Uttar Pradesh Rajarshi Tandon Open University

Bachelor of Business Administration

BBA-107

Business Statistics

Block-1	**BASIC STATISTICAL CONCEPTS**	**3-56**
UNIT-1	**Meaning and Scope of Statistics**	**5**
UNIT-2	**Organising a Statistical Survey**	**19**
UNIT-3	**Sampling**	**33**
UNIT-4	**Accuracy of Data**	**49**
Block-2	**COLLECTION AND CLASSIFICATION OF DATA**	**57-136**
UNIT-5	**Approximation and Errors**	**61**
UNIT-6	**Collection of Data-Primary Source**	**73**
UNIT-7	**Collection of Data-Secondary Source**	**91**
UNIT-8	**Classification of Data**	**113**
Block-3	**PRESENTATION OF DATA**	**137-206**
UNIT-9	**Tabular presentation**	**141**
UNIT-10	**Diagrammatic Presentation**	**149**
UNIT-11	**Graphical Presentation**	**167**
UNIT-12	**Ratio, Percentage and Rates**	**179**
Block-4	**MEASURES OF TENDENCY**	**207-306**
UNIT-13	**Measure of Central Tendency and Mean**	**211**
UNIT-14	**Median**	**233**
UNIT-15	**Mode**	**263**
UNIT-16	**-----**	**283**

Block-5 **307-352**

UNIT-17	**Measures of Dispersion-I**	**309**
UNIT-18	**Measures of Dispersion-II**	**323**
UNIT-19	**Measures of Skewness-I**	**337**
UNIT-20	**Measures of Skewness-II**	**341**

Uttar Pradesh Rajarshi Tandon Open University

Bachelor of Business Administration

BBA-107

Business Statistics

BLOCK

1

BASIC STATISTICAL CONCEPTS

UNIT-1 **5**

Meaning and Scope of Statistics

UNIT-2 **19**

Organising a Statistical Survey

UNIT-3 **33**

Sampling

UNIT-4 **49**

Accuracy of Data

Curriculum Design Committee

Prof. Omji Gupta **Director**

School of Management Studies, UPRTOU, Prayagraj

Dr Gyan Prakash Yadav

Assistant Professor

School of Management Studies, UPRTOU, Prayagraj

Dr Devesh Ranjan Tripathi

Assistant Professor

School of Management Studies, UPRTOU, Prayagraj

Course Preparation Committee

Dr T.M. Raini **Author**

Associate professor, Firoj Gandhi PGCollege, Raibarely

Dr Shiv Bushan Gupta

Associate professor, MMM PG college, Kalakankar, Pratapgarh

Dr Gaurav Sankalp

Academic consultant, School of Management Studies, UPRTOU, Prayagraj

Dr Ajay Tiwari

Assistant Professor

MONIRBA, Department of Commerce, University of Allahabad

Prof J.N. Mishra **Editor**

Professor, MONIRBA, Department of Commerce, University of Allahabad

Dr Gyan Prakash Yadav **Coordinator**

Assistant Professor

School of Management Studies, UPRTOU, Prayagraj

©UPRTOU, Prayagraj-2021

ISBN :

©All Rights are reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the **Uttar Pradesh Rajarshi Tondon Open University, Prayagraj.**
Printed and Published by Vinay Kumar Registrar, Uttar Pradesh Rajarshi Tandon Open University, 2023.

Printed By : Chandrakala Universal Pvt. Ltd. 42/7 Jawahar Lal Neharu Road, Prayagraj .

UNIT-1 MEANING AND SCOPE OF STATISTICS

Objectives

After reading this unit, you should be able to :

- Present a broad overview of statistics as a subject.
- Bring out applications of statistics as its usefulness in management decision making.
- Understand the limitation and distrust of statistics.

Structure

1.1 Introduction

1.2 Meaning of Statistics

1.3 Definition of Statistics in Plural & Singular Sense.

1.4 Type of Statistical Method-Descriptive and Inferential Statistics.

1.5 Function of Statistics

1.6 Importance and Scope of Statistics

1.7 Limitation of Statistics

1.8 Distrust of Statistics

1.9 Self Assessment Questions

1.1 INTRODUCTION

Statistics is a fast growing subject. These days there is hardly any subject or branch of study that deals with large numbers or data, but does not use the methodology of statistics in one or the other form. Therefore, We can say that statistics constitutes an integral part of every scientific and economic inquiry.

Social an economic studies without statistics are in conceivable. Statistics is indispensible in almost all the spheres of human activity and knowledge. Statistics, thus plays a multifarious role and as **Tippett** has rightly remarked, *"it affects everybody and touches life at many points."*

The word 'Statistics' of English language has either been derived from the Latin word **status** or Italian word statistica or German word statistic. In each case it means *'an organized political state'*. In this sense

this word is found used in the seventeenth century in the famous drama Hamlet (1602) of Shakespeare. Milton used this word in his famous epic Paradise Regained. In the early Propounders of the discipline, the name of *Gatt Fried Achenwall (1719-1772)* who was professor of Law and political science at **Gottington (Germany)**, is worth mentioning. He in 1749, recognized statistics as specialized branch of human knowledge. This is why he is called **'Father of Statistics.'**

1.2 MEANING OF STATISTICS

Broadly speaking the word statistics means:

(a) Statistics as Statistical Data.

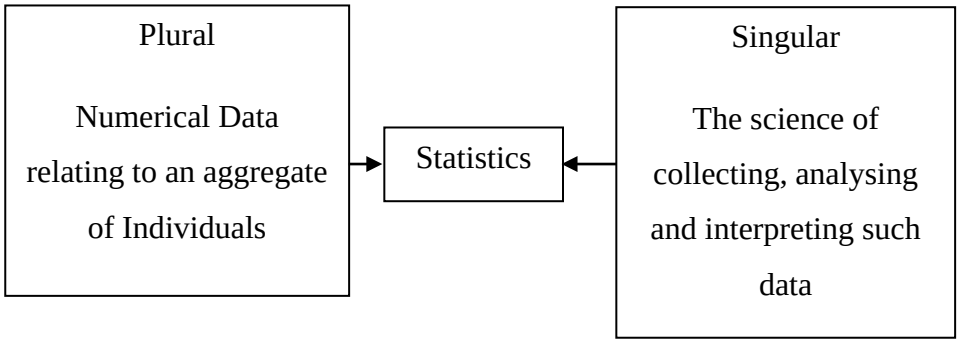
(b) Statistics Methods or science of statistics.

(c) Statistics, Plural of the word statistic.

The common man refers the word statistics as numerical data. For example, Statistics of National Income, statistics of Automobile, Accidents, Employment statistics, Market statistics, Production statistics, Import and Export statistics, etc.

In this sense the word 'statistics' is considered in plural noun form. In a wide and singular from statistics refers as a science of statistics. In this sense the science of statistics is commonly known as statistical methods.

In **Kendall and Buckland statistical Dictionary** the word statistics is explained as :



The three meaning of the word 'statistics' are contained in the statement due to Tate:

"You compute statistics by statistics from statistics."

1.3 DEFINITION OF STATISTICS

It is very difficult to assign any precise definition of statistics as various definitions have been given by different authors on the subject. In

1869, **Quetlet** prepared a list of about 180 different definitions. In this context Bowley says that, "very many definition have been given of the word 'statistics' and each author who has written on the subjects has assigned new limits to the field which should be included in its scope."

Kendall remarked. "among the many subjects about which statistics disagree in the definition of their science."

On the basis of these definition the term 'statistics' can be expressed in two ways.

1. **In Plural sense :** Statistics refers to statistical data.

2. **In Singular Sense :** Statistics refers to statistics as a science or branch of knowledge – (a) Classical view (b) Modern view

1.3.1 IN THE FORM OF PLURAL SENSE

Achenwell, the father of statistics had said, *"statistics are collection of noteworthy facts concerning state-both historical and descriptive".*

Statistics refer to statistical data :

According to **Wappaus**, *"An aggregate of knowledge brought together for a practical end namely the knowledge of a concrete state."* Above two definitions are very narrow because they are limited to state only.

