

11. Calculate index numbers from the following data by :

i) Laspeyre's method

ii) Paasche's method

iii) Marshall Edgeworth method

iv) Fisher's method

v) Dorbish and Bowley method

vi) Walsch's method

| | Base year | | Current year | |
|---|---|---|---|---|
| Commodities | Price.Rs. | Qty.kg. | Price Rs. | Qty. kg |
| A | 5 | 50 | 8 | 50 |
| B | 2 | 100 | 2 | 100 |
| C | 4 | 50 | 5 | 75 |
| D | 10 | 40 | 10 | 30 |
| E | 7 | 50 | 12 | 40 |

(Ans: i. 132.14 ii. 131.95 ii. 132.05 iv. 132.04 v. 132.05 vi. 132.1)

12. Calculate Index Number for 1988 on the basis of 1985 by Fisher's ideal formula, for the data given below.

| | 1985 | | 1988 | |
|---|---|---|---|---|
| Article | Price.Rs. | Qty.kg. | Price.Rs. | Qty.kg. |
| I | 5 | 10 | 4 | 12 |
| II | 8 | 6 | 7 | 5 |
| III | 6 | 3 | 3 | 5 |

(Ans : 76.9)

## 29.9 RECOMMENDED BOOKS


1. Gupta,S.P. : "Satistical Methods," Sultan Chand & Company, New Delhi.
2. Gupta,B.N. : "Statistics",Sahitya Bhavan, Agra.
3. Gupta, S.C : "Fundamentals of Statistics", Himalaya Publishing House, Bombay.
4. Simpson and Kafka : "Basic Statistics", Oxford and IBH. Publishing Company, Calcutta.


## 29.10 GLOSSARY

1. Weighted Aggregative Index Numbers : In this method, both the prices and quantities of the commodities are taken into account to construct the index numbers.
2. Weighted Average of Relatives Index Numbers : It is ascertained by multiplying price relatives of different commodities by their respective weights and averaging the total product by using either arithmetic mean or geometric mean.

## UNIT-30 : TESTS OF INDEX NUMBERS

**Contents**





## 30.0 AIMS AND OBJECTIVES

This unit aims at explaining the various tests of adequacy or consistency of index number formulae and describing the chain indices, base shifting, splicing and deflating of index numbers. On completion of this unit, you should be able to :

- prove the adequacy of the formula of index number by applying various tests
- explain the meaning of chain index numbers
- explain the terms like base shifting, splicing and deflating of index numbers.

## 30.1 INTRODUCTION

There are several formulae to construct index numbers. Generally, it is a problem to select the most appropriate formula among all these to measure the price level changes with any perfection. A number of mathematical tests have been designed to overcome this problem. According to these tests any formula which satisfies them can be considered perfect. If the formula does not satisfy these tests, it is regarded as an imperfect formula to measure the price

level changes. These mathematical tests are also known as "Tests of consistency or adequacy of an index formula or criteria for a good index number".

## 30.2 TESTS OF INDEX NUMBER FORMULAE

The following are the tests of adequacy of index number formulae.

30.2.1 Unit Test

30.2.2 Time Reversal Test

30.2.3 Factor Reversal Test

30.2.4 Circular Test

### 30.2.1 UNIT TEST

In this test, formulae used to compute an index number should be independent of units in which the prices and quantities of various commodities are quoted. For instance, the prices are quoted in rupees, whereas quantities are quoted in metres, litres, kilograms, etc., which are absolute units of measurement. If the index is constructed by using these absolute values, the result must be independent of them. This implies that the value of index number should be in relative terms i.e, percentages, rather than in absolute terms. Because the absolute values are not useful for comparison and drawing valid inferences unless they are expressed in relative terms. All the formulae used for constructing the index numbers, except simple aggregative method, satisfy this test.

### 30.2.2 TIME REVERSAL TEST

Time reversal test was propounded by prof. Irving Fisher to test time consistency of index number formulae. This test enables us to determine whether the formulae can work both ways in time i.e., backward and forward.

According to Fisher" The formula for calculating an index number should be such that it gives the same ratio between one point of comparison and the other, no matter which of the two is taken as a base" or putting it the other way, "The index number reckoned forward should be the reciprocal of that reckoned backward".

As per this test, the product of two indices computed by interchanging current year as a base year and base year as a current year should be unity. For example, the index for 1983 with 1980 as base and index for 1980 with 1983 as base and their porduct should be unity. If the product is not unity, it indicates that the formula used for the construction of index number does not satisfy the Time Reversal Test. It implies that there is an inherent bias in the formula.

The Time Reversal Test is symbolically expressed as

$$P_{01} \times P_{10} = 1$$

Where, $P_{01}$ = Index (without factor 100) for year '1' on year '0 as base;

$P_{10}$ = Index (without factor 100) for year '0' on year'1'as base.

For instance, if the price of a commodity has increased to Rs. 4 in 1983 as compared to Rs. 2 in 1980, it means that the prices of 1983 is 200 per cent of 1980 prices and the prices of 1980 is 50 per cent of 1983 prices. These two figures are reciprocals of one another and their product is equal to 1 i.e., $\frac{200}{100} \times \frac{50}{100} = 1$.

The following index number formulae satisfy the time reversal test.

(i) Fisher's ideal formula

(ii) Marshall-Edgeworth formula

(iii) Kelly's formula

(iv) Simple geometric mean of price relatives

(v) Simple aggregative index with fixed weights

(vi) The weighted geometric mean of price relatives

(vii) Walsch Price Index

However, the test is applied to Fisher's ideal Index formula only in this Unit.

Let us see how fisher's ideal formula satisfies the Time Reversal Test.

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$$

Changing the year '0' to 1' and '1' to '0'

$$P_{01} = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}}$$

$$P_{01} \times P_{10} = \sqrt{1}$$

$$= 1$$

Since, the product of $P_{01} \times P_{10}$ is unity, the Fisher's ideal Index satisfies the Time Reversal Test.

**Illustration - 1**

From the following data, compute Fisher's ideal Index and prove, whether it satisfies Time Reversal Test.

| | Base Year 1980 | | Current Year 1983 | |
|---|---|---|---|---|
| Commodity | Price (Rs.) | Quantity | Price (Rs.) | Quantity |
| A | 8 | 50 | 10 | 60 |
| B | 4 | 100 | 6 | 120 |
| C | 6 | 120 | 8 | 140 |
| D | 10 | 80 | 12 | 100 |
| E | 12 | 50 | 14 | 80 |

**Solution :**

COMPUTATION OF FISHER'S IDEAL INDEX

| Commodity | Base Year 1980 | | Current Year 1983 | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $p_0$ | $q_0$ | $p_1$ | $q_1$ | $p_0q_0$ | $p_0q_1$ | $p_1q_0$ | $p_1q_1$ |
| A | 8 | 50 | 10 | 60 | 400 | 480 | 500 | 600 |
| B | 4 | 100 | 6 | 120 | 400 | 480 | 600 | 720 |
| C | 6 | 120 | 8 | 140 | 720 | 840 | 960 | 1120 |
| D | 10 | 80 | 12 | 100 | 800 | 1000 | 960 | 1200 |
| E | 12 | 50 | 14 | 80 | 600 | 960 | 700 | 1120 |
| | | | | | $\Sigma p_0q_0$ = 2920 | $\Sigma p_0q_1$ = 3760 | $\Sigma p_1q_1$ = 3720 | $\Sigma p_1q_1$ = 4760 |

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

Here, $\Sigma p_0q_0 = 2920$ ; $\Sigma p_0q_1 = 3760$ ;

$\Sigma p_1q_0 = 3720$ ; $\Sigma p_1q_1 = 4760$.

Substituting the values in the formula,

$$P_{01} = \sqrt{\frac{3720}{2920} \times \frac{4760}{3760}} \times 100$$

$$= \sqrt{1.27 \times 1.27} \times 100$$

$$= 1.27 \times 100$$

$$= 127$$

$\therefore$ Price index for 1983 = 127.

Fisher's ideal Index satisfies the Time Reversal Test if

$$P_{01} \times P_{10} = 1$$

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$$

$$P_{10} = \sqrt{\frac{\Sigma p_0 q_0}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}}$$

Substituting the values in the formula,

$$P_{01} \times P_{10} = \sqrt{\frac{3720}{2920} \times \frac{4760}{3760} \times \frac{3760}{4760} \times \frac{2920}{3720}}$$

$$= \sqrt{1}$$

$$= 1$$

Hence, the Fisher's ideal index satisfies the Time Reversal Test.

### 30.2.3 FACTOR REVERSAL TEST

Another test suggested by prof. Irving Fisher to examine the consistency of index number formula is "Factor Reversal Test". According to this test, the product of price and quantity indices (without factor 100) should give true value ratio, when these two indices were computed for the same data and for the same current and base periods.

According to prof. Fisher "Just as our formula should permit the interchange of the two times without giving inconsistent result, so it ought to permit interchanging the prices and quantities without giving inconsistent results, i.e., the two results multiplied together should give the true value ratio". In other words the change in the prices multiplied by the change in the quantity should give us total change in the value i.e., true value.

Factor Reversal Test can be symbolically stated as

$$P_{01} \times Q_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

Where, $P_{01}$ = price index for the year '1' on the year '0' as base.

$Q_{01}$ = Quantity index for the year '1' on the year '0' as base.

$\Sigma p_1 q_1$ = Total value in the current year.

$\Sigma p_0 q_0$ = Total value in the base year.

If the product of the price and quantity indices is not equal to the true value ratio, it is assumed that there is an error in the formula. Except Fisher's Ideal Index, none of the formulae satisfies the Time Reversal Test.

Let us see how Fisher's Ideal Index satisfies the Factor Reversal Test.

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$$

Changing P to q and q to P.

$$Q_{01} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}}$$

$$P_{01} \times Q_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}}$$

$$= \sqrt{\frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} \times \frac{\Sigma q_1 p_1}{\Sigma p_0 q_0}}$$

$$= \sqrt{\left(\frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}\right)^2}$$

$$P_{01} \times Q_{01} = \sqrt{\frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}}$$

Since $P_{01} \times Q_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$, Factor Reversal Test is satisfied by the Fisher's ideal Index.

**Illustration - 2**

Using the following data construct Fisher's Ideal Index and show how it satisfies Factor Reversal Test.

| | Price (in (Rs.) per unit | | Number of units | |
|---|---|---|---|---|
| Items | Base period | Current period | Base period | Current period |
| A | 6 | 10 | 50 | 60 |
| B | 2 | 2 | 100 | 120 |
| C | 4 | 6 | 60 | 80 |
| D | 10 | 12 | 30 | 20 |
| E | 8 | 10 | 60 | 50 |

**Solution :**

CONSTRUCTION OF FISHER'S IDEAL INDEX

| | Base period | | Current period | | | | | |
|---|---|---|---|---|---|---|---|---|
| Items | $p_0$ | $q_0$ | $p_1$ | $q_1$ | $p_0q_0$ | $p_0q_1$ | $p_1q_0$ | $p_1q_1$ |
| A | 6 | 50 | 10 | 60 | 300 | 360 | 500 | 600 |
| B | 2 | 100 | 2 | 120 | 200 | 240 | 200 | 240 |
| C | 4 | 60 | 6 | 80 | 240 | 320 | 360 | 480 |
| D | 10 | 30 | 12 | 20 | 300 | 200 | 360 | 240 |
| E | 8 | 60 | 10 | 50 | 480 | 400 | 600 | 500 |
| | | | | | $\Sigma p_0q_0$ = 1520 | $\Sigma p_0q_1$ = 1520 | $\Sigma p_1q_0$ = 2020 | $\Sigma p_1q_1$ = 2060 |

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

Here, $\Sigma p_0 q_0 = 1520$; $\Sigma p_0 q_1 = 1520$ ;

$\Sigma p_1 q_0 = 2020$ : $\Sigma p_1 q_1 = 2060$.

Substituting the values in the formula,

$$P_{01} = \sqrt{\frac{2020}{1520} \times \frac{2060}{1520}} \times 100$$

$$= \sqrt{1.328 \times 1.355} \times 100$$

$$= \sqrt{1.79944} \times 100$$

$$= 134.1$$

∴Price index for current period = 134.1.

Factor Reversal Test is satisfied if

$$P_{01} \times Q_{01} = \sqrt{\frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}}$$

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$$

$$Q_{01} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}}$$

$$P_{01} \times Q_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}}$$

Substituting the values in the formula,

$$P_{01} \times Q_{01} = \sqrt{\frac{2020}{1520} \times \frac{2060}{1520} \times \frac{1520}{1520} \times \frac{2060}{2020}}$$

$$= \sqrt{\frac{2060}{1520} \times \frac{2060}{1520}}$$

$$= \frac{2060}{1520} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

Hence, Fisher's ideal Index satisfies the Factor Reversal Test.

While computing various kinds of indices, the Factor Reversal Test is completely ignored on account of its strong restrictions involved to prove this test.

**Check Your progress - 1**

Using the following data construct Fisher Ideal Index and show how it satisfies Factor Reversal Test.

| | 1986 | | 1988 | |
|---|---|---|---|---|
| Items | $(p_0)$ | $(q_0)$ | $(p_1)$ | $(q_1)$ |
| A | 8 | 40 | 9 | 50 |
| B | 4 | 80 | 4 | 90 |
| C | 5 | 60 | 6 | 70 |
| D | 9 | 20 | 10 | 20 |

## 30.2.4 CIRCULAR TEST

Circular Test first proposed by Westergaard and was highly favoured by C.M. Walsch who has named it "Circular Test". This is an extension of Time Reversal Test. According to this test, the formula used to compute the index should work in a circular manner. In order to satisfy this test, the product of price indices over a period of years should be unity. The indices for different years are computed without referring back to orginal base in each time. The circular test suggests that if an index is computed for year 'A' on the base year 'B' and for the year 'B' on the base year 'C' and for the year 'C' on the base year 'A' and the product of all these indices should be unity. For example, there are three indices such as $P_{01}, P_{12}, P_{20}$ the circular test is said to be satisfied if ;

$$P_{01} \times P_{12} \times P_{20} = 1$$

This test is satisfied only by simple geometric mean of price relatives, simple aggregative

method and fixed weights aggregative method. This test is not satisfied by other formulae used to compute indices such as Laspeyres, Paasche's, Dorbish and Bowley's and fisher's ideal Index.

The simple aggregative price index satisfies the Circular Test which is shown below :

$$P_{01} \times P_{12} \times P_{20} = 1$$

$$\therefore \frac{p_1}{p_0} \times \frac{p_2}{p_1} \times \frac{p_0}{p_2} = 1$$

Fixed weights aggregative method also satisfies the Circular Test which is shown below :

$$P_{01} \times P_{12} \times P_{20} = 1$$

$$\therefore \frac{p_1 q}{p_0 q} \times \frac{p_2 q}{p_1 q} \times \frac{p_0 q}{p_2 q} = 1$$

**Illustration - 3**

From the following data, show how simple aggregative index satisfies Circular Test ?

| Year | Prices of Commodities | | | | |
|---|---|---|---|---|---|
| | A | B | C | D | E |
| 1983 | 8 | 6 | 10 | 12 | 15 |
| 1980 | 6 | 4 | 8 | 10 | 12 |
| 1978 | 4 | 3 | 5 | 8 | 10 |

**Solution :**

| Commodities | Price (in Rs.) 1978 | Price (in Rs.) 1980 | Price (in Rs.) 1983 |
|---|---|---|---|
| A | 4 | 6 | 8 |
| B | 3 | 4 | 6 |
| C | 5 | 8 | 10 |
| D | 8 | 10 | 12 |
| E | 10 | 12 | 15 |
| | $\Sigma P_0 = 30$ | $\Sigma P_1 = 40$ | $\Sigma P_2 = 51$ |

In order to prove that simple aggregative index satisfies the Circular test, we should get,

$P_{01} \times P_{12} \times P_{20} = 1$

$$P_{01} = \frac{\Sigma p_1}{\Sigma p_0}$$

$$P_{12} = \frac{\Sigma p_2}{\Sigma p_1}$$

$$P_{20} = \frac{\Sigma p_0}{\Sigma p_2}$$

Here, $\Sigma p_0 = 30$, $\Sigma p_1 = 40$ and $\Sigma p_2 = 51$.

Substituting the values in the formula,

$$P_{01} \times P_{12} \times P_{20} = \frac{40}{30} \times \frac{51}{40} \times \frac{30}{51} = 1$$

$$\therefore \quad P_{01} \times P_{12} \times P_{20} = 1$$

Hence, the Simple Aggregative Method of Index satisfies the Circular Test.

**Illustration - 4**

Show with the help of the following data that fixed Weighted Aggregative Method satisfies the Circular Test.

| | Price (in Rs.) | | | |
|---|---|---|---|---|
| Commodities | 1976 | 1980 | 1983 | Weights |
| A | 20 | 30 | 40 | 21 |
| B | 30 | 40 | 50 | 23 |
| C | 20 | 30 | 40 | 12 |
| D | 10 | 20 | 30 | 21 |
| E | 40 | 50 | 60 | 23 |

**Solution :**

| | Prices (in Rs.) | | | | | | |
|---|---|---|---|---|---|---|---|
| Commodities | 1976 | 1980 | 1983 | Weights | | | |
| | $p_0$ | $p_1$ | $p_2$ | q | $p_0q$ | $p_1q$ | $p_2q$ |
| A | 20 | 30 | 40 | 21 | 420 | 630 | 840 |
| B | 30 | 40 | 50 | 23 | 690 | 920 | 1150 |
| C | 20 | 30 | 40 | 12 | 240 | 360 | 480 |
| D | 10 | 20 | 30 | 21 | 210 | 420 | 630 |
| E | 40 | 50 | 60 | 23 | 920 | 1150 | 1380 |
| | | | | | $\Sigma p_0q$ = 2480 | $\Sigma p_1q$ = 3480 | $\Sigma p_2q$ = 4480 |

The Fixed Weights Aggregative index satisfies the Circular Test if,

$$P_{01} \times P_{12} \times P_{20} = 1$$

i.e. $\frac{\Sigma p_1q}{\Sigma p_0q} \times \frac{\Sigma p_2q}{\Sigma p_1q} \times \frac{\Sigma p_0q}{\Sigma p_2q} = 1$

Here, $\Sigma p_0q = 2480$, $\Sigma p_1q = 3480$ and $\Sigma p_2q = 4480$.

Substituting the values in the fomula,

$$P_{01} \times P_{12} \times P_{20} = \frac{3480}{2480} \times \frac{4480}{3480} \times \frac{2480}{4480} = 1$$

$$\therefore \quad P_{01} \times P_{12} \times P_{20} = 1$$

Hence, the Fixed Weights Aggregative index satisfies the Circular Test.

## 30.3. THE CHAIN INDEX NUMBERS

According to Croxton and Cowden, "The Chain Index is one in which the figures for each year (or Sub-periods thereof) are first expressed as percentages of the preceding year. These percentages are then chained together by successive multiplication to form a chain index".

The need for using chain indices is due to the following reasons :

(i) Old commodities may be withdrawn from and new ones may be introduced into the market continuously. Therefore, it is desirable to revise the list of items and the system of weights of index numbers from time to time.

(ii) If the base year is quite distant form the current year, the comparison on the basis of fixed base indices may be unrealistic, unreliable and may sometimes mislead the conclusions.

In order to compute the chain indices, the following procedure is used :

(i) Express the prices of each year as percentage of the preceding year and obtain link relatives.

(ii) Chain together these link relatives by successive multiplication to form a chain index. Chain index of any year is the average link relative of that year multiplied by chain index of its preceding year divided by 100. This is shown in the following formula.

$$\text{Chain base index for current year} = \frac{\text{Current Year Link relative} \times \text{Preceding Year's chain base index}}{100}$$

**Illustration - 5**

From the following data of the wholesale prices of cement, compute chain base indices.

| Year | 1975 | 1976 | 1977 | 1978 | 1979 | 1980 | 1981 | 1982 | 1983 |
|---|---|---|---|---|---|---|---|---|---|
| Price of Cement (Rs.per quintal) | 70 | 50 | 60 | 75 | 70 | 50 | 40 | 60 | 75 |

**Solution :**

CONSTRUCTION OF BASE INDICES

| Year | Price of Cement (Rs.per quintal) | Link relatives $\left(\frac{P_1}{P_2} \times 100\right)$ | Chain indices |
|---|---|---|---|
| 1975 | 70 | – | 100.00 |
| 1976 | 50 | $\frac{50}{70} \times 100 = 71.42$ | $\frac{71.42 \times 100}{100} = 71.42$ |
| 1977 | 60 | $\frac{60}{50} \times 100 = 120.00$ | $\frac{120 \times 71.42}{100} = 85.70$ |
| 1978 | 75 | $\frac{75}{60} \times 100 = 125.00$ | $\frac{125 \times 85.70}{100} = 107.12$ |
| 1979 | 70 | $\frac{70}{75} \times 100 = 93.33$ | $\frac{93.33 \times 107.12}{100} = 99.97$ |
| 1980 | 50 | $\frac{50}{70} \times 100 = 71.42$ | $\frac{71.42 \times 99.97}{100} = 71.39$ |
| 1981 | 40 | $\frac{40}{50} \times 100 = 80.00$ | $\frac{80 \times 71.39}{100} = 57.11$ |
| 1982 | 60 | $\frac{60}{40} \times 100 = 150.00$ | $\frac{150 \times 57.11}{100} = 85.66$ |
| 1983 | 75 | $\frac{75}{60} \times 100 = 125.00$ | $\frac{125 \times 85.66}{100} = 107.07$ |

**Illustration - 6**

Compute chain indices from the following link relatives :

| Year | 1977 | 1978 | 1979 | 1980 | 1981 | 1982 | 1983 |
|---|---|---|---|---|---|---|---|
| Link Relatives | 100 | 120 | 110 | 125 | 130 | 115 | 95 |

**Solution :**

COMPUTATION OF CHAIN INDICES

| Year | Link relatives | Chain indices |
|---|---|---|
| 1977 | 100 | – = 100.00 |
| 1978 | 120 | $\frac{120 \times 100}{100} = 120.00$ |
| 1979 | 110 | $\frac{110 \times 120}{100} = 132.00$ |
| 1980 | 125 | $\frac{125 \times 132}{100} = 165.00$ |
| 1981 | 130 | $\frac{130 \times 165}{100} = 214.50$ |
| 1982 | 115 | $\frac{115 \times 214.5}{100} = 246.67$ |
| 1983 | 95 | $\frac{95 \times 246.67}{100} = 234.33$ |

**Illustration - 7**

Compute chain index numbers with 1978 as base relating to average wholesale prices of 4 commodities from 1978 to 1983.

Average wholesale prices (in Rs.)

| Commodities | 1978 | 1979 | 1980 | 1981 | 1982 | 1983 |
|---|---|---|---|---|---|---|
| A | 15 | 20 | 25 | 30 | 25 | 30 |
| B | 20 | 15 | 30 | 25 | 30 | 35 |
| C | 25 | 25 | 20 | 20 | 20 | 25 |
| D | 20 | 30 | 25 | 15 | 25 | 20 |

**Solution :**

COMPUTATION OF CHAIN INDICES - RELATIVES BASED ON PRECEDING YEAR.