According to **A.L. Bowley***"Statistics are numerical statements of facts in any department of enquiry, placed in relation of each other."* This definitions highlights three characteristics –

1. Statistical data may relate to any filed of investigation.

2. Statistics are numerically expressed.

3. Statistics are comparable.

As the above definitions does not contain other characteristics hence it is incomplete.

Definition given by **Hares Secrist** is more universal and clear which also posses all the characteristics of statistics.

Secrist has given the most exhaustive and clear definition in which all the characteristics are exposed. According to him, *"By statistics we mean aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in systematic manner for a pre-determined purpose and placed in relation to each other."*

1.3.2 THE DEFINITION OF STATISTICS IN SINGULAR FORM

Statistics has been defined by various scholars in different ways. But many of them have expressed it as the statistical methods. They have widely accepted the singular form of statistics. Definitions expressing it as a science can be divided in two groups –

(a) Classical Definitions- According to **Beildfeld**, *"The Science that teachers us what is the political arrangement of all the modern states of the known world."* This definition is not perfect because it limits the scopes of statistical science upto statecraft.

Kendall defines statistics as – *"Statistics is the branch of scientific method which deals with the properties of population."*

But further he himself condemns is saying that it is very ordinary and adds that – *"Statistics is the branch of Scientific method which deals with data obtained by counting or measuring the properties of populations."* But again he considers it incomplete. Third time he defines it as – *"Statistics is the branch of scientific method which deals with the data obtained by counting or measuring the properties of population of natural phenomena."*

Here natural phenomena activities of external world as a whole, either individual or God oriented. **Boddington** defines statistics as *"The science of estimates and probabilities."* This is also an incomplete and probabilities though being important estimates and probabilities are only a part of statistical methods. Actual counting or enumeration is also important. When the numbers are has than they are actually counted.

(b) Modern Definitions : Modern definitions have widened the scope of statistical science. Some of them are as follows –

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

   - Lovitt

2. *"Statistics is the science and the art of handling aggregate of facts."*

   - Person, Harlows

3. *"Statistics is the body of methods which are used when we wish to study masses of numerical data and to extract from them a few simple facts."*

   - Waugh

4. *"Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data."*

- Croxton and Cowden

5. *"Statistical methods are methods specially adopted to the elucidation of quantitative data affected by multiplicity of causes."*

- Yule

All the above definition conclude that statistics is the collection of scientific methods for drawing conclusions out of the numerical facts. These methods deal with collection of data, classification, tabulation, comparison, analysis and interpretation. Statistics is both science and art. Hence its scope is very wide.

1.4 TYPES OF STATISTICAL METHODS

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

(i) Descriptive statistics

(ii) Inferential statistics

(i) **Descriptive statistics** includes statistical methods involving the collection, presentation and characterization of a set of data in order to describe the various features of that set of data. In general, methods of descriptive statistics include graphic methods and numeric measures. Bar charts, line graphs, and pie charts comprise the graphic methods, whereas numeric measures include measures of central tendency, dispersion, skewness, and kurtosis.

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

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

conclusions about the characteristics of a large population under study.

1.5 FUNCTIONS OF STATISTICS

In the modern age, there is no field wherein statistics is not used. There are many persons that are uneducated and they know nothing about statistics, still they use statistics in their everyday life, although at the outset, statistics develops as a state science the chief purpose of which was to collect the information concerning the state. But today the use of statistics has become essential in every field, or we can say, that without using statistics no achievement can be gained. In other words, the planning of future will become impossible without the help of statistics.

Famous thinker **M.M. Blair** has gone to the extent of writing that *'we live in statistics'*, Blair's view cannot be disagreed with If the human welfare is to be done to the maximum and the means are to be used properly, to take help of statistics is most essential. Main functions of statistics are as follows –

1. **To simplify complex facts :** The complex mass of data are made simple and understandable with the help of statistical methods. **Dr. A.L. Bowley** states that, *"The object of a statistical estimate of complex groups is to present an outline to enable the mind to comprehend with a single effort the significance of the whole."* Human mind is unable to remember huge facts and figures. Statistical methods make these data easy to grasp. Figures are boring. A man is bound to be confused with figures. Statistical techniques like average, variation, graph, diagram, etc., make these figures intelligible and understandable. With the help of statistical methods it is possible to understand the whole thing in a short time and in a better way.

2. **Comparison of facts :** The science of statistics does not merely mean counting but also comparison. Comparison in quantitative terms is easy. **Boddington** states that. *"The object of statistics is to enable comparison to be made between the past and present results with a view to ascertaining the reasons for changes which have taken place and the effect of such changes in the future."* Hence the chief function of statistics is comparison. Statistics devices like averages, ratios, percentages, rates, coefficients, standard error, etc. offer the best way of comparison between two phenomena.

3. **Establishment of relationship :** The relationship existing between demand and supply, money and price-level, rainfall and agriculture production can best be measured with the help of statistical methods. Statistics also help in finding out the association between two or more attributes.

4. **To enlarge individual knowledge and experience :** In the words of **Whipple**, *"Statistics enables one to enlarge his horizon."* Dr. A.L. Bowley also observes that, *"The proper functions of statistics indeed it to enlarge individual experience."* Knowledge becomes precise and cosy to understand with the help of statistical techniques.

5. **To formulate policies in different fields :** Statistics help in formulating policies in different fields, especially in social, economic and business fields. The governmental policies are framed on the basis of statistics. Policies in the fields of taxation, import, export, social welfare, wages, etc. are formulated on the basis of statistical data and inferences drawn from their analysis.

6. **To measure the effect :** Not only in formulating a policy but also is measuring the effects of a policy. Statistics acts as a guide. The effects of a change in the Bank rate or tax rate, etc., can be properly studies with the help of statistics.

7. **To test a hypothesis :** Statistical methods are also employed to test the hypothesis in theory and discover newer theory. This is the most important theoretical function of statistics. According to **Prof. Keynes**, *"The function of statistics is first to suggest empirical laws which may or may not be capable of subsequent deductive explanation and secondly to supplement deductive reasoning by checking its results and submitting them to the best of experience."*

8. **Forecating :** The future is uncertain. Statistics directly helps in forecasting. On the basis of facts and figures related to the past and the present one can foresee the future with fair amount of certainty. Statistics does this job with forecasting function. Statistics help in planning and finding solutions to be problems which may arise in future.

1.6 IMPORTANCE AND SCOPE OF STATISTICS

Importance and utility of statistic has widened significantly in the modern days. As a matter of fact statistics affects every person and touches several aspects of his life. **Croxten** and **Cowden** has said that, *"Today there is hardly a phase of endeavour which does not find statistical devices at least occasionally useful."*

Being a citizen our entry and exit in this world is also maintained in the form of statistical data. In every phase and activity of society the importance of statistics is acknowledge and for this reason it can be said that statistics is no more a *"Science of Kings"* but has become the *"Science of Human Welfare."* Moreover, our culture has also been influenced by statistics. According to **Walkar***"To a very striking extent our culture has become a statistical culture."*

It is not possible to understand basic principles of Psychology, Sociology, Economics and Physical Sciences without the knowledge of statistics. Almost all the branches of science depend on statistical methods for their development. Economists, scientists, social workders, businessmen, farmers, doctors and lawyers all are indebted to statistical science. Sometimes musicians, artists and writers also use statistical facts. According to **Croxton and Cowden**, *"Without an adequate understanding of statistics the investigator in the social sciences may frequently be like the blind grouping in the dark closet for a black that isn't there."*

Blair observed that *"We live in the midst of statistics, from prices to the details of weather our life is guided by statistical data. From routine daily activities to the formation of national laws, almost all activities depend upon statistics."*

It will not be inappropriate to say that knowingly or unknowingly we use the basic principle of statistics in the selection of our friends, spouse, business partner, foods, clothes, professions as well as to save ourselves from the apprehension of losses etc.

Actually the scope of statistics has become so vast that we generally apply statistical data in every field of life. For example parents are always eager to know the weight of their new born child so that by comparing the average weight they can know whether their child is healthy or not. Similarly as the child grows up they want to know whether the growth is normal or not. When we buy essential items for him the shopkeeper after knowing the age of the child provides all the items, insurance company insure him. But how all this is known? Answer is very simple – with the help of applying the statistical method.

Importance of statistical science is more clear from the following statements –

"Statistics are the straw out of which one, like every other economist, have to make bricks."

- Marshall

"Statistical enable one to enlarge his horizon."

- Whipple

"When you can measure what you are speaking about and express it in numbers, you know some thing about it, but when you can not measure, when you can not express in numbers, your knowledge is of a meagre and unsatisfactory kind.

- Kelvin

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

1.6.1 STATISTICS AND THE STATE

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

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

1.6.2 STATISTICS IN ECONOMICS

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

(i) Time-series analysis is used for studying the behaviour of prices, production and consumption of commodities, money in circulation, and bank deposits and clearing.

(ii) Index numbers are useful in economic planning as they indicate the changes over a specified period of time in (a) prices of commodities, (b) imports and exports, (c) industrial/agricultural production, (d) cost of living, and the like.

(iii) Demand analysis is used to study the relationship between the price of a commodity and its output (supply).

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

1.6.3 STATISTICS IN BUSINESS MANAGEMENT

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

Statistical reports provide a summary of business activities which improves the capability of making more effective decisions regarding future activities. Discussed below are certain activities of a typical

organization where statistics plays an important role in their efficient execution.

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

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

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

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

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

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

1.6.4 STATISTICS IN PHYSICAL SCIENCE

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

1.6.5 STATISTICS IN SOCIAL SCIENCES

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

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

- Bowley

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

- W.I. King

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

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

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

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

1.6.6 STATISTICS IN MEDICAL SCIENCE

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

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

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

1.6.7 STATISTICS IN COMPUTER SCIENCE

Computers and information technology, in general, have had a fundamental effect on most business and service organizations. Over the last decade or so, however, the advent of the personal computer (PC) has revolutionized both the areas to which statistical techniques are applied. PC facilities such as spreadsheets or common statistical packages have now made such analysis readily available to any business decisions-maker. Computers help in processing and maintaining past records of operations

involving payroll calculations, inventory management, railway/airline reservations, and the like. Use of computer softwares, however, presupposes that the user is able to interpret the computer outputs that are generated.

1.7 LIMITATIONS OF STATISTICS

Statistical methods are now used in almost all the branches of study. But in spite of the wide scope of the subject of statistics, it is not fully free of limitations. In reality, statistics suffers from a few grave limitations. Some of them being –

(1) Statistical results may be non-uniform : If we wish to arrive at an uniform results, we have to use homogeneous data and information as the raw materials of statistical enquiry. If, however, the available material is not uniform, some measures may be undertaken to give it uniformity, but the statistical measures alone cannot bring in perfect uniformity out of heterogeneous elements. It is a severe limitation of statistics or that of statistical methods.

(2) Statistical methods are not applicable to qualitative studies : An other major limitation of statistics is the fact that statistical methods are not applicable to qualitative studies like 'good', 'bad', 'intelligence' etc. For a proper, coordinated and wholesome coverage of the whole scope the methods must be applicable to quantitative as well as qualitative studies.

(3) Statistical laws are mostly dependent on average which may give fallacious result : Statistical laws are mostly derived from and dependent on average of any kind. The averages, unfortunately, sometimes represent fallacious, contradictory and unwholesome results. Sometimes they give impracticable and abnormal results that have no meaning in practical life.

(4) Jurisdiction of statistics cannot be reduced to individuals : A proper and judicious analysis of the definitions of statistics would indicate that the subject deals with large mass of data naturally, if all the figures and numerical are reduced to a single, isolated figure, it cannot be termed as statistics. Hence isolated or single figures, however, important they might be, cannot be studied within the purview of statistics. It is a major handicap of the subject.

(5) Statistical results cannot always be treated as the sole determinant of the value of a group : Some men, erroneously think that statistical methods are the only applicable tools of ascertaining truths or values of a group. It is not so. In reality, statistical measures are some ways to ascertain results from the mentioned groups. It is one of the methods out of several ones to analyse a 'determinate'.

(6) All statical methods are subject to bias : This may give a wrong or a biased result. On account of all these defects and bias, sometimes a statistical enquiry may give abnormal or absurd results.

(7) Statistical laws are not correct in the short period : To be a perfect science, its laws must always be perfectly correct and accurate, irrespective of time element. But statistical laws are correct only on the average and in the long run.

Thus, from all the above, it is evident that the subject suffers from some limitations also.

1.8 DISTRUST OF STATISTICS

It is a general belief that, *"Statistics can prove anything."* The statement, though common is not cent percent true. Similarly, we cannot say that the above is fully wrong – due to the fact that there is some truth in it. Statistical methods being most frequently used in almost all the branches of study, are used also by unscrupulous men to achieve their personal ends. It is on account of this that statistics are deliberately put in a cooker form by the unscrupulous to give a wrong, biased and deceitful picture.

It is also said that, "There are three kinds of lies – lies, condemned lies and statistics. The statement in very much sarcastic though there is a seed of truth in it. Statistics, as expressed in the remark, is a lie of superlative degree-indicating the fact that statistical data and methods may be easily and favourably twisted to give a desired result. If it is true, how the subject can be relied upon? Obviously, it cannot be easily admitted as cent percent true.

Absence of an objective test that may be applied to a statistical data to know its worth and intrinsic values also helps in spreading a general distrust against statistics. Some of the statistical methods, themselves are not very appropriate and correct and cause distrust for the subject itself.

But, whatever may be the causes of distrust of the subject, we cannot deny the truth that statistics is being put to more and more uses in almost all the branches of study with the rigidity of the fixation of the term of reference of the enquiry, selection and training of the enumerators for better and unbiased enquiry, improvements in technique of collection, tabulation, classification, processing, analysis and interpretation of data, distrust towards statistics is going down.

Bowley, thus, rightly asserted that statistics are only tools. One uses them according to his knowledge and experience. The science of statistics is most useful servant. Its distrust lies in its uses. For those who understand its proper use, the science of statistics is of great value and to those who do not understand its use, naturally they give rise to the distrust

of statistics. Thus, we can conclude with the hopeful remarks that, *"With the study of statistics as a science, with the recognition of its limitations and with improvements in its techniques, the causes for its distrust would gradually wane."* The day is not for off.

1.9 SELF ASSESSMENT QUESTION

(a) Short Answer Questions :

(i) Discuss the importance of statistics.

(ii) What are the causes of distrust of statistics?

(iii) Discuss the limitation of statistics.

(iv) Comment on the following statements –

(a) At its best science is statistics.

(b) Statistics is the science A estimates and probabilities.

(c) Statistics can prove anything.

(v) Discuss the scope of statistics.

(vi) What statistics, is art and science both?

(b) Long Answer Questions :

(i) Define statistics and discuss its nature and scope.

(ii) Prove that statistics is a science and and both.

(iii) Discuss and scope and division of statistics.

(iv) Write an essay/note on the following –

(a) Statistics in the service of state.

(b) Distrust of statistics.

(c) Important of statistics in trade and Business.

(d) Scope of statistics.

(e) Function of statistics.

(f) Limitation of statistics.

UNIT-2 ORGANISING A STATISTICAL SURVEY

Objectives

After reading this unit, you should be able to understand :

- The steps or pre-determinants of organising a statistical survey.
- The concept of sources of data and their collections.
- Understand the types of enquiry.
- The process of actual collection of data relating to the given enquiry.