AVERAGE WHOLESALE PRICES

| Commodity | 1978 | 1979 | 1980 | 1981 | 1982 | 1983 |
|---|---|---|---|---|---|---|
| A | 100 | $\frac{20}{15}\times 100 = 133.33$ | $\frac{25}{20}\times 100 = 125.00$ | $\frac{30}{25}\times 100 = 120.00$ | $\frac{25}{30}\times 100 = 83.33$ | $\frac{30}{25}\times 100 = 120.00$ |
| B | 100 | $\frac{15}{20}\times 100 = 75.00$ | $\frac{30}{15}\times 100 = 200.00$ | $\frac{25}{30}\times 100 = 83.33$ | $\frac{30}{25}\times 100 = 120.00$ | $\frac{35}{30}\times 100 = 116.67$ |
| C | 100 | $\frac{25}{25}\times 100 = 100.00$ | $\frac{20}{25}\times 100 = 80.00$ | $\frac{20}{20}\times 100 = 100.00$ | $\frac{20}{20}\times 100 = 100.00$ | $\frac{25}{20}\times 100 = 125.00$ |
| D | 100 | $\frac{30}{20}\times 100 = 150.00$ | $\frac{25}{30}\times 100 = 83.33$ | $\frac{15}{25}\times 100 = 60.00$ | $\frac{25}{15}\times 100 = 166.67$ | $\frac{20}{25}\times 100 = 80.00$ |
| Total of Link Relatives | 400 | 458.33 | 488.33 | 363.33 | 470.00 | 441.67 |
| Average Link Relatives | 100 | 114.58 | 122.08 | 90.83 | 117.5 | 110.42 |
| Chain Index Numbers 1978 = 100 | | $\frac{114.58\times 100}{100}$ | $\frac{122.08\times 114.58}{100}$ | $\frac{90.83\times 139.88}{100}$ | $\frac{117.5\times 127.05}{100}$ | $\frac{110.42\times 149.28}{100}$ |
| | | = 114.58 | = 139.88 | = 127.05 | = 149.28 | = 164.83 |

**Distinction between Chain Base and Fixed Base Indices**

The following are the differences between chain base and fixed base indices.

(i) Calculation of chain indices is very cumbersome whereas the fixed base are easily computed form the original data.

(ii) In chain indices, immediately preceding the year of each year is taken as base, whereas, in the case of fixed base indices, the base period is fixed and is chosen arbitrarity.

(iii) When new items are included and absolute items are deleted from the list of commodities, the Chain indices can be easily computed without recasting the entire calculations. Alternatively, in the case of fixed base method, the entire series of indices have to be recast when changes in the items occur.

**Merits**

The following are the merits of the Chain base indices :

(i) Comparison of economic and business data is made by using chain indices at one point of time to its previous period without any difficulty.

(ii) Chain index permits to calculate indices without recasting the calculations completely when there is a change in the commodities.

(iii) Weights can be adjusted as frequently as possible. This type of flexibility is very significant in the computation of several types of indices.

(iv) Chain base indices eliminate the impact of the seasonal forces.

**Limitations**

(i) Chain indices are very difficult to understand and cumbersome to compute.

(ii) Strictly speaking, the long range comparisons of chained percentages are not valid.

(iii) The chain indices for subsequent periods cannot be computed, if the data for any one year is missing.

(iv) An error in an index leads to wrong computation of the entire series of indices.

**Conversion of Chain Indices into Fixed Base indices**

The Chain base indices can be converted into fixed base indices by using the procedure given below :

(i) For the first Year, the chain base index will be taken as the fixed base index for that year. When index numbers are constructed by taking the first year as base, then the index for the frist year is taken as 100.

(ii) To calculate indices for other years, the following formula is used.

$$\text{Current year F.B.I.} = \frac{\textit{Current Year C.B.I.} \times \textit{Previous year F.B.I.}}{100}$$

Where, F.B.I. = Fixed base index number

C.B.I. = Chain base index number.

**Illustration - 8**

From the chain base index numbers given below, compute fixed base index numbers.

| Year | 1976 | 1977 | 1978 | 1979 | 1980 | 1981 | 1982 | 1983 |
|---|---|---|---|---|---|---|---|---|
| Chain indices | 90 | 110 | 120 | 115 | 130 | 120 | 150 | 140 |

**Solution :**

COMPUTATION OF FIXED BASE INDICES

| Year | Chain base indices | Fixed base indices |
|---|---|---|
| 1976 | 90 | – 90.00 |
| 1977 | 110 | $\frac{110\times90}{100}$ = 99.00 |
| 1978 | 120 | $\frac{120\times99}{100}$ = 118.8 |
| 1979 | 115 | $\frac{115\times118.8}{100}$ = 136.62 |
| 1980 | 130 | $\frac{130\times136.62}{100}$ = 177.61 |
| 1981 | 120 | $\frac{120\times177.61}{100}$ = 213.13 |
| 1982 | 150 | $\frac{150\times213.13}{100}$ = 319.69 |
| 1983 | 140 | $\frac{140\times319.69}{100}$ = 447.57 |

## 30.4 BASE SHIFTING

According to Morris Hamburg "For a variety of reasons, it is often necessary to change the reference base of an index number series from one time period to another without returning to the original raw data and recomputing the entire series. This change of reference base period is usually referred to as shifting the base".

There are two important resons for base shifting :

(i) The Previous base might have become too old or too distant from the current year and is not helpful to make meaningful and valid comparisons. By shifting the base, the index number series can be stated in terms of more recent base period.

(ii) Indices of different base periods are converted into common base to make valid and quick inferences pertaining to various economic and business variables.

The possible way for shifting the base is to recompute the entire series of indices with new base. But this is a very difficult process. Alternatively, a simple and approximation method is followed, according to which the index numbers of various years of old base are divided by index number corresponding to the new base period and multiplying the quotient with 100. This results in recasting of old base indices into new base indices.

**Illustration - 9**

You are given below price indices computed with 1975 as base. Shift the base from 1975 to 1980 and recast the indices.

| Year | 1975 | 1976 | 1977 | 1978 | 1979 | 1980 | 1981 | 1982 | 1983 |
|---|---|---|---|---|---|---|---|---|---|
| Index (1975=100) | 100 | 120 | 130 | 110 | 140 | 150 | 160 | 180 | 200 |

**Solution :**

BASE SHIFTING FROM 1975 TO 1980

| Year | Index | Index number 1980=100 |
|---|---|---|
| 1975 | 100 | $\frac{100}{150} \times 100 = 66.67$ |
| 1976 | 120 | $\frac{120}{150} \times 100 = 80.00$ |
| 1977 | 130 | $\frac{130}{150} \times 100 = 86.67$ |
| 1978 | 110 | $\frac{110}{150} \times 100 = 73.33$ |
| 1979 | 140 | $\frac{140}{150} \times 100 = 93.33$ |
| 1980 | 150 | $\frac{150}{150} \times 100 = 100.00$ |
| 1981 | 160 | $\frac{160}{150} \times 100 = 106.67$ |
| 1982 | 180 | $\frac{180}{150} \times 100 = 120.00$ |
| 1983 | 200 | $\frac{200}{150} \times 100 = 133.33$ |

The new series of indices with 1980 as base is obtained by dividing all indices by 150 i.e, the values of index for 1980 with old base and multiplying the quotient by 100.

For instance,

(i) Index for 1975 with 1980 as base

$$= \frac{Index\ for\ 1975}{Index\ for\ 1980} \times 100$$

i.e., $\frac{100}{150} \times 100 = 66.67$

(ii) Index for 1982 with 1980 as base

$$= \frac{Index\ for\ 1982}{Index\ for\ 1980} \times 100$$

i.e., $\frac{180}{150} \times 100 = 120.00$

## 30.5. SPLICING

Another application of the principle of base shifting is the technique of splicing the indices. Splicing involves combining two or more series of overlapping indices to obtain a single continuous series of indices enabling us to make valid and meaningful comparisons with a common base period.

According to Ya-Lun-Chou "When the weights of an index number becomes out of date, we may construct another index with new weights. Thus, two indices result on occasion, we may also wish to convert these two indices into a continuous series. The procedure employed for this conversion is called Splicing, which is mainly a problem of finding proportions".

The splicing of two or more series of indices with different base periods is done with the help of the following formula.

$$\text{Spliced index number} = \frac{\text{Index number of current year} \times \text{Old index of new base year}}{100}$$

**Illustration - 10**

The following are the two sets of indices 'A' with 1975 and 'B' with 1978 as base periods. The index number 'A' with 1975 as base discontinued in 1978. Splice the index 'B' with 'A'.

| Year | Index 'A' | Year | Index'B' |
|---|---|---|---|
| 1975 | 220 | 1978 | 100 |
| 1976 | 240 | 1979 | 105 |
| 1977 | 250 | 1980 | 120 |
| 1978 | 260 | 1981 | 90 |
| | | 1982 | 110 |
| | | 1983 | 120 |

**Solution :**

SPLICING OF INDEX B WITH INDEX A

| Year | Index A | Year | Index B | Index 'B' spliced to Index 'A' (1975 as base) |
|---|---|---|---|---|
| 1975 | 100 | 1978 | 100 | $\frac{260\times100}{100} = 260$ |
| 1976 | 240 | 1979 | 105 | $\frac{260\times105}{100} = 273$ |
| 1977 | 250 | 1980 | 120 | $\frac{260\times120}{100} = 312$ |
| 1978 | 260 | 1981 | 90 | $\frac{260\times90}{100} = 234$ |
| | | 1982 | 110 | $\frac{260\times110}{100} = 286$ |
| | | 1983 | 120 | $\frac{260\times120}{100} = 312$ |

The technique of splicing is very useful to join and compare new and old series of indices. But the accurate results of splicing are obtained by using geometric mean only rather than arithmetic mean. However, in practice, the geometric mean is not used owing to difficulties in computing the values. Though the artihmetic mean is simple to splice the indices the comparison

and interpretation of spliced indices over a long period of time will become extremely difficult due to the inclusion of items such as frozen food, clothing made from synthetic fibre, television and similar types of recently developed products in the later indices, where as the spliced indices for the earlier period did not include these items. In spite of this type of difficulties, still the splicing is the only frequently used practical method providing for comparability in similar variables measured by indices of different time periods.

## 30.6 DEFLATING

One of the most useful applications of price index numbers is deflating. It means adjusting series of rupee figures for price level changes. The technique of deflating is desirable, when inflationary tendencies are prevailing in the economy, to convert money values into real values. An increase in the prices of commodities over a period of years results in a fall in the purchasing power of money. On the other hand, a rise in the money or nominal income may not accompany a rise in the corresponding price index to get the real income. Thus the purchasing the real income due to inflationary situation. Therefore, it is necessary to adjust or correct nominal wages in accordance with the rise in the corresponding price index to get the real income. Thus the purchasing power of money is the reciprocal of the price index. In deflating , the original rupee figures are stated in terms of 'constant rupee values'. The real income is computed from money or nominal income with the help of the following formula :

$$\text{Real wages} = \frac{\textit{Money or nominal wages}}{\textit{Price index}} \times 100$$

$$\text{Real wage index or Real income index} = \frac{\textit{Index of money wages}}{\textit{Consumer price index}} \times 100$$

**Illustration - 11**

Following table gives the money wages and price index numbers. Compute the index number of real wages.

| Year | 1978 | 1977 | 1980 | 1981 | 1982 | 1983 | 1984 | 1985 | 1986 |
|---|---|---|---|---|---|---|---|---|---|
| Wages (in Rs.) | 150 | 180 | 200 | 230 | 250 | 275 | 300 | 325 | 350 |
| Price index numbers | 100 | 150 | 160 | 200 | 220 | 240 | 280 | 300 | 320 |

**Solution :**

COMPUTATION OF INDEX NUMBERS OF REAL WAGES

| Year | Wages (in Rs.) | Price Index Numbers | Real Wages | Real wages indices (1978 as base) |
|---|---|---|---|---|
| 1978 | 150 | 100 | $\frac{150}{100} \times 100 = 150$ | — 100.00 |
| 1979 | 180 | 150 | $\frac{180}{150} \times 100 = 120$ | $\frac{120 \times 100}{150} = 80.00$ |
| 1980 | 200 | 160 | $\frac{200}{160} \times 100 = 125$ | $\frac{125 \times 100}{150} = 83.33$ |
| 1981 | 230 | 200 | $\frac{230}{200} \times 100 = 115$ | $\frac{115 \times 100}{150} = 76.67$ |
| 1982 | 250 | 220 | $\frac{250}{220} \times 100 = 113.64$ | $\frac{113.64 \times 100}{150} = 75.76$ |
| 1983 | 275 | 240 | $\frac{275}{240} \times 100 = 114.58$ | $\frac{114.58 \times 100}{150} = 76.38$ |
| 1984 | 300 | 280 | $\frac{300}{280} \times 100 = 107.14$ | $\frac{107.14 \times 100}{150} = 71.43$ |
| 1985 | 325 | 300 | $\frac{325}{300} \times 100 = 108.33$ | $\frac{108.33 \times 100}{150} = 72.22$ |
| 1986 | 350 | 320 | $\frac{350}{320} \times 100 = 109.37$ | $\frac{109.37 \times 100}{150} = 72.91$ |

The monetary wages have increased from Rs. 150 in 1978 to Rs. 350 in 1986, whereas the real wages have declined from Rs. 150/- to Rs. 109.37 and the real wage indices have gone down from 100 to 72.91 respectivelly during the same period.