Structure

2.1 Need of organising a statistical survey

2.2 Steps in statistical survey

2.3 Primary and secondary data

2.4 Sources of data

2.5 Types of enquiry

2.6 Self help questions

2.1 NEED OF ORGANISING A STATISTICS SURVEY

For conducing any investigation it is necessary to make a proper planning in absence of which some unnecessary facts which might be collected, can give confusing results. It will also be misuse of time, money and energy. Hence, it is clear that before commencing any investigation every aspect of the subject should be thoroughly checked. What will be the object and scope of investigation, how much accuracy is needed, from where the data will be collected, which method of data collection will be adopted, what will be the definition of unit for fulfilling the object, whether investigation will be conducted by census method or by sampling method, if by sampling then which specific method etc. problems should be paid attention and investigation should be planned. For example, if the problem of unemployment of educated persons in a particular area is to be investigated then it should be finalised first that how many such persons of the area should be investigated, what questions should be asked, what is the concept of education and unemployment, what will be the method of

data collection and how the collected data will be exhibited, etc. So an investigation planning of includes all the above facts.

2.2 ORGANISING OF INVESTIGATION

For planning, as Statistical Investigation the following aspects must be considered:

1. Object and Scope,
2. Nature and Type,
3. Extent of Accuracy,
4. Statistical Unit,
5. Sources of Data,
6. Method of Collection,
7. Devising the Questionnaire,
8. Editing the Data,
9. Presentation and Analysing of Data
10. Preparations of Report

2.2.1 OBJECT AND SCOPE

It is important to determine the object and scope of investigation before its beginning. For example, if any labour union collects the data of wages then probably its object is to compare the wages of that area with that to the wages of the labour of other sector and if it is less than to demand for its increase. Like this the object and scope of investigation is determined and for this it is necessary to contact the labour of that particular area. **Prof. Neiswanger**, in his book **Elementary Statistical Methods**, has stressed the need for determining the object of investigation. He is of the opinion that if the object is not determined then it is difficult to collect the correct data in absence of which the subject and its assessment is indefinite. Hence for the accuracy of investigation and for proper utilisation of time and money, the object and area of investigation must be clearly determined.

According to **Neiswanger**, *"Statement of objective is of basis importance because it determines the data which are to be collected, the characteristics of the data which are relevant, relations which are to be explored, the choice of the technique to be used in these explorations and the form of the final report."*

The object of investigation can be general or specific. Investigations having a general object can be useful for all such as the investigations related to population, national income, literacy etc. Specific investigations are conducted on small scales and are useful for specific persons, such as unemployment problem of engineers, problem of indiscipline at the university level, or problem of labour involved in lock trade of Aligarh etc.

Scope of investigation can be divided into economic, political and natural, state, district and city are the political part. Industry, agriculture,

minerals and transportation are economic and mountains, forest and fields are the natural parts.

2.2.2 NATURE AND TYPE

There are various kinds of statistical investigation. Hence, it is very important to know the nature and type of investigation. Investigation is specific or general. If specific then basic facts are to be collected but if it is general then pre-collected facts will be sufficient. Likewise, it should be determined whether investigation will continue in future or is irregular. Is it of secret nature or for general use? Whether the collection of facts will be direct or indirect? Will it be completed by personal efforts or is government-aided? Is it preliminary or in continuation of past investigation or after some amendments it is being continued? Whether it will be conducted by census method or by sample method? If to be conducted by sample method then which of these– deliberate, random, stratified etc. will be used? Different investigation are ideal for various situation. Therefore the nature and type of investigation be determined according to the circumstances.

2.2.3 EXTENT OF ACCURACY

Generally total accuracy is not possible in statistical investigation. Statistical methods cannot be compared with scientific methods. There is an important place of probabilities in statistics. It is not necessary that even after investment of more money and time complete accuracy can be achieved and if at all it is achieved it will be some use. Generally, the relative error should be less. Actually the accuracy depends on the object of investigation. For example, it will be sufficient to indicate the number of Kumbh visitors in thousands or lacs but the slightest change in the decimal part of the rate of interest will not be correct. Similarly, figures related to price value will need more accuracy in comparison to the figures of National income. It is sufficient to indicate the weight of coal in metric tonnes but for gold even one gram of weight is very important.

> According to **King**, *"For every statistical problem, there should be determined in advance a definite standard of accuracy ..... but this standard, by no means, needs to correspond to the highest degree of accuracy attainable."*

2.2.4 STATISTICAL UNIT

Statistical Unit is accurate basis of numerical measurement. It is that base of measurement according to which the data is collected. To define statistical unit for the purpose of investigation seems very easy but actually it is very difficult. If the definition is illusive, errorful or is not clear then it affects the object of investigation. For example, if the data related to wages of labour is to be collected and the 'wages' and 'labour' is not defined then it is ought to create problems-such as who should be the labour, total working labour, or workers getting payment to a certain limit,

or the labours involved in cottage industries, or factories or in agricultural jobs, permanent workers or casual workers, whether children, women and men all are to be included or should be separately considered because of the difference in their average income. Similar problems are faced in case of salary. Whether salary includes only cash or other benefits too. If other benefits are included then what are they. In the same way investigation related to the unemployment of educated persons then what does 'educated' mean. Even slightest educated person should be considered as literate or not. Then the unemployment is either total or partial. Some problems are faced in defining profits, income, price, etc.

In short statistical unit must possess the following qualities to avoid any hindrance in fulfilling the object of investigation :

(1) The definition of Unit should be easy, appropriate, in short and clear.

(2) Unit should be easily traceable and should be stationary, certified and accurate.

(3) Unit should be similar in all respect so that can be easily compared.

(4) Unit should be according to the object of investigation.

(5) Once the unit is finalised it should be continued till the investigation is completed.

2.2.5 SOURCES OF INFORMATION (DATA)

Having decided about the objectives and scope of the enquiry and the statistical units to be used, the next problem is to decide about the sources from which the information (data) can be obtained or collected. For any statistical enquiry, the investigator may collect the data first hand or he may use the data from other published sources such as the publications of the government/semi-government organisations, periodicals, magazines, newspaper, research journals etc. If the data are collected originally by the investigator for the given enquiry it is termed as **primary data** and if he makes use of the data which had been earlier collected by some one else, it is termed as **secondary data**. For example, the vital rates, i.e., the rates of fertility and mortality in India prepared by the office Registrar-General of India, New Delhi, are primary data but if the same data are reproduced in the U.N. Statistical Abstract (a publication of the United Nations Organisation), it becomes a secondary data. Obviously the type of enquiry needed for primary data is bound to be of a totally different nature than the type of enquiry needed for the use of secondary data. In case of primary data, the type of enquiry requires laying down the definitions of the various terms and the statistical units used in the enquiry, keeping in mind the objectives and scope of the enquiry. However, in the use of secondary data there are no such problems since the data have already been collected under a given set of definitions of

various terms and units used. However, before using secondary data for statistical investigation under study, it must be subjected to careful editing and scrutiny with respect to their reliability, suitability and adequacy. [For details see 2.6 in this Chapter]. For a given enquiry, the use of either primary or secondary data or both may be made, depending upon the purpose and scope of the enquiry.

2.2.6 METHOD OF DATA COLLECTION

The problem does not arise if secondary data are to be used. However, if primary data are to be collected a decision has to be taken whether **(i) census method or (ii) Sample technique**, is to be used for data collection. In the census method, we resort to 100% inspection of the population and enumerate each and every unit of the population. In the sample technique we inspect or study only a selected representative and adequate fraction (finite subset) of the population and after analysing the result of the sample data we draw conclusions about the characteristics of the population. In some situations such as population being infinite or very large, census method fails. Moreover, it is not practicable if the enumeration or testing of the units (object) is destructive e.g. for testing the breaking strength of chalk, testing the life of electric bulbs or tubes, testing of crackers and explosives etc. Even if practicable, it may not be feasible from considerations of time money. Thus a choice between the sample method and census method is to be made depending upon the objective and scope of the survey, the limitations of resources in terms of time, money, manpower etc, and the degree of accuracy desired. In case of sample method the size of the sample and the technique of sampling like simple random sampling, stratified random sampling, systematic sampling etc., are to be decided.

2.2.7 DEVISING THE QUESTIONNAIRE

After knowing the subject of investigation it is necessary to devise a questionnaire. Questions must be easily and correctly answerable. After preparing the questionnaire a schedule should be prepared. A paper indicating directions for answering is annexed with the schedule for making it easier. The following things should be kept in mind while preparing the schedule :

(1) Questions are easy and clear specially of those schedules which are sent by post.

(2) Questions are answerable in 'yes' or 'no'.

(3) Questions should not be confusive.

(4) Questions should be directly connected with investigation.

(5) Questions should amicably answerable.

(6) They should not be more in numbers.

(7) They should be in proper serial and can easily be adjusted.

(8) If necessary, questions should be confirmatory.

2.2.8 EDITING THE COLLECTED DATA

After collecting the data comes the editing. If investigation is small then only one person can do this job but in case of a big investigation it is conducted by two or more persons. Incomplete schedules are either left unconsidered or are returned to the sender for there completion. Controversial answers are checked. Some figures are approximated. To serialise the information, to check the entries, put down the code, to find the multiple, to add and classification all these are related to editing. Their proper arrangement can only produce correct tables and accurate results.

2.2.9 PRESENTATION AND ANALYSIS OF DATA

The collected data needs presentation. But before presentation they are analysed with the help of different available methods-averages etc., so as to draw conclusion and present them in a manner in which it is most suitable according to the nature of the enquiry.

2.2.10 PREPARATION OF REPORT

In any statistical investigation preparation of report occupies a key position because of simple reason that it is the conclusion on the basis of which the decisions may be taken. Policy framing, further action required decision and also implementation depends upon the result of statistical enquiry. Usually investigator himself does this job. He proceeds to prepare the report. The report should be in simple and lucid language, brief, to the point, clear and as far as possible self-explanatory. The report should also throw light on the problem investigated, its possible solution and future course of action which may help society at large.

2.3 PRIMARY AND SECONDARY DATA

After going through the preliminaries discussed in the above section, we come to the problem of data collection. The most important factor in any statistical enquiry is that the original collection of data is correct and proper. If there are inadequacies, shortcomings or pitfalls at the very source of the data, no useful and valid conclusions can be drawn even after applying the best and sophisticated techniques of data analysis and presentation of the results. In this context, it may be interesting to quote the remarks made by a judge on Indian statistics. "Cox, when you are bit older you will not quote Indian statistics with the assurance. The governments are very keen on amassing statistics – they collect them, add them, raise them to the ***n*th** power, take the cube root and prepare

wonderful diagram. But what you must never forget is that every one of those figures comes in the first instance from the 'Chowkidar' (i.e., the village watchman) who just puts down what he damn pleases."

It may be remarked that this quotation applies to India of very old days when no definite statistical set up existed in India. Today in India, we have a fairly sound and systematic method of data collection on almost all problems relating to various diversified fields such as economics, business, industry, demography, social, physical and natural sciences. As already pointed out the data may be obtained from the following two sources :

(i) The investigator or the organising agency may conduct the enquiry originally or

(ii) He may obtain the necessary data for his enquiry from some other sources (or agencies) who had already collected the data on that subject.

The data which are originally collected by an investigator or agency for the first time for any statistical investigation and used by them in the statistical analysis are termed as **primary** data. On the other hand the data (published or unpublished) which have already been collected and processed by some agency or person and taken over from there and used by any other agency for their statistical work are termed as **secondary** data as far as second agency is concerned. The second agency if and when it published and files such data becomes the **secondary source** to any one who later uses these data. In other words secondary source is the agency who publishes or releases for use by others the data which was not originally collected and processed by it.

It may be observed that the distinction between primary and secondary data is a matter of degree or relativity only. The same set of data may be secondary in the hands of one and primary in the hands of others. In general, the data are primary to the source who collects and processes them for the first time and are secondary for all other who later use such data. For instance, the data relating to mortality (death rates) and fertility (birth rates) in India published by the Office of Registrar General of India, New Delhi are primary whereas the same reproduced by the United Nations Organisation (U.N.O.) in its U.N. Statistical Abstract become secondary in as far as later agency (U.N.O.) is concerned. For this data, the office of Registrar General of India, is the primary source while U.N.O. is the secondary source. Likewise, the data collected by C.S.O. and N.S.S.O. for various surveys are primary as far as these departments are concerned but they become secondary if such data are used by other departments or organisations.

Choice Between Primary and Secondary Data – Obviously there is lot of difference in the method of collection of primary and secondary data. In the case of primary data which is to be collected originally, the entire scheme of the plan starting with the definitions of various terms used, units to be employed, type of enquiry to be conducted, extent of

accuracy aimed at etc., is to be formulated whereas the collection of secondary data is in the form of mere compilation of the existing data. A proper choice between the type of data (primary or secondary) needed for any particular statistical investigation is to be made after taking into consideration the nature, objective and scope of the enquiry ; the degree of precision aimed at and the status of the agency (whether government – state or central – or private institution or an individual).

2.4 SOURCES OF DATA

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

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

1. **Primary Data Sources :** Individuals, focus groups, and/or panels of respondents specifically decided upon and set up by the investigator for data collection are examples of primary data sources. Any one or a combination of the following methods can be chosen to collect primary data:

   (i) Direct personal observations

   (ii) Direct or indirect oral interviews

   (iii) Administrating questionnaires

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

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

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

Interviewing Interviews can be conducted either face-to-face or over telephone. Such interviews provide an opportunity to establish a rapport with the interview and help extract valuable information. Direct interviews are expensive and time-consuming if a big sample of

respondents is to be personally interviewed. Interviewers' biases also come in the way. Such interviews should be conducted at the exploratory stages of research to handle concepts and situational factors.

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

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

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

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

(a) External secondary data sources

- Government publications, which include

(i) The National Accountants Statistics, published by the Central Statistical Organization (CSO). It contains estimates of national income for several years, growth rate, and rate on major economic activities such as agriculture, industry, trade, transport, and so on;

(ii) Wholesale Price Index, published by the office of the Economic Advisor, Ministry of Commerce and Industry;

(iii) Consumer Price Index;

(iv) Reserve Bank of India bulletins;

(v) Economic Survey.

- Non-Government publications include publications of various industrial and trade associations such as

    (i) The Indian Cotton Mills Association

    (ii) The various Chamber of Commerce

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

(b) Internal secondary data sources :

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

Advantages and Disadvantages of Secondary Data :

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

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

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

2.5 TYPES OF ENQUIRY

Another important point one has to bear in mind before embarking upon the process collect on of data is to decide about the types of enquiry. The statistical enquiries may be different types as outlined below :

(i) Official, Semi-official or Un-official.

(ii) Initial or Repetitive.

(iii) Confidential or Non-confidential.

(iv) Direct or Indirect.

(v) Regular or Ad-hoc.

(vi) Census or Sample.

2.5.1 OFFICIAL, SEMI-OFFICIAL OR UN-OFFICIAL

A very important factor in the collection of data is 'the sponsoring agency of the survey or enquiry.' If an enquiry is conducted by or on behalf of the central, state or local governments it is termed as official enquiry. A semi-official enquiry is one that is conducted by organisations enjoying government patronage like the Indian Council of Agricultural Research (I.C.A.R.), New Delhi; Indian Agricultural Statistical Research Institute (I.A.S.R.I.), New Delhi; Indian Statistical Institute (I.S.I.), Calcutta; and New Delhi and so on. An un-official enquiry is one which is sponsored by private institutions like the F.I.C.C.I., trade unions, universities or the individuals, Obviously the facilities available for each type of the above en-enquiries differ considerably. In an official enquiry, legal or statutory compulsions can be exercised asking the public or respondents to furnish the requisite information in time and that too at their own cost. In semi-official type enquiries also, the necessary information may be obtained without much difficulty. However, in un-official enquiries the investigator is faced with serious problems in getting information from the respondents. He can only persuade and request them for information. In such enquiries there is only moral obligation and no legal compulsion on the respondents. Things are still worse if the enquiry is conducted by an individual who, at stages, has even to beg for information. Moreover, there are lot of differences in the financial positions of these three sponsoring agencies. Obviously, the state or central governments can afford to spend much more on an enquiry as compared with private institutions, which in turn, can generally spend more than an individual. Consequently, there is bound to be difference in the types of enquiries depending upon the sponsoring agency and also its financial implications.