## 30.7 SUMMING UP

The adequacy and appropriateness of an index number formula can be tested by applying mathematical tests such as Unit Test, Time Reversal Test, Factor Reversal Test and Circular Test. According to Unit Test, the formula used to compute an index number should be independent of units in which the prices and quantities are quoted. As per the Time Reversal Test, the formula gives the same ratio between one point and comparison and the other no matter which of the two is taken as base. On the other hand, according to Factor Reversal Test, the product of price and quantity indices gives the true value ratio when these two indices are computed for the same data and for the same current and base periods. According to the circular test, the formula used to compute the index works in a circular manner, i.e., the product of price indices over a period of years is equal to unity.

## 30.8 CHECK YOUR PROGRESS : MODEL ANSWERS

| | | 1986 | | 1988 | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1. | Items | $p_0$ | $q_0$ | $p_1$ | $q_1$ | $p_1q_0$ | $p_0q_0$ | $p_1q_1$ | $p_0q_1$ |
| | A | 8 | 40 | 9 | 50 | 360 | 320 | 450 | 400 |
| | B | 4 | 80 | 4 | 90 | 320 | 320 | 360 | 360 |
| | C | 5 | 60 | 6 | 70 | 360 | 300 | 420 | 350 |
| | D | 9 | 20 | 10 | 20 | 200 | 180 | 200 | 180 |
| | | | | | | 1240 | 1120 | 1430 | 1290 |

Fisher Ideal Index $(P_{01}) = \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} \times 100 = 110.8$

Factor Reversal Test $(P_{01} \times Q_{01})$

$$= \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1} \times \frac{\Sigma q_1p_0}{\Sigma q_0p_0} \times \frac{\Sigma q_1p_1}{\Sigma q_0p_1}} = \frac{\Sigma p_1q_1}{\Sigma p_0q_0} = \frac{1430}{1120}$$

## 30.9. MODEL EXAMINATION QUESTIONS

**A.Short Questions**

1.What are chain base index numbers ?

2. What is meant by Base shifting ?

3.What is meant by Splicing ?

4. What is meant by Deflating ?

5. What is meant by Fixed base index number ?

6. What are the uses of Chain base index numbers ?

7. What do you mean by tests of consistency of an index number ?

8. Distinguish between Time Reversal and Factor Reversal Tests.

**B. Essay Questions**

9. Explain the Time Reversal and Factor Reversal Tests and prove them with a suitable illustration.

10.Discuss the advantages of chain indices over fixed base indices. What are their limitations?

EXERCISES

11. Shift the base years from 1975 to 1980 for the following indices :

| Year | 1975 | 1976 | 1977 | 1978 | 1979 | 1980 | 1981 | 1982 | 1983 |
|---|---|---|---|---|---|---|---|---|---|
| Index Numbers | 100 | 110 | 175 | 225 | 250 | 300 | 325 | 350 | 400 |

12. From the chain base index numbers given below prepare fixed base index numbers.

| Year | 1978 | 1979 | 1980 | 1981 | 1982 | 1983 | 1984 |
|---|---|---|---|---|---|---|---|
| Index Numbers | 80 | 110 | 120 | 105 | 125 | 130 | 135 |

13. You are given below two sets of indices one with 1975 as base and the other with 1980 as base, Splice index 'B' with index 'A'.

| Year | Index 'A' | Year | Index 'B' |
|---|---|---|---|
| 1975 | 100 | 1978 | 100 |
| 1976 | 110 | 1979 | 140 |
| 1977 | 120 | 1980 | 160 |
| 1978 | 125 | 1981 | 180 |
| | | 1982 | 200 |
| | | 1983 | 250 |

14. Construct Fisher's ideal index and show how it satisfies the Time Reversal and Factor Reversal Tests.

| | Base year 1980 | | Current year 1983 | |
|---|---|---|---|---|
| Commodities | Price | Quantity | Price | Quantity |
| A | 12 | 100 | 20 | 120 |
| B | 4 | 200 | 4 | 240 |
| C | 8 | 120 | 12 | 120 |
| D | 10 | 60 | 24 | 50 |
| E | 20 | 80 | 25 | 60 |

( Ans : fisher Index = 154.64, Factor Reversal Test = $\frac{7500}{4760}$)



## 30.10 RECOMMENDED BOOKS


1. Gupta, S.P. : 'Statistical Methods ', Sulthan Chand & Company, New Delhi.
2. Gupta,B.N. : 'Statistics',sahitya Bhavan, Agra.
3. Gupta, S.C. : 'Fundamentals of Statistics', Himalaya pub. House, Bombay.
4. Simpson and Kafka : 'Basic Statistics', Oxford and I.B.H. Publishing Company, Calcutta.

## 30.11 GLOSSARY

| | | |
|---|---|---|
| 1 | Circulat Test : | It is an extension of time reversal test for more than two periods and is based on the shiftability of base period. |
| 2. | Deflating : | It means adjusting series of rupee figures for price level changes. |
| 3. | Factor Reversal : Test | The change in the prices multiplied by the change in the quantity should be the total change in value i.e., real value. |
| 4. | Splicing : | It involves the construction of another index by shifting base. It is done by combining two or more overlapping series of index numbers to obtain a single continuous series. |
| 5. | Unit Test : | It states that the index number formula should be independent of the units in which the prices of quantities of various commodities that are expressed. |
| 6. | Time Reversal : Test | The formula for calculating an index number should be such that it gives the same ratio between one point of comparison and other, no matter which of the two is taken as a base. |

# UNIT-31 : COST OF LIVING INDEX NUMBERS

**Contents**





## 31.0 AIMS AND OBJECTIVES

The aims of this unit are to expalin the meaning, utility and methods of construction of cost of living idex numbers.

After going through this unit, you should be able to :

- explain the meaning of cost of living indices
- recognise the need for construction of cost of living indices
- describe the utility of cost of living indices
- explain the procedure for construction of cost of living indices
- list the methods of construction of cost of living indices.
- identify the precautions while using cost of living indices.

## 31.1 INTRODUCTION

Wholesale price index numbers measure the changes that take place in the general price level only. These index numbers may not reflect the true cost of living of different categories of people. As such there is a need to construct cost of living index numbers. They are also known

as consumer price index numbers. Consumer price index numbers measure the changes that take place in the cost of living at different time periods.

## 31.2 MEANING OF COST OF LIVING INDICES

Cost of living indices are also termed as 'Consumer Price Index Numbers' or 'Retail Price Index Numbers'. They are intended to measure the effect of changes in the prices of a group of goods and services on the purchasing power of a particular section or class of the society during any given period with reference to some fixed (base) period.

According to John I. Griffin 'Cost of living index is the ratio of the monetary expenditure of an individual which services for him the same 'Standard of living or total utility' in two situations differing only in respect of prices'. The objective of cost of living indices is to find out how much the consumers of a particular class have to pay more for a certain group of goods and services in a given period when compared with the base period.

## 31.3 NEED FOR CONSTRUCTION OF COST OF LIVING INDICES

The need for construction of cost of living indices arises due to the following reasons :

(i) The wholesale price indices measure the general price level changes only and they fail to reflect the effect of price level changes on the cost of living of different classes or groups of people in the society.

(ii) It is necessary to construct separate indices for different classes of people and also for different geographical regions such as town, city, rural area, urban area, hill area etc., because the variations in the prices affect these people differently from time to time.

(iii) The Consumption pattern of different people in the society differs from time to time and the proportions of consumption of same type of commodity vary significantly from period to period. Hence, such type of changes in the proportions and commodities usage is studied with the help of cost of living indices only.



## 31.4 UTILITY OF COST OF LIVING INDICES

The following are the uses of cost of living indices :

(i) Cost of living indices helps us to determine the effect of the rise and fall in the prices of different classes of consumers living in different areas.

(ii) They enable us to find out the purchasing power of money and real income.

(iii) These indices are used to deflate income and value series in national accounts.

(iv) The cost of living indices are used by the Government for the formulation of price policy, wage policy, rent control, taxation and geneal economic policies.

(v) the Central and State Governments, industrial and business houses use the cost of living indices to regulate the dearness allowance (D.A) and payment of bonus to the employees to meet the rise in the cost of living due to rise in the prices.

(vi) These indices are widely used in wage negotiations and wage contracts.

## 31.5 PROCEDURE FOR CONSTRUCTION OF COST OF LIVING INDICES

The procedure given below is followed to collect data and to construct the cost of living indices.

*(a) Scope and Coverage*

The frist step in the construction of cost of living index is to determine the class of people for whom the index is being constructed. For example, whether the index is related to industrial workers. agricultural workers, teachers, officers etc. The scope of the index is clearly defined for each such type of categories. For instance, Central Government officers or State government officers or all the officers taken together. Apart from this, the geographical area such as urban area, city, town or a locality of a town, hill area etc., of the people is to be defined.

*(b) Family Budget Enquiry*

Once the scope of the index is defined the next step is to conduct family budget enquiry. This is conducted by selecting an adequate number of representative families from the class of people for whom the index is being designed. The main aim of conducting family budget enquiry is to determine the amount, that an average family of the group included in the index, spends on different items of consumption. Generally, the family budget enquiry is conducted in a period of economic stability.

The enquiry should include the aspects such as the nature, quality and quantities of commodities consumed by the given class of people. In this case, the commodities are broadly classified into the following five major groups viz., (i) Food, (ii) Clothing, (iii) fuel and Lighting, (iv) House rent and (v) Miscellaneous. Each of these groups are further sub-divided into smaller groups and termed as 'Sub-groups. For instance, the food group is sub-divided into wheat, rice, pulses, sugar, etc. While conducting family budget enquiry, care is taken to include only those commodities which are geneally used by the people for whom the index is computed. With the help of family budget enquiry an average standard budget is prepared and such budget should not be affected by wide variations in quality, quantity and seasonal forces of short supply.

*(c) Obtaining Price Quotations*

The next step is to obtain price quotations of commodities. The price quotations are obtained from retail shops or local markets where the class of people reside or from where they are frequently purchasing their commodities. The collection of pr-e quotations is very imporant and at the same time it is a very difficult task, beacause price quotations vary from place to place and from shop to shop. While collecting retail price quotations, the folowing principles are considered.

(i) The retail prices should relate to a fixed list of items and for each item, the quality should be fixed by means of suitable specifications.

(ii) Retail prices should be those which are actually charged to customers.

(iii) If discount is given to all customers, it should be taken into account.

(iv) In a period of price controls and rationing, where illegal prices are charged, such illegal prices also must be taken into account along with the controlled prices.

## 31.6 METHODS OF CONSTRUCTION OF COST OF LIVING INDICES

The methods of construction of cost of living indices are brodally grouped into two categoies, viz.,

(a) Aggregate expenditure method or Weighted aggregative method.

(b) Family Budget method or the method of weighted relatives.

### 31.6.1 AGGREGATTIVE EXPENDITURE METHOD OR WEIGHTED AGGREGATIVE METHOD

In this method, cost of living index number is constructed by using Laspeyres' fromula. The quantities consumed in the base year are taken as weights. This method is widely used for construction of cost of living index numbers.

The procedure to compute cost of living index number is given below :

(i) Multiply the base year prices of various commodities by base year weights and obtain the current year aggregate expenditure .

(ii) Multiply the base year prices of various commodities by base year weights and obtain base year aggregate expenditure.

(iii) Divide the current year aggregate expenditure by base year aggregate expenditure and multiply the quotient with 100. The resulting value is cost of living index number.