2.5.2 INITIAL OR REPETITIVE

As the name suggests, an *initial* or o*riginal* enquiry is one which is conducted for the first time while a *repetitive* enquiry is one which is carried on in continuation or repetition of some previously conducted enquiry (enquiries). In conducting an original enquiry the entire scheme of the plan starting with definitions of various terms, the units the method of collection etc., has to be formulated afresh whereas in repetitive enquiry

there is no such problem as such as plan already exists and only the original enquiry is to be modified to suit the current situation and on the basis of the experience gained in the past enquiry. However, for making valid conclusions in a repetitive enquiry, it should be ascertained that there is not any material change in the definitions of various terms used in the original enquiry.

2.5.3 CONFIDENTIAL OR NON-CONFIDENTIAL

In a confidential enquiry, the information collected and the results obtained are kept confidential and they are not made known to the public. The findings of such enquiries are meant only for the personal records of the sponsoring agency. The enquiries conducted by private organisations like trade unions, manufacturers' associations private business concerns are usually of confidential nature. On the other hand, the types of enquiries whose results are published and made known to the general public are termed as non-confidential enquiries. Most of the enquiries conducted by the state, private bodies or even individuals are of this type.

2.5.4 DIRECT OR INDIRECT

An enquiry is termed as *direct* if the phenomenon under study is capable of quantitative measurement such as age, weight, income, prices, quantities consumed and so no. However if the phenomenon under study in of a qualitative nature which is not capable of quantitative measurement like honesty, beauty, intelligence etc., the corresponding enquiry is termed as *indirect* one. In such an enquiry, the quantitative characteristic is converted into quantitative phenomenon by assigning appropriate standard which may represent the given attribute (qualitative phenomenon) indirectly. For example, the study of the attribute of intelligence may be made through the Intelligence Quotient (I.Q.) score of a group of individuals in a given test.

2.5.5 REGULAR OR AD-HOC

If the enquiry is conducted periodically at equal intervals of time (monthly, quarterly, yearly etc.) it is said to be *regular* enquiry. For example, the census is conducted in India periodically every 10 every 10 years. Similarly a number of enquiries are conducted by the Central Statistical Organisation (C.S.O.) and their results are published periodically such as Monthly Abstract to Statistics, Monthly Statistics of Production of Selected Industries of India, Statistical Abstract, (Annually) ; Statistical Pocket Book, India (Annually). On the other hand if an enquiry is conducted as and when necessary without any regularity or periodicity, it is termed as *ad-hoc.* For instance C.S.O. and N.S.S.O. (National Sample Survey Organisation) conduct a number of ad-hoc enquiries.

2.5.6 CENSUS (OR COMPLETE) OR SAMPLE

A census investigation is one in which all the elementary units committed with the problem are investigated. A sample investigation is one in which only a selected group of elementary units is investigated.

2.6 SELF HELP QUESTIONS

(a) Short Answer Questions –

(i) What is statistical investigation?

(ii) Is it necessary to plan a statistical survey?

(iii) What do you understand by data?

(iv) What are primary and secondary data?

(b) Long Answer Questions –

(i) Describe the preliminary steps which should be taken before commencing a statistical investigations.

(ii) What is statistical enquiry? Describe the types of statistical enquiry.

(iii) You have been asked to conduct a statistical enquiry. Outline the preliminary steps that your should take before commencing the work of collection of data.

UNIT-3 SAMPLING

Objectives

After studying this unit, you should be able to

- Distinguish between population parameter and sample statistics.
- Know the various procedures of sampling that provide an attractive means of learning about a population or process.
- Develop the concept of a sampling distribution that helps you to understand the methods and understanding thinking of statistical inference.

Structure

3.1 Introduction

3.2 Universe or population

3.3 Sampling

3.4 Census versus sample enumeration.

3.5 Limitation of sampling

3.6 Methods of sampling

3.7 Which method of sampling should be adopted?

3.8 How to decide the sample Size

3.9 Law of statistical regularity

3.10 Principle of inertia of large numbers

3.11 Self help questions

3.1 INTRODUCTION

The science of Statistics may be broadly studies under the following two headings:

(a) Descriptive,

(b) Inductive

So far, we have confined the discussion to *Descriptive Statistics* which consists in describing some characteristics of the numerical data. The *Inductive Statistics*, also known as *Statistical Inference*, may be termed as the logic of drawing statistically valid conclusions about the

totality of cases or items termed as *population*, in any statistical investigation on the basis of examining a part of the population, termed as sample, and which is drawn from the population in a scientific manner. In modern *'decision making process'* in different fields of human activity, including the ordinary actions of our daily life, most of our decisions and attitudes depend very much upon the inspection or examination of only a few objects or items out of the total lot. This process of studying only the sample data and then generalising the results to the population (i.e., drawing inferences about the population on the basis of sample study) involves an elements of risk, the risk of making wrong decisions. In this chapter will shall discuss the various techniques of drawing samples from the population.

3.2 UNIVERSE OR POPULATION

In any statistical investigation the interest usually lies in studying the various characteristics relating to items or individuals belonging to a particular group. This group of individuals under study is known as the *population* or *universe*. For example, if an enquiry is intended to determine the average per capita income of the people in a particular city the population will comprise all the earning people in that city. On the other hand if we want to study the expenditure habits of the families in that city, then the population will consist of all the households in that city. Further, if we want to study the quality of the manufactured product in an industrial concern during the day, then the population will consist of the day's total production. Thus, *"In Statistics, population is the aggregate of objects, animate or inanimate, under study in any statistical investigation"*. In sampling theory, the population means the larger group from which the samples are drawn.

A population containing a finite number of objects or items is known as *finite population* e.g., the students in a college, the day's production in an industrial concern, the population of a city or a town etc. On the other hand, a population having an infinite number of objects or with the number of objects so large as to appear practically infinite, is termed as an *infinite population*, e.g., the population of temperatures at various points of the atmosphere; the population of heights, weight or ages of people in the country (each of these variables can take any numerical value in a particular interval), the population of stars in the sky, etc. Infinite populations are better for sampling studies, the population may further be classified as *existent* or *hypothetical*. A population consisting of concrete objects is known as *existent* population, e.g., the population of (i) the books in a library, (ii) the aeroplanes in the Indian Air Force, (iii) the scooters in Delhi, etc. On the other hand, if the population does not consist of concrete objects, i.e., it consists of imaginary objects then it is called *hypothetical* population. for instance, the populations of the throes of a die or a coin, thrown infinite number or times are hypothetical populations.

3.3 SAMPLING

A finite subset of population, selected from it with the objective of investigating its properties is called a *sample* and the number of units in the sample is known as the *sample size*. Sampling is a tool which enables us to draw conclusions about the characteristics of the population after studying only those objects or items that are included in the sample.

The main objectives of the sampling theory are :

(i) To obtain the optimum results, i.e. the maximum information about the characteristics of the population with the available sources at our disposal in terms of time, money and manpower by studying the sample values only.

(ii) To obtain the best possible estimates of the population parameters.

Although the scientific development of the theory of sampling has taken place only during the last few decades, the idea of sampling is very old. From times immemorial, people have been using it without knowing that some scientific procedure has been used in arriving at the conclusions. On inspecting the sample of a particular stuff, we arrive at a conclusion about accepting or rejecting it. For example, the consumer examines only a handful of the rice, pulses or any commodity in a shop to assess its quality and then decides to buy it or not. The housewife, usually tastes a spoonful of the cooked products to ascertain if it is properly cooked and also to see if it contains proper quantity of salt or sugar. The consumer ascertains the quality of the grapes by testing one or two from the seller's basket. The intelligence of the individuals in a subject is estimated by the university by giving them a 3 hour test. A businessman orders for the products after examining only a sample from it. In fact, the entire business is done on the basis of display of a few specimen samples only. The error involved in approximations about the population characteristics on the basis of the sample is known as *sampling error* and is inherent ant unavoidable in any sampling scheme.

3.4 CENSUS VERSUS SAMPLE ENUMERATION

For any statistical enquiry in any field of human activity, whether it is in business, economics or social sciences, the basic problem is to obtain adequate and reliable data relating to the particular phenomenon under study. There are two methods of collecting the data:

(i) The Census Method or Complete Enumeration.

(ii) The Sample Method or Partial Enumeration.

1. Census Method. In the census method we resort to 100% inspection of the population and enumerate each and every unit of the population. In the sample method we inspect only a selected representative and adequate fraction (finite subset) of the

population and after analysing the results of the sample data we draw conclusions about the characteristics of the population.

The census method seems to provide more accurate and exact information as compared to sample enumeration as the information is collected from each and every unit of population. Moreover, it affords more extensive and detailed study. For instance, the population census conducted by the Government of India every 10 years collects information not only about the population but also obtains data relating to age, marital status, occupation, religion, education, employment, income, property, etc. The census method has its obvious limitations and drawback given below:

(i) The complete enumeration of the population requires lot of time, money, manpower and administrative personnel. As such this method can be adopted only by the government and big organisations who have vast resources at their disposal.

(ii) Since the entire population is to be enumerated, the census method is usually very time consuming. If the population is sufficiently large, then it is possible that the processing and the analysis or the data might take so much time that when the results are available they are not of much use because of changed conditions.

When to use Census Method? Census method is recommended in the following situations :

(a) If the information is required about each and every unit of the population, there is no way but to resort to 100% enumeration.

(b) In any manufacturing process in industry, 100% enumeration should be taken recourse to under the following conditions :

(i) The occurrence of a defect may cause loss of life serious casually to personnel.

(ii) A defect may cause serious malfunction of the equipment.

It may also be desirable to carry out complete census if,

(i) N, the lot size is small and

(ii) the incoming lot quality is poor or unknown.

2. Sample Method. The sample method has a number of distinct advantages over the complete enumeration method. **Prof. R.A. Fisher** sums up the advantages of sampling techniques over complete census in just four words: **Speed, Economy, Adaptability and Scientific Approach**. A property designed and carefully executed sampling plan yields fairly good results, often

better than those obtained by the census method. We summarise below the merits of the sample method over the census method:

1. **Speed, i.e., less time.** Since only a part of the population is to be inspected and examined, the sample method results in considerable amount of saving in time and labour. There is saving in time not only in conducting the sampling enquiry but also in the processing, editing and anyalysing the data. This is a very sensitive and important point for the statistical investigations where the results are urgently and quickly needed.

2. **Economy, i.e., Reduced Cost of the Enquiry.** The sample method is much more economical than a complete census. In a sample enquiry, there is reduction in the cost of collection of the information, administration, transport, training and man hours. Although, the labour and the expenses of obtaining information per unit are generally large in a sample enquiry than in the census method, the overall expenses of a sample survey are relatively much less, since only a fraction of the population is to be enumerated. This is particularly significant in conducting socio-economic surveys in developing countries with budding economies who cannot afford a complete census because of lack of finance.

3. **Administrative Convenience.** A complete census requires a very huge administrative set up involving lot of personnel, trained investigators and above all the co-ordination between the various operating agencies. On the other hand, the organisation and administration of a sample survey is relatively much convenient as it requires less personnel staff and the field of enquiry is also limited.

4. **Reliability.** In the census, the sampling errors are completely absent. If the non-sampling errors are also absent, the result would be 100% accurate. On the other hand, a sample enquiry contains both sampling and non-sampling errors. In spite of this weakness, a carefully designed and scientifically executed sample survey gives results which are more reliable than those obtained from a complete census. This because of the following reasons:

   (i) It is always possible to ascertain the extent of sampling error and degree of reliability of the results. Even the desired degree of accuracy can be achieved through sampling using different devices.

   (ii) The non-sampling errors such as due to measuring and recording observations, inaccuracy, or incompleteness of information, location of units, non-response or incomplete response, training of investigators, interpretation of questions, bias of the investigators, etc., are of a more serious nature in a complete census. In a sample survey these errors can be effectively controlled and minimised by (i) Employing highly qualified, skilled and trained personnel, (ii) imparting

adequate training to the investigators for conducting the enquiry, (iii) better supervision, (iv) using more sophisticated equipment and statistical techniques for the processing and analysis of the relatively limited data.

Lastly, follow-up work in case of non-response or incomplete response can be effectively undertaken in a sample survey than in a census. The effective reduction of the non-sampling errors in a sample survey more than compensates the errors in the estimates due to sampling procedure and thus provides relatively more accurate and reliable results. The efficiency of the sampling procedure lies in the fact that quite often the accuracy of the results of a complete census is ascertained by using some type of sampling check procedure.

5. Greater Scope. It appears that there is possibility of obtaining detailed information only in a complete census where each and every unit in the population is enumerated. But in practice because of our limitations in any statistical enquiry in terms of time, money and man hours and because of the fact that sampling procedure results in considerable savings in time, money and labour it is possible to obtain more detailed and exhaustive information from the limited few units selected in the sample. Obviously, it is relatively easy to collect information on, say 25 questions from each of 100 units selected in the sample than to obtain the information on, say, 10 questions from each of 1,000 units in the population. Moreover, complete enumeration in impracticable, rather inconceivable if the enquiry requires highly trained personnel and more sophisticated equipment for the collection, processing and analysis of the data. The sampling procedure is more readily adaptable than census for statistical investigations.

6. Infinite or/and Hypothetical Population. If the population is infinite or too large, then sampling procedure is the only way of estimating the parameters of a population. For instance, the number of fish in the sea or the number of wild elephants in dense forest can be estimated only by sampling method.

Similarly, in case of hypothetical population, as for example in the problem of throwing a die or tossing a coin where the process may continue large number of times or indefinitely, the sampling procedure is the only scientific technique of estimating the parameters of the population.

7. Destructive Testing. If the testing of units is destructive, i.e., if in the course of inspection the units are destroyed or affected adversely, then we are left with no other way but to resort to sampling. For example:

(i) to estimate the average life of the bulbs or tubes in a given consignment.

(ii) to determine the composition of a chemical salt.

(iii) to test the breaking strength of chalks manufactured in a factory or to estimate the tensile strength of the steel rods.

(iv) to test the quality of explosives, crackers, shells, etc.,

We have to inspect a representative sample since complete census will destroy all the items.

3.5 LIMITATIONS OF SAMPLING

The sample method of investigation has the following shortcomings/limitations:

1. **Unsuitable for higher order accuracy –** Sample investigation is not suitable if higher order accuracy is required.
2. **Misleading results –** The results of sample investigation may be false, inaccurate and misleading : (i) if sample is not drawn properly, (ii) if sample is not a representatives of the population, (iii) due to the absence of qualified and trained enumerators.
3. **False illusory conclusion –** The investigator's personal bias regarding the choice of units and drawing of sample lead to false conclusion.
4. **Unnecessary in case of small populations –** The sample method is unnecessary in case of small population. Sometimes the population may be so small that it may be impossible to draw a representatives sample from it.
5. **Unsuitable in case of heterogeneity –** When the population is not homogeneous then sample investigation is not so useful.
6. **Information for each and every unit is not available –** Sample investigation method is not suitable if information is required about each individual of the population.
7. **Details of the population are sacrificed –** The various samples taken from the same population with the same procedure would also be different in their values. Whatever the technique, the details of the population are sacrificed through sample investigation.
8. **Special knowledge required –** Sample investigation is a special technique and everyday cannot use it. Its use requires a specialised knowledge.

3.6 METHODS OF SAMPLING

The process of taking a sample is called sampling. The nomenclature of sample depends on the method of sampling. For example, the sample taken by random sampling method is known as random

sample. The sampling method or technique mainly depends upon the purpose of the investigation.