Symbolically,

$$\text{Cost of Living Index Number} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$

$\Sigma p_1 q_0$ = Current year aggregate expenditure

$\Sigma p_0 q_0$ = Base year aggregate expenditure

**Illustration - 1**

From the data given below, compute the cost of living index number by using aggregate expenditure method.

| Commodities | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| Quantity in 1980 (in units) | 80 | 20 | 15 | 25 | 30 | 40 |
| Price per unit in 1980 (Rs.) | 6 | 5 | 8 | 20 | 16 | 10 |
| Price per unit in 1983(Rs.) | 8 | 9 | 10 | 25 | 20 | 25 |

**Solution :**

COMPUTATION OF COST OF LIVING INDEX NUMBER BY AGGREGATE EXPENDITURE METHOS.

| Commodities | Quantity 1980 $q_0$ | Price (in Rs.) 1980 $p_0$ | 1983 $p_1$ | Aggregate Expenditure $p_0q_0$ | $p_1q_0$ |
|---|---|---|---|---|---|
| A | 80 | 6 | 8 | 480 | 640 |
| B | 20 | 5 | 9 | 100 | 180 |
| C | 15 | 8 | 10 | 120 | 150 |
| D | 25 | 20 | 25 | 500 | 625 |
| E | 30 | 16 | 20 | 480 | 600 |
| F | 40 | 10 | 25 | 400 | 1000 |
| | | | | $\Sigma p_0q_0$ = 2080 | $\Sigma p_1q_0$ = 3195 |

Cost of Living Index Number $= \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$

Here, $\Sigma p_0 q_0 = 2080$ and $\Sigma p_1 q_0 = 3195$.

Substituting the values in the formula,

Cost of Living index number $= \frac{3195}{2080} \times 100$

$= 153.6$

$\therefore$ Cost of living index number for 1983 = 153.6.

**Illustration - 2**

Construct cost of living indices from the following data :

| Items | Food | Rent | Cloth. g | Fuel and Lighting | Others |
|---|---|---|---|---|---|
| Quantities in 1978 | 50 | 20 | 15 | 10 | 5 |
| Prices in (Rs.)1978 | 10 | 14 | 9 | 8 | 15 |
| Prices in (Rs.)1980 | 11 | 12 | 10 | 12 | 20 |
| Prices in (Rs.)1983 | 12 | 14 | 12 | 15 | 25 |

**Solution :**

Computation of cost of living indices for 1980 and 1983 with 1978 as base, by aggregate expenditure method.

| Items | Quantities in 1978 $q_0$ | Prices (in Rs.) 1978 $p_0$ | 1980 $p_1$ | 1983 $p_2$ | Aggregate Expenditure $p_0q_0$ | $p_1q_0$ | $p_2q_0$ |
|---|---|---|---|---|---|---|---|
| Food | 50 | 10 | 11 | 12 | 500 | 550 | 600 |
| Rent | 20 | 14 | 12 | 14 | 280 | 240 | 280 |
| Clothing | 15 | 9 | 10 | 12 | 135 | 150 | 180 |
| Fuel and Lighting | 10 | 8 | 12 | 15 | 80 | 120 | 150 |
| Others | 5 | 15 | 20 | 25 | 75 | 100 | 125 |
| | | | | | $\Sigma p_0q_0$ = 1070 | $\Sigma p_1q_0$ = 1160 | $\Sigma p_2q_0$ = 1335 |

(i) Cost of living index for 1980 $= \frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times 100$

Here, $\Sigma p_0q_0 = 1070$ and $\Sigma p_1q_0 = 1160$.

Substituting the values in the formula,

Cost of living index for 1980 $= \frac{1160}{1070} \times 100$

$= 108.4$

(ii) Cost of living index for 1983 $= \frac{\Sigma p_2q_0}{\Sigma p_0q_0} \times 100$

Here, $\Sigma p_2q_0 = 1335$ and $\Sigma p_0q_0 = 1070$.

Substituting the values in the formula,

Cost of living index for 1983 $= \frac{1335}{1070} \times 100$

$= 124.8$.

## 31.6.2 FAMILY BUDGET METHOD OR THE METHOD OF WEIGHTED RELATIVES

The construction of this index number is based on weigthed average of price relatives method. In this method, the value of prices and quantities in the base year are taken as weights. The price relatives are obtained by expressing current year prices of various commodities as percentage of base year prices. Further, the weighted price relatives are obtained by mulitiplying the price relatives with value weights. To obtain the cost of living index number divide the sum of the product of price relatives and value weights by total of value weights. This method is the same as the weighted average of price relatives method.

Symbolically,

Cost of Living Index Number $= \frac{\Sigma PV}{\Sigma V}$

where, $P = \frac{p_1}{p_0} \times 100$

$V$ = Value weights i.e., $p_0 \times q_0$.

$\Sigma PV$ = Sum of the products of price relatives and value weights.

$\Sigma V$ = Sum of the value weights.

**Illustration - 3**

From the following data, calculate cost of living index number by using family budget method for 1983 with 1975 as base year.

| Commodities | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| Quantity in units in 1975 | 100 | 25 | 10 | 20 | 25 | 30 |
| Price per unit in 1975 (Rs.) | 8 | 5 | 5 | 20 | 10 | 5 |
| Price per unit in 1983 (Rs.) | 12 | 7 | 5 | 12 | 16 | 10 |

**Solution :**

CONSTRUCTION OF COST OF LIVING INDEX NUMBER BY USING FAMILY BUDGET METHOD.

| Commodities | Quantity in units in 1975 $q_0$ | Price per unit (Rs.) 1975 $p_0$ | 1983 $p_1$ | $\left(\frac{p_1}{p_0} \times 100\right)$ P | $p_0q_0$ V | PV |
|---|---|---|---|---|---|---|
| A | 100 | 8 | 12 | $\frac{12}{8} \times 100 = 150$ | 800 | 120000 |
| B | 25 | 5 | 7 | $\frac{7}{5} \times 100 = 140$ | 125 | 17500 |
| C | 10 | 5 | 5 | $\frac{5}{5} \times 100 = 100$ | 50 | 5000 |
| D | 20 | 20 | 12 | $\frac{12}{20} \times 100 = 60$ | 400 | 24000 |
| E | 25 | 10 | 16 | $\frac{16}{10} \times 100 = 160$ | 250 | 40000 |
| F | 30 | 5 | 10 | $\frac{10}{5} \times 100 = 200$ | 150 | 30000 |
| | | | | | $\Sigma V =$ 1775 | $\Sigma PV =$ 236500 |

Cost of living index number for 1983 $= \frac{\Sigma PV}{\Sigma V}$

Here, $\Sigma V = 1775$ and $\Sigma PV = 2{,}36{,}500$.

Substituting the values in the formula,

Cost of living index number for 1983 $= \frac{236500}{1775}$

$= 133.24$

**Illustration - 4**

An enquiry into the budgets of Bank officers families in Hyderabad city gave the following information.

| Expenses on | Food | Fuel | Clothing | Rent | Misc. |
|---|---|---|---|---|---|
| | 40% | 15% | 25% | 10% | 10% |
| Prices in 1980 (Rs.) | 200 | 20 | 80 | 25 | 50 |
| Prices in 1983 (Rs.) | 250 | 25 | 100 | 30 | 60 |

Construct cost of living index number by using Family budget method.

**Solution :**

Construction of cost of living index number by using Family budget method.

In this problem, the percentage of expenditure on different items is taken as value weights.

| | Prices (in Rs.) | | $\left(\frac{p_1}{p_0} \times 100\right)$ | | |
|---|---|---|---|---|---|
| Expenses | 1980 | 1983 | | | |
| | $p_0$ | $p_1$ | P | V | PV |
| Food | 200 | 250 | 125 | 40 | 5000 |
| Fuel | 20 | 25 | 125 | 15 | 1875 |
| Clothing | 80 | 100 | 125 | 25 | 3125 |
| Rent | 25 | 30 | 120 | 10 | 1200 |
| Misc. | 50 | 60 | 120 | 10 | 1200 |
| | | | $\Sigma V = 100$ | | $\Sigma PV = 12400$ |

Cost of living index number $= \frac{\Sigma PV}{\Sigma V}$

Here, $\Sigma V = 100$ and $\Sigma PV = 12400$.

Substituting the values in the formula,

Cost of Living Index $= \frac{12400}{100}$

$= 124$

Cost of Living Index number $= 124$.

**Check Your Progress - 1**

Calculate cost of living index number from the following data.

| Items | food | rent | clothing | fuel | Mis. |
|---|---|---|---|---|---|
| Expenses | 20% | 10% | 25% | 35% | 10% |
| prices (Rs)'84 | 30 | 50 | 60 | 10 | 40 |
| Prices (Rs)'87 | 60 | 70 | 80 | 20 | 60 |

**Illustration - 5**

Construct cost of living index number from the data given below :

| Group | A | B | C | D | E |
|---|---|---|---|---|---|
| Index for 1980 | 525 | 220 | 200 | 175 | 250 |
| Expenditure | 45% | 20% | 10% | 15% | 10% |

**Solution :**

CONSTRUCTION OF COST OF LIVING INDEX

| Group | Index Number I | Expenditure V | IV |
|---|---|---|---|
| A | 525 | 45 | 23625 |
| B | 220 | 20 | 4400 |
| C | 200 | 10 | 2000 |
| D | 175 | 15 | 2625 |
| E | 250 | 10 | 2500 |
| | | $\Sigma V = 100$ | $\Sigma IV = 35150$ |



$$\text{Cost of living index number} = \frac{\Sigma IV}{\Sigma V}$$

Here, $\Sigma IV = 35150$ and $\Sigma V = 100$.

Substituting the values in the formula,

$$\text{Cost of Living Index Number} = \frac{35150}{100}$$

$$= 351.5$$

**Illustration - 6**

Construct the cost oof living index number using the weighted geometric mean of price relatives method.

| Group | Food | Fuel and Lighting | Clothing | Rent | Others |
|---|---|---|---|---|---|
| Index Number | 325 | 175 | 225 | 125 | 150 |
| Weights | 10 | 15 | 20 | 30 | 10 |

**Solution :**

CONSTRUCTION OF COST OF LIVING INDEX NUMBER

| Group | Index Number P | Weights V | log P | V log P |
|---|---|---|---|---|
| Food | 325 | 10 | 2.5119 | 25.119 |
| Fuel and Lighting | 175 | 15 | 2.2430 | 33.645 |
| Clothing | 225 | 20 | 2.3522 | 47.044 |
| Rent | 125 | 30 | 2.0969 | 62.907 |
| Others | 150 | 10 | 2.1761 | 21.761 |
| | | $\Sigma V = 85$ | | $\Sigma V \log P = 190.476$ |

$$\text{Cost of living index number} = \text{Antilog } \frac{\Sigma V \log P}{\Sigma V}$$

Here, $\Sigma V = 85$ and $\Sigma V \log P = 190.476$.

Substituting the values in the formula,

$$\text{Cost of living index number} = \text{Antilog } \frac{190.476}{85}$$

$$= \text{Antilog } 2.2408$$

$$= 174.1$$

## 31.7 PRECAUTIONS WHILE USING COST OF LIVING INDEX NUMBERS

The following precautions are to be taken while using and interpreting the cost of living index numbers, as they are generally misinterpreted.

(i) Usually consumer price indices measure changes in the retail prices in the given period as compared to base period. But these do not reflect the variations in living standards at two different places. For example, if the cost of living index of workers at Hyderabad is 200 and that of Madras is 300 for the same period and same class of people, it does not imply that the living costs are higher in Madras as compared to Hyderabad.

(ii) The cost of living indices are computed on the presumption that the group of goods does not change from year to year. But this presumption is seldom correct in practice. Because of shortages and changes in the prices of commodities the people change their buying habits from time to time.

(iii) **The data used to construct cost of living index is collected on the principle of sampling. Hence, the reliability of the index number depends on the correctness of the sampling technique. If the sample is not a representative of the class of people for whom the index is computed, naturally the data collected and index computed from it does not represent such a class of people.**

## 31.8 SUMMING UP

**The cost of Living Indices are designed to measure the effect of changes in the prices of group of commodities and services on the purchasing power of particcular section of people during a given period with reference to some base period, To compute cost of living indices, certain preliminaries such as the scope and coverage of the index, conducting family budget enquiry, obtaining price quotations, etc., are to be taken into consideration. The cost of living indices can be computed either by using Aggregative Expenditure Method or by Family Budget Method.**

## 31.9 CHECK YOUR PROGRESS : MODEL ANSWERS

| Expenses on | Prices (in Rs.) 1984 $p_0$ | 1987 $p_1$ | $\left(\frac{p_1}{p_0} \times 100\right)$ P | V | PV |
|---|---|---|---|---|---|
| Food | 30 | 60 | 200.00 | 20 | 4000.00 |
| Rent | 50 | 70 | 140.00 | 10 | 1400.00 |
| Clothing | 60 | 80 | 133.33 | 25 | 3333.25 |
| Fuel | 10 | 20 | 200.00 | 35 | 7000.00 |
| Miscellaneous | 40 | 60 | 150.00 | 10 | 1500.00 |
| | | | | 100 | 17233.25 |

$$\text{Cost of Living Index} = \frac{\Sigma PV}{\Sigma V} = \frac{17233.25}{100} = 172.3325$$

## 31.10 MODEL EXAMINATION QUESTIONS

**Essay Questions**

1. What is a cost of living index number ? What does it measure? Discuss its uses and limitations.
2. Describe the various steps involved in the construction of cost of living index numbers.
3. Explain the methods of constructing cost of living index numbers for the working classes in an industrial area of Hyderabad.
4. "Cost of living index number is essentially a consumer price index". Discuss.

## EXERCISES

5. From the following information compute real earnings :

| Year | Index of money earning | Consumer price index |
|---|---|---|
| 1978 | 100 | 100 |
| 1979 | 107 | 98 |
| 1980 | 110 | 96 |
| 1981 | 115 | 100 |
| 1982 | 120 | 105 |
| 1983 | 122 | 108 |
| 1984 | 130 | 110 |

6. The following table gives the per capita income and the cost of living index numbers. Deflate the per capita money income.

| Year | 1976 | 1977 | 1978 | 1979 | 1980 | 1981 | 1982 | 1983 |
|---|---|---|---|---|---|---|---|---|
| per capita income(Rs.) | 60 | 65 | 70 | 110 | 120 | 135 | 138 | 140 |
| Cost of living (1975=100) | 105 | 107 | 110 | 115 | 120 | 125 | 130 | 135 |

7. Compute cost of living index number from the following data.

| Group | Index Number | Weights |
|---|---|---|
| Food | 152 | 50 |
| Fuel and lighting | 130 | 10 |
| Clothing | 100 | 12 |
| House rent | 110 | 5 |
| Miscellaneous | 80 | 25 |

(Ans : 126.29)

8. Calculate the cost of living index number from the following data :

| Items | Weights | Price Relatives |
|---|---|---|
| A | 55 | 150 |
| B | 25 | 120 |
| C | 8 | 175 |
| D | 12 | 170 |

(Ans : 146.9)

**9. Calculate the cost of living index number from the following data :**

| Items | Price (in Rs.) 1980 | 1983 | Weight |
|---|---|---|---|
| Food | 30 | 45 | 4 |
| Fuel | 8 | 16 | 6 |
| Clothing | 14 | 21 | 10 |
| House rent | 20 | 25 | 20 |
| Miscellaneous | 25 | 30 | 30 |

(Ans : 134.28)

**10. From the data given below ; calculate the cost of living index number for the current year by the :**

**i) Aggredgate Expenditure Method, and ii) Family Budget Method**

| Article | Qty. Consumed in Base year | Price in Base year | Price in Current Year |
|---|---|---|---|
| Rice | 5Qtls | 24 | 30 |
| Wheat | 1 Qtls | 16 | 20 |
| Pulses | 2 " | 12 | 18 |
| Ghee | 4 Kg | 5 | 8 |
| Oil | 20 Kg | 40 | 50 |
| Clothing | 40 mtrs | 1 | 2 |
| Firewood | 10 Qtls | 2 | 4 |
| House Rent | — | 20 | 27 |



11. Construct cost of living index for 1988 based on 1980 from the following data :

| Group | Food | Housing | Clothing | Fuel& lighting | Misc. |
|---|---|---|---|---|---|
| Group Index No. for 1988 (based on 1980) | 122 | 140 | 112 | 116 | 106 |
| Weights | 32 | 10 | 10 | 6 | 42 |

(Ans : 114.88)

280

12. An enquiry into the budgets of the middle class families in a city of AP gave the following information :

| Expenses of | Food (35%) Rs. | Rent (15%) Rs. | Clothing (20%) Rs. | Fuel (10%) Rs. | Misc. (20%) Rs. |
|---|---|---|---|---|---|
| Prices in 80 | 150 | 30 | 75 | 25 | 40 |
| Prices in 89 | 145 | 30 | 65 | 23 | 45 |

What changes in cost of living figures of 1989 as compared with that of 1980 are seen ?

(Ans ; 97.86)

## 31. 11 RECOMMENDED BOOKS


1. Gupta, S.P. : 'Statistical Methods' , Sultan Chand & Company, New Delhi.
2. Gupta, B.N. : 'Statistics', sahitya Bhavan, Agra.
3. Gupta, S.C. : 'Fundamentals of Statistics', Himalaya Pub. House, Bombay.
4. Simpson and Kafka : ' Basic Statistics', Oxford and I.B.H. publishing company, calcutta.


## 31.12 GLOSSARY

Cost of Living Index : This measures the cahange in the cost of maintaining unchanged pattern of living of a particular group of persons that is, the change in the cost of consuming fixed quantities and qualities of goods and services. These index numbers relate to a particular class of people having similar consumption habits and pattern and to a definite region with more or less economic homogenity.

Family Budget method : Under this method, the prices and quantities in the base year are multiplied for each commodity. These are called weights. Then the price relatives are found as usual for each commodity. Later price relatives are multiplied with weights. The total of the product thus obtained is devided by total weights.

# APPENDIX :1

## LOGARITHMS

In logarithms, the value of a numbbr is expressed in terms of a common base of 10. The logarithm of any given number is the power to which the base (i.e 10) must be raised to obtain that number. Thus the logarithm or simply log of 10 is 1 because $(10) = 10$, log of 100 is 2 because $(10)^2 = 100$ and log of 1000 is 3 because $(10)^3 = 1000$. It is very easy to obtain the logarithms of such numbers as 10, 100, 1000, 10000. But it is not that easy to find out the logarithms of numbers like 84 or 184.2. In such cases we can consult the logarithm tables which are specially prepared for this purpose.

The logarithm of any given number comprises two parts viz., (1) the characteristic and (ii) the mantissa. While the characteristic of logarithm refers to the integral power of 10 of that number, the mantissa of logarithm refers to the fraction value.

**Determining the charateristic :**

To determine the characteristic of a number, we need not consult the logarithm tables. The characteristic of any number is determined by deducting 1 from the number of digits to the left side of the decimal point. Thus the characteristic of any number greate than 1 is always positive. On the other hand, the characteristic of any number less than 1 is always negative and is obtained by adding 1 to the number of zeros which follow the decimal point. The characteristic for some of the numbers is given below.

| Number | Characterstic |
|---|---|
| 9 | 0 |
| 95 | 1 |
| 956 | 2 |
| 9567 | 3 |
| 95676 | 4 |
| 0.9 | $\bar{1}$ |
| 0.09 | $\bar{2}$ |
| 0.009 | $\bar{3}$ |
| 0.0009 | $\bar{4}$ |
| 0.0802 | $\bar{2}$ |
| 0.9101 | $\bar{1}$ |

Conventionally the minus sign of the characteristic is written on the top of the figure and is read as bar 1, bar 2, bar 3, etc.

**Determining the Mantissa**

The Mantissa of any number is always positive and is not affected by the position of decimal point. Thus the value of mantissa of 956, 95.6, and 0.0956 will be the same.

To determine the value of the mantissa of a number, we consult the logarithm tables. These tables are useful to find out the mantissa of all numbers which contain 4 or less than 4 digits. Since it is not possible to obtain the value of the mantissa for the number which contains more than 4 digits, such numbers should be approximated and reduced to 4 digits. For example, to find out the mantissa of 4237.8, it should be approximated and written as 4238. Now the characteristic of this number would be 3 and to determine the mantissa, the frist two digits i.e. 42 should be located on the left hand vertical coloumn of logarithm tables and its corresponding value for the third digit i.e. 3 is taken from the horizontal columns (6263). To this figure, the 8th mean difference is (8) added. Thus the result (3.6271) would be the logarithm of 4237.8

Following are the logarithms of some of the numbers :

| Number | Logarithms |
|---|---|
| 3562 | 3.5516 |
| 356 | 2.5514 |
| 35.6 | 1.5514 |
| 3.56 | 0.5514 |
| 0.356 | $\bar{1}$.5514 |
| 0.036 | $\bar{2}$.5563 |
| 0.004 | $\bar{3}$.6021 |

**Antilogarithm**

The natural number of any logarithm is obtained by consulting the antilogarithm tables specially prepared for this purpose. To find out the antilogarithm of a logarithm we consult the table for the mantissa part only. After obtaining the antilogarithm value for a logarithm, the place of decimal is determined with the help of the characteristic i.e. the number of digits before the decimal point. We add 1 to the characteristic for finding out the natural number. The procedure for consulting antilog tables is the same as we follow in case of logarithms.

**Utility of Logarithms**

The utility of logarithms is clear from the following illustrations.

**1. Multiplication**

The product of two or more numbers is equal to the antilog of the sum of their respective logarithms :

Symbolically,

$A \times B = Antilog(LogA + LogB)$

**Illustration 1 :** Multiply 942 with 225.

**Solution :**

$942 \times 225$ = Antilog (Log 942 + Log 225)

= Antilog (2.9741 + 2.3522)

= Antilog 5.3263

= 221900

**Illustration 2 :** Multiply 32.07 with 2.136

**Solution :**

$32.07 \times 2.136$ = Antilog (Log 32.07 + Log 2.136)

= Antilog (1.5060 + 0.3296)

= Antilog 1.8356

= 68.48

**II. Division**

To divide one number by another, subtract the logarithm of latter number from the former number and find out the antilogarithm of the difference.

Symbolically,

$A \div B$ = Antilog (Log A—Log B )

**Illustration 3 :** Divide 936 by 42

**Solution :**

$\frac{936}{42}$ = Antilog (Log 936—Log 42)

= Antilog (2.9713—1.6232)

= Antilog 1.3481

=22.29

**III. Raising a number to a certain power**

To raise a number to a certain power,multiply the logarithm of the number by the exponent of the power and find out the antilogarithm of the product.

Symbolically,

$A^n$ = Antilog (n Log A)

**Illustration 4 :** Solve $(5.62)^4$

**Solution :**

$(5.62)^4$ = Antilog (4 log 5.62)

= Antilog (4 × 0.7497)

= Antilog 2.9988

= 997.2

**IV. Extracting the suare root of a number**

To find out the square root of a number, divide the logarithm of the number by 2 and find ut its antilogarithm.

Symbolically,

A = Antilog $\frac{1}{2}$(Log A)

**Illustration 5** : simplify $\sqrt{5231.6}$

**Solution** : $\sqrt{5231.6}$ = Antilog (1/2 log 5231.6)

= Antilog (1/2 × 3.7187)

= Antilog (1.8593)

= 72.33

**Illustration 6** : Solve $\frac{329.5}{\sqrt{232}\ \sqrt{115}}$

**Solution :**

$\frac{329.5}{\sqrt{232}\ \sqrt{115}}$ = Antilog (log 329.5−1/2 (log 232 + log115)

= Antilog (2.5179)− 1/2 (2.3655 + 2.0607)

= Antilog (2.5179)−1/2 (4.4262)

= Antilog (2.5179−2.2131)

=Antilog 0.3048

= 2.018



## RECIPROCALS

The reciprocal of a given number is defined as one divided by that number. Thus the reciprocal of 5 is $\frac{1}{5}$ = 0.2, reciprocal of 10 is $\frac{1}{10}$ = 0.1. In order to find out reciprocals, reciprocal tables are to be consulted. While consulting reciprocal tables for a given number, locate the firsit two digits in the left hand vertical column of the table and its corresponding figure for the third digit in the top horizontal column. From this the mean difference for the fourth digit should be subtracted to obtain the value of reciprocal. If the given number contains more than four digits, it must be approximated and reduced to four digits for the purpose of consulting reciprocal tables.

It should be noted that if the decimal point moves by one digit to the right in the give.. number it moves by one digit to the left in the reciprocal. The reciprocals of some of the numbers are given below :

| Number | Reciprocal | Number | Reciprocal |
|---|---|---|---|
| 7 | 0.1429 | 0.482 | 2.0747 |
| 9 | 0.1111 | 0.048 | 20.8333 |
| 15 | 0.0667 | 0.008 | 125.0000 |
| 20 | 0.0500 | 0.0008 | 1250.0000 |
| 325 | 0.0030 | | |

# APPENDIX - II

## SYMBOLS, ABBREVIATIONS AND FORMULAE

### *SYMBOLS AND ABBREVIATIONS*

| | | |
|---|---|---|
| A | = | Assumed Mean |
| A.M | = | Arithmetic Mean |
| $A_w$ | = | Assumed Weighted Mean |
| $b_{xy}$ | = | Regression coefficient of X on Y |
| $A_w$ | = | Assumed Weighted Mean |
| $b_{yx}$ | = | Regression coefficient of Y on X |
| C | = | Common factor |
| C. V. | = | Coefficient of Variation |
| C.f | = | Cumulative frequency |
| $d^1$ | = | $\frac{(X-A)}{C}$ , i.e., deviation of a value of X from an assumed mean taking a common factor |
| d | = | (X−A), i.e., deviation of a value of X from an assumed mean |
| /d/ | = | deviation of items from median or mean ignoring signs |
| $D_1, D_2, D_3$, etc. | = | 1 st , 2nd, 3rd deciles, etc. |
| f | = | frequency |
| G.M | = | Geometric Mean |
| W.G.M. or $G.M._w$ | = | Weighted Geometric Mean |
| H.M. | = | Harmonic Mean |
| $H.M_w$ | = | Weighted Harmonic Mean |
| i | = | Class interval |
| Log | = | Logarithm |
| L | = | Lower limit of a class |
| m | = | Mid-point of a class |
| Med | = | Median |
| M.D. | = | Mean Deviation |
| N | = | Number of observations or sum of frequency, i.e., $\Sigma f$ |
| $P_1, P_2, P_3$ etc. | = | 1st, 2nd, 3rd Percentiles, etc. |

| | | |
|---|---|---|
| $p_0, p_1$ | | prices in the base year and prices in the current year respectives |
| $Q_1, Q_2, Q_3,$ | = | 1st, 2nd, 3rd Quartiles |
| $Q_{01}$ | = | Quantity index for current year on the basis of base year |
| $q_0$ | = | Quantity in the base year |
| $q_1$ | = | Quantity in the current year |
| Q. D. | = | Quartile Deviation |
| R | = | Range |
| r | = | Coefficient of correlation |
| $r_k$ | = | Rank correlation |
| $r_c$ | = | Correlation by concurrent deviation method |
| $r^2$ | = | Coefficient of Determination |
| $SK_P$ | = | Karl Pearson's coefficient of skewness |
| $SK_B$ | = | Bowley's coefficient of skewness |
| V | = | Variance |
| W | = | Weights |
| $\overline{X}$ | = | Arithmetic Mean |
| x | = | $(X-\overline{X})$,i.e., deviations of items from actual mean |
| $\overline{X}_{12}$ | = | Combined Mean of two series |
| $\overline{X}_w$ | = | Weigthed arithmetic mean |
| $\overline{Y}$ | = | Arithmetic Mean of Y series |
| Y | = | $(Y-\overline{Y})$, i.e, deviations of items from actual mean |
| Z | = | Mode |

$\sigma$ (pronounced as sigma) = Standard deviation

$\Delta$ Pronounced as 'Delta'

$\mu$ Pronounced as 'Mu'

$\Sigma$ (Pronounced as sigma) = summation or the sum of.

## FORMULAE

## MEASURES OF CENTRAL TENDENCY

**Arithmetic Mean**

Individual Series

$\overline{X} = \frac{\Sigma X}{N}$ (Direct Method)

$\overline{X} = A + \frac{\Sigma dx}{N}$ (short-cut Method)

Discrete Series

$\overline{X} = \frac{\Sigma fX}{N}$ (Direct Method)

$\overline{X} = A + \frac{\Sigma fdx}{N}$ (Short-cut Method)

Continuous Series

$\overline{X} = \frac{\Sigma fm}{N}$ (Direct Method)

$\overline{X} = A + \frac{\Sigma fd}{N}$ (Short-cut Method)

$\overline{X} = \text{A} + \frac{\Sigma fd^1}{N} \times \text{C}$ (Step deviation method)

$\overline{X} = m - i(F - 1)$ (Summation Method)

Combined Mean

$\overline{X}_{12} = \frac{N_1\overline{X}_1 + N_2\overline{X}_2}{N_1 + N_2}$

**Weighted Arithmetic Mean :**

$\overline{X}_w = \frac{\Sigma WX}{\Sigma W}$ (Direct Method)

$\overline{X}_w = A_w + \frac{\Sigma Wdx}{\Sigma W}$ (short-cut Method)

**Median :**

Individual and Discrete Series

Median = Size of $\frac{N+1}{2}$ th item

Continuous Series

Median $= \text{L} + \frac{\frac{N}{2} - C.f}{f} \times \text{i}$

or $\text{U} - \frac{\frac{N}{2} - C.f}{f} \times \text{i}$

**Quartiles :**

Individual and Discrete Series

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

$Q_3$ = Size of $3\left(\frac{N+1}{4}\right)$ th item

Continuous Series

$Q_1 = \text{L} + \frac{\frac{N}{4} - C.f}{f} \times \text{i}$

$$Q_3 = L + \frac{3\frac{N}{4} - C.f}{f} \times i$$

**Mode :**

Continuous Series

$$Z = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

$$or\, Z = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

**Geometric Mean :**

Individual Series

$$G.M. = \text{Antilog} \quad \frac{\Sigma log X}{N}$$

Discrete Series

$$G.M. = \text{Antilog} \quad \frac{\Sigma f log X}{N}$$

Continuous Series

$$G.M. = \text{Antilog} \quad \frac{\Sigma f log m}{N}$$

**Weighted Geometric Mean :**

$$G.M_w = \text{Antilog} \quad \left(\frac{\Sigma W log X}{\Sigma W}\right)$$

**Harmonic Mean :**

Individual Series

$$H.M. = \frac{N}{\Sigma\left(\frac{1}{X}\right)}$$

Discrete Series

$$H.M. = \frac{N}{\Sigma\left(\frac{f}{X}\right)}$$

Continuous Series

$$H.M. = \frac{N}{\Sigma\left(f \times \frac{1}{m}\right)} \quad \text{or} \quad \frac{N}{\Sigma\left(\frac{f}{X}\right)}$$

**Weighted Harmonic Mean :**

$$H.M._w = \frac{\Sigma W}{\Sigma\left(\frac{W}{X}\right)}$$

**MEASURES OF VARIATION**

$$\text{Range} = L - S$$

$$\text{Coefficient of Range} = \frac{L-S}{L+S}$$

$$\text{Inter Quartile Range} = Q_3 - Q_1$$

Quartile Deviation : Q.D. $= \frac{Q_3 - Q_1}{2}$

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

**Mean Deviation :**

Individual Series

M.D. $= \frac{\Sigma |d|}{N}$ (Direct Method)

M.D. $= \frac{\Sigma X_A - \Sigma X_B - (n_A - n_B)\ Average}{N}$ (Short-cut Method)

Discrete and continuous series

M.D. $= \frac{\Sigma f|d|}{N}$

Coefficient of M.D. $= \frac{Mean\ Deviation}{Median\ or\ Mean}$

**Standard Deviation :**

Individual Series

$\sigma = \sqrt{\frac{\Sigma x^2}{N}}$ (Direct Method)

$\sigma = \sqrt{\frac{\Sigma dx^2}{N} - \left(\frac{\Sigma dx}{N}\right)^2}$ (Short-cut Method)

Discrete Series

$\sigma = \sqrt{\frac{\Sigma fd^2}{N}}$ (Direct Method)

$\sigma = \sqrt{\frac{\Sigma fdx^2}{N} - \left(\frac{\Sigma fdx}{N}\right)^2}$ (Short-cut Method)

$\sigma = \sqrt{\frac{\Sigma fd^{1^2}}{N} - \left(\frac{\Sigma fd^1}{N}\right)^2} \times C$ (Step deviation Method)

Coefficient of Standard deviation $= \frac{\sigma}{\bar{X}}$

Coefficient of variation

C.V. $= \frac{\sigma}{\bar{X}} \times 100$

Variance $= \frac{\Sigma(X - \bar{X})^2}{N}$ or variance $= \sigma^2$

$\sigma = \sqrt{variance}$

Combined Standard deviation

$\sigma_{12...n} = \frac{N_1\sigma_1^2 + N_2\sigma_2^2 ... N_n\sigma_n^2 + N_1d_1^2 + N_2d_2^2 ... N_nd_n}{N_1 + N_2 + ... N_n}$

## SKEWNESS

**Absolute Skewness**

$$\text{Mean} - \text{Mode}$$

$$\text{or } Q_3 + Q_1 - 2 \text{ Median}$$

Relative Skewness

(i) Coefficient of SK $= \frac{Mean - Mode}{\sigma}$

....(Karl pearson's Method)

(ii) Coefficient of SK $= \frac{3(Mean - Median)}{\sigma}$

(When Mode is ill defined)

(iii) Coefficient of SK $= \frac{Q_3+Q_1-2Median}{Q_3-Q_1}$

(Bowley's Method)

## CORRELATION ANALYSIS

*(i) Karl Pearson's Method*

(a) When deviations are taken from actual mean

$$r = \frac{\Sigma xy}{N\sigma x\sigma y}$$

(or)

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2.\Sigma y^2}}$$

(b) When deviations are taken from assumed mean

$$r = \frac{\Sigma dxdy - \frac{\Sigma dx.\Sigma dy}{N}}{\sqrt{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}}\sqrt{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}}$$

(c) Coefficient of correlation in grouped data

$$r = \frac{\Sigma fdxdy - \frac{\Sigma fdx.\Sigma fdy}{N}}{\sqrt{\Sigma fdx^2 - \frac{(\Sigma fdx)^2}{N}.\Sigma fdy^2 - \frac{(\Sigma fdy)^2}{N}}}$$

$$\text{P.E}_r = 0.6745\ \frac{1-r^2}{\sqrt{N}}$$

(ii) Rank correlation :

(a) $r_k = 1 - \frac{6\Sigma D^2}{N^3 - N}$

(b $r_k = 1 - \frac{6\{\Sigma D^2 + \frac{1}{12}(m^3 - m)\}}{N^3 - N}$ (When ranks are repeated)

(iii) Concurrent Deviation Method

$$r_c = \pm\sqrt{\pm\left(\frac{2C-N}{N}\right)}$$

## REGRESSION ANALYSIS

Regression line of Y on X is given by

$$Y = a+bx$$

Regression line of X on Y is given by

$$X = a + by$$

Regression coefficeints :

(i) By solving normal equations

Regression coefficient of X on Y

$$\Sigma X = Na + b\Sigma Y$$

$$\Sigma XY = a\Sigma X + b\Sigma Y^2$$

Regression coefficient of X on Y i.e., bxy or $b_1$

$= r\frac{\sigma x}{\sigma y} = \frac{\Sigma xy}{\Sigma Y^2}$ (When deviations are taken from actual mean )

$bxy = \frac{\Sigma dxdy - \frac{\Sigma dx.\Sigma dy}{N}}{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}$ ( When deviations are taken from assumed mean)

$bxy = \frac{\Sigma XY - N\overline{XY}}{\Sigma Y^2 - N(\overline{Y})^2}$ (When figures are given in original values )

Regression coefficient of Y on X i.e, byx or $b_2$

$= r\frac{\sigma y}{\sigma x} = \frac{\Sigma xy}{\Sigma x^2}$ (When deviations are taken from actual mean )

byx $= \frac{\Sigma dxdy - \frac{\Sigma dx.\Sigma dy}{N}}{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}}$ ( When deviations are taken from assumed mean)

byx $= \frac{\Sigma XY - N\overline{XY}}{\Sigma Y^2 - N(\overline{X})^2}$ (When figures are given in original values )

$r = \sqrt{bxy \times byx}$

## INDEX NUMBERS

**1. Unweigthed Index Numbers :**

(a) Simple Aggregate Method

$$P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100$$

**(b) Simple Average of Price Reltive Method**

$$P_{01} = \frac{\Sigma P}{N}$$

$$\text{Where, } P = \frac{p_1}{p_0} \times 100$$

**(c) Geometric Mean of price Relative Method**

$$P_{01} = \text{Antilog} \quad \frac{\Sigma log P}{\Sigma N}$$

**2. Weighted Index Numbers :**

**(a) Weigthed Aggregate Method**

**(i) Laspeyer's Method**

$$P_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$

**(ii) Paasche's method**

$$P_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$$

**(iii) Dorbish and Bowley's Method**

$$P_{01} = \frac{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}{2} \times 100$$

$$\text{or } P_{01} = \frac{L+P}{2}$$

**(iv) Fisher's ideal method**

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

(v) Marshall - Edgeworth's Method

$$P_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$$

(vi) Kell's Method

$$P_{01} = \frac{\Sigma p_1 Q}{\Sigma p_0 Q} \times 100$$

(b) Weighted Average of price Relatives

$$P_{01} = \left(\frac{\Sigma log PV}{\Sigma V}\right)$$

$$\text{or } \left(\frac{\Sigma log \frac{p_1}{p_0} \times 100 \times V}{\Sigma V}\right)$$

**3. Quantity or Volume Index Numbers**

$$Q_{01} = \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times 100$$ (When Laspeyre's method is used)

$$Q_{01} = \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1} \times 100$$ (When Paasche's method is used )

$$Q_{01} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}} \times 100$$ (When Fisher's method is used)

Time Reversal Test

$P_{01} \times P_{10} = 1$

$$P_{01} \times P_{10} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}} = 1$$

Factor Reversal Test

$$P_{01} \times Q_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

$$P_{01} \times Q_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}}$$

$$= \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

Circular Test :

$$P_{01} \times P_{12} \times P_{20} = 1$$

**4. Consumer Price Index :**

(a) Aggregate Expenditure Method

Cost of Living Index $= \frac{\Sigma P_1 q_0}{\Sigma P_0 q_0} \times 100$

(b) Family Budget Method

Cost of Living Index $= \frac{\Sigma PV}{\Sigma V}$

**ANDHRA PRADESH OPEN UNIVERSITY**

**B . COM . IIND YEAR**

**COMMERCE : GROUP-B : COURSE-I**

**BUSINESS STATISTICS**

**SYLLABUS**

BLOCK - I : STATISTICS - COLLECTION - CLASSIFICATION AND PRESENTATION

Unit - 1 : Origin, Growth and Definition of Statistics

Unit - 2 : Scope, Importance and Limitations of Statistics

Unit - 3 : Nature of Data

Unit - 4 : Collection of Data

Unit - 5 : Sampling Techniques

Unit - 6 : Classification of Data

Unit - 7 : Seriation of Data

Unit - 8 : Tabulation of Data

Unit - 9 : Diagrammatic presentation of Data

Unit -10 : Graphic Presentation of Data

BLOCK -II : MEASURES OF CENTRAL TENDENCY

Unit -11 : Introduction to Averages

Unit -12 : Artihmetic Mean

Unit -13 : Median and Quartiles

Unit -14 : Mode

Unit -15 : Geometirc and Harmonic Mean

ANDHRA PRADESH OPEN UNIVERSITY

B . COM . IIND YEAR

COMMERCE : GROUP-B : COURSE-I

# BUSINESS STATISTICS

## SYLLABUS

**F A C U L T Y O F C O M M E R C E**

B.Com. IInd Year (3YDC) Examination

MODEL QUESTION PAPER

GROUP -B : PAPER - I : BUSINESS STATISTICS

Time = 3 Hours

Max.Marks=100

Min.Marks=35

PART - A : Essay Questions

(Marks- 4 × 15 = 60 )

PART - A

Answer any four of the following

1. What do you mean by statistics and explain the limitations of statistics ?

2. Distiguish between dispersion and skewness ?

3. Briefly explain the problems to be faced in the construction of Index Numbers ?

4. Determine Median of the following data by graphic method.

| Class Interval | Frequency |
|---|---|
| 0 - 10 | 6 |
| 10 - 20 | 10 |
| 20 - 30 | 25 |
| 30 - 40 | 28 |
| 40 - 50 | 13 |
| 50 - 60 | 8 |

5. Calculate standard deviation and co-efficient of variation of the following data.

| Weights | Number of boys |
|---|---|
| 40 | 12 |
| 44 | 16 |
| 48 | 17 |
| 50 | 19 |
| 52 | 10 |
| 56 | 8 |
| 60 | 3 |

6. Calculate Skewness of the following data by Karl pearson's method

| wages(Rs) | No. of Workers |
|---|---|
| 250 - 300 | 4 |
| 300 - 350 | 5 |
| 350 - 400 | 8 |
| 400 - 450 | 12 |
| 450 - 500 | 9 |
| 500 - 550 | 4 |
| 550 - 600 | 3 |



7. Calculate coefficient of correlation of the following data

| | | | | | | |
|---|---|---|---|---|---|---|
| X | 12 | 14 | 16 | 20 | 24 | 22 |
| Y | 14 | 15 | 17 | 19 | 21 | 20 |

8. From the following information, calculate Fisher's Ideal Index.

| | Base Year 1985 | | Current Year | |
|---|---|---|---|---|
| Item | Price | Quantity | Price | Quantity |
| A | 4 | 40 | 5 | 50 |
| B | 7 | 30 | 7 | 40 |
| C | 6 | 70 | 7 | 60 |
| D | 8 | 60 | 9 | 70 |
| E | 10 | 80 | 11 | 90 |

## PART - B : SHORT QUESTIONS

(Marks $5 \times 8 = 40$)

Answer any four of the following

1. Distinguish between primary data and Secondary data.

2. Briefly explain the requisites of a good average .

3. What do you mean by 'Range'? List out the merits and limitations of it ?

4. What are the features of 'Symmetrical' distribution ?

5. Distinguish between correlation and regression .

6. Calculate Artihmetic Mean of the following data

| Class Interval (Marks in statistics) | 0-20 | 20-40 | 40-60 | 60-80 | 80-100 |
|---|---|---|---|---|---|
| No. of Students | 5 | 8 | 16 | 9 | 2 |

7. Following data relate to the daily turnover of two factories. Find out which factory is more consistent.

| | Factory-A | Factory-B |
|---|---|---|
| Standard Deviation | 25 | 30 |
| Mean | 45 | 50 |
| Number of observations | 100 | 100 |

8. Calculate Geometric Mean of the following data :

| X | f |
|---|---|
| 14 | 12 |
| 28 | 10 |
| 40 | 20 |
| 64 | 8 |
| 125 | 4 |

9. For a moderately skewed distribution the Arithmetic Mean is 200, the coefficeint of variation is 8, and the pearson's coefficient of skewness is 0.3. Find out Mode .

10. Karl pearson's coefficient of correlation between X and Y series is 0.4, their covariance is 9 and the standard deviation of X and Y series is 4 and 5. Find out the two regression coefficients.

**ANDHRA PRADESH OPEN UNIVERSITY**

**UNDERGRADUATE COURSE-II YEAR**

**SUBJECT : COMMERCE**

**COURSE : BUSINESS STATISTICS**

**ASSIGNMENT - I**

**N. B.**

1. Do not copy the answer directly from any of the books
2. As far as possible try to answer the questions independently in your own words
3. If it is necessary to quote from any source, give the correct reference
4. Use your own foolscap pages for writing the assignment
5. Leave sufficient margin for the comments of the evaluator.
6. Completion of this assignment normally should not take more than two hour's time .

ESSAY QUESTIONS

1. What are the functions of Statistics ?

2. During the years from 1984-85 to 1987-88 the number of students in a college are as follows :-

| Year | Arts | Commerce | Science | Total |
|---|---|---|---|---|
| 1984-85 | 100 | 75 | 50 | 225 |
| 1985-86 | 150 | 100 | 75 | 325 |
| 1986-87 | 200 | 150 | 100 | 450 |
| 1987-88 | 250 | 200 | 150 | 600 |

Present the data in a suitable diagram

3. Compute Mode of the following data

| Marks in Accountancy | No. of Students |
|---|---|
| 10 - 20 | 4 |
| 20 - 30 | 12 |
| 30 - 40 | 14 |
| 40 - 50 | 18 |
| 50 - 60 | 30 |
| 60 - 70 | 10 |
| 70 - 80 | 8 |
| 80 - 90 | 4 |

## SHORT QUESTIONS

1. What is Law of 'Statistical Regularity' ?

2. Calculate Mean of the following data . ?

| X | f |
|---|---|
| (Wages in Rs. ) | ( No. of Workers) |
| 200 | 7 |
| 400 | 8 |
| 600 | 9 |
| 800 | 6 |
| 1000 | 4 |
| 1200 | 2 |

3. Find out the combined Mean of the following data.

| | B.Com. | |
|---|---|---|
| | Section - A | Section- B |
| No. of students | 50 | 60 |
| Average | 65 | 55 |

## COURSE : BUSINESS STATISTICS

## ASSIGNMENT - II

**N.B.**

1. Do not copy the answer directly from any of the books
2. As far as possible try to answer the questions independently in your own words.
3. If it is necessary to quote from any source, give the correct reference
4. Use your own foolscap pages fro writing the assignment
5. Leave sufficient margin for the comments of the evaluator
6. Completion of this assignment normally should not take more than two hours.

### ESSAY QUESTIONS

1. What are the cahracteristics of a satisfactory measure of dispersion ?
2. Calculate standard deviation of the following data

| Class Interval | Frequency |
|---|---|
| 0 - 5 | 15 |
| 5 - 10 | 16 |
| 10 - 15 | 20 |
| 15 - 20 | 26 |
| 20 - 25 | 14 |
| 25 - 30 | 7 |
| 30 - 35 | 2 |

3. Calculate coefficient of skewness of the following data based on Karl pearson's formula :

| Income (Rs.) | 400-500 | 600-800 | 800-1000 | 1000-1200 | 1200-1400 |
|---|---|---|---|---|---|
| No . of Persons | 14 | 18 | 26 | 12 | 10 |



### SHORT QUESTIONS

1. Expalin the objectives of studying skewness

2. Find the combined standard deviation of the following two groups

| | Group - A | Group - B |
|---|---|---|
| Arthmetic Mean | 25 | 28 |
| Standard Deviation | 4 | 5 |
| Number of observations | 20 | 20 |

3. In a certain distribution $Q_3 = 75$, $Q_1 = 25$ and Median = 50. Find out the coefficient of skewness.

## COURSE BUSINESS STATISTICS

## ASSIGNMENT - III

N.B.

1. Do not copy the answer directly from any of the books
2. As far as possible try to answer the questions independently in your own words.
3. If it is necessary to quote from any source, give the correct reference
4. Use your own foolscap pages for writing the assignment
5. Leave sufficient margin for the comments of the evaluator
6. Completion of this assignment normally should not take more than two hours

### ESSAY QUESTIONS

1. Distinguish between correlation and regression
2. Calculate Karl pearson's coefficient of correlation of the following data

| | | | | | | |
|---|---|---|---|---|---|---|
| X | 24 | 26 | 28 | 22 | 25 | 30 |
| Y | 12 | 16 | 17 | 15 | 20 | 21 |

3. From the following data, calculate Fisher's Ideal Index.

| | Base Year (1984) | | Current Year (1988) | |
|---|---|---|---|---|
| Item | Price (Rs.) | Quantity | Price (Rs.) | Quantity |
| A | 4 | 60 | 5 | 70 |
| B | 8 | 80 | 9 | 60 |
| C | 12 | 30 | 16 | 40 |
| D | 16 | 10 | 20 | 20 |

## SHORT QUESTIONS

1. What are the uses of cost of living Index Numbers ?
2. In a distribution $\Sigma X = 21$, $\Sigma X^2 = 147$, $\Sigma Y = 25$, $\Sigma Y^2 = 185$, $\Sigma XY = 140$ and $N = 10$. Find out the Regression coefficint of X on Y.
3. Construct price Index for 1988 on the basis of 1984 with the help of Average Price Relatives Method by using Mean

| Items | A | B | C | D |
|---|---|---|---|---|
| Price in 1984 (Rs) | 80 | 40 | 60 | 10 |
| Price in 1988(Rs) | 100 | 80 | 90 | 40 |