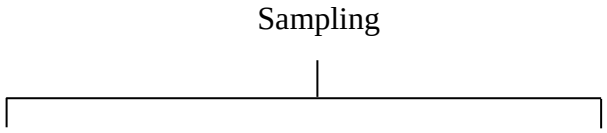
There are various methods of taking samples from a population. We shall consider a few or them. The three main methods of taking a sample are :

(1) Subjective or Purposive or Deliberate or Conscious Sampling,

(2) Probability or Objective Sampling,

(3) Mixed Sampling.

The sampling methods may also be classified as follows :



Probability Sampling (Random Sampling)	Non-probability Sampling (Non-random Sampling)
1. Simple Random Sampling	1. Purposive Sampling
2. Stratified Random Sampling	2. Quota Sampling
3. Systematic Sampling	3. Convenience Sampling
4. Multistage Random Sampling	4. Judgment Sampling
5. Cluster Sampling	5. Snowball Sampling

Remark : When we call simple random sampling as unrestricted random sampling then the stratified random sampling, systematic sampling, multistage random sampling, cluster sampling are put in the category restricted random sampling.

3.6.1 PROBABILITY SAMPLING METHODS (RANDOM SAMPLING)

Several probability sampling methods for selecting samples from a population or process are as follows:

(1) **Simple Random (Unrestricted) Sampling :** In this method, every member (or element) of the population has an equal and independent chance of being selected again and again when a sample is draw from the population. To draw a random sample, we need a complete list of all elements in the population of interest so that each element can be identified by a distinct number. Such a list is called *frame for experiement.* The frame for experiment allows us to draw elements from the population by randomly

generating the numbers of the elements to be included in the sample.

For instance, in drawing random sample of 50 students from a population of 3500 students in a college we make a list of all 3500 students and assign each student an identification number. This gives us a list of 3500 numbers, called frame for experiment. Then we generate by computer or by other means a set of 50 random numbers in the range of values from 1 and 3500. The procedure gives every set of 50 students in the population and equal chance of being included in the sample. Selecting a random sample is analogous to using a gambling device to generate numbers from this list.

This method is suitable for sampling, as many statistical tests assume independence of sample elements. One disadvantage with the method is that all elements of the population have to be available for selection, which many a times in not possible.

(2) **Stratified Sampling –** This method is useful when the population consists of a number of heterogeneous subpopulations and the elements within a given subpopulation are relatively homogeneous compared to the population as a whole. Thus, population is divided into mutually exclusive groups called *strata* that are relevant, appropriate and meaningful in the context of the study. A simple random sample, called a *sub-sample*, is then drawn from each *strata* or *group*, in proportion or a non-proportion to its size. As the name implies, a proportional sampling procedure requires that the number of elements in each stratum be in the same proportion as in the population. In non-proportional procedure, the number of elements in each stratum are disproportionate to the respective numbers in the population. The basis for forming the strata such as location, age, industry, gross sales, or number of employees, is at the discretion of the investigator. Individual stratum samples are combined into one to obtain an overall sample for analysis.

This sampling procedure is more efficient than the simple random sampling procedure because, for the same sample size, we get more representativeness from each important segment of the population and obtain more valuable and differentiated information with respect to each strata.

Disproportionate sampling decisions are made either when strata are either too small, too large, or when there is more variability suspected within a particular stratum. For example, the educational levels in a particular strata might be expected to influence perceptions, so more people will be sampled at this level. Disproportionate sampling is done when it is easier, and less expensive to collect data from one or more strata than from others.

For this method of sampling to be more effective in terms of reliability, efficiency, and precision, any stratification should be done which ensures.

(a) maximum uniformity among members of each strata.

(b) largest degree of variability among various strata.

(3) **Systematic Sampling.** This procedure is useful when elements of the population are already physically arranged in some order, such as an alphabetized list of people with driving licenses, list of bank customers by account numbers. In these cases one element is chosen at random from first *k* element and then every kth element (member) is included in the sample. The value *k* is called the sampling interval. For example, suppose a sample size of 50 is desired from a population consisting of 1000 accounts receivable. The sampling interval is k=N/n = 1000/50=20. Thus a sample of 50 account is identified by moving systematically through the population and identifying every 20th account after the first randomly selected account number.

(4) **Multistage Sampling.** This method of sampling is useful when the population is very widely spread and random sampling is not possible. The researcher might stratify the population in different regions of the country, then stratify by the urban and rural and then choose a random sample of communities within these strata. These communities are then divided into city areas as clusters and randomly consider some of these for study. Each element in the selected cluster may be contacted for desired information.

For example, for the purpose of a national pre-election opinion poll, the *first stage* would be to choose as a sample a specific state (region). The size of the sample, that is the number of interviews, from each region would be determined by the relative populations in each region. In the *second stage*, a limited number of towns/cities in each of the regions would be selected, and then in the *third stage*, within the selected towns/cities, a sample of respondents could be drawn from the electoral roll of the town/city selected at the second stage.

The essence of this type of sampling is that a subsample is taken from successive groups or strata. The selection of the sampling units at each stage may be achieved with or without stratification. For example, at the second stage when the sample of towns/cities is being drawn, it is customary to classify all the urban areas in the region in such a way that the elements (towns/cities) of the population in those areas are given equal chances of inclusion.

(5) **Cluster Sampling.** This method, sometimes known as *area sampling method*, has been devised to meet the problem of costs or inadequate sampling frames (a complete listing of all elements in

the population so that each member can be identified by a distinct number). The entire population to be analysed is divided into smaller groups or chunks of elements and a sample of the desired number of areas selected by a simple random sampling method. Such groups are termed as *clusters*. The elements of a cluster are called *elementary units.* These clusters do not have much heterogeneity among the elements. A household where individuals live together is an example of cluster.

If several groups with intragroup heterogeneity and intergroup homogeneity are found, then a random sampling ot the cluster or groups can be done with information gathered from each of the elements in the randomly chosen clusters. Cluster samples offer more heterogeneity within groups and more homogeneity among groups – the reverse of what we find in stratified random sampling, where there is homogeneity within each group and heterogeneity across groups.

For instance, committee formed from various departments in an organization to offer inputs to make decisions on product development, budget allocations, marketing strategies, etc are examples of different clusters. Each of these clusters or groups contains a heterogeneous collection of members with different interest, orientations, values, philosophy, and vested interests. Based on individual and combined perceptions, it is possible to make final decision on strategic moves for the organization.

3.6.2 NON-PROBABILITY SAMPLING (NON-RANDOM SAMPLING METHODS)

Several non-random sampling methods for selecting samples from a population or process are as follows :

(1) **Purposive Sampling :** Instead of obtaining information those who are most conveniently available, it sometimes becomes necessary to obtain information from specific targets-respondents who will be able to provide the desired information either because they are the only ones who can give the desired information or because they satisfy to some criteria set by researcher.

(2) **Quota Sampling :** Quota Sampling is a form of proportionate stratified sampling in which a predetermined proportion of elements are sampled from different groups in the population, but on convenience basis. In other words, in quota sampling the selection of respondents lies with the investigator, although in making such selection he/she must ensure that each respondent satisfies certain criteria which is essential for the study. For example, the investigator may choose to interview ten men and ten women in such as way that two of them have annual income of more than two lakh rupees, five of them have annual income between one and two lakh rupees and thirteen whose annual

income is below one lakh rupees. Furthermore, some of them should be between 25 and 35 years of age, others between 36 and 45 years of age, and the balance over 45 years. This that the investigator's choice of respondent is partly dictated by these 'controls'.

Quota sampling has been criticized because it does not satisfy the fundamental requirement of a sample, that is, it should be random. Consequently, it is not possible to achieve precision of results on any valid basis.

(3) **Convenience Sampling.** In this procedure, units to be included in the sample are selected at the convenience of the investigator rather than by any prespecified or known probabilities of being selected. For example, a student for this project on 'food habits among adults' may use his own friends in the college to constitute a sample simply because they are readily available and will participate for little or no cost. Other examples are, public opinion surveys conducted by any TV channel near the railway station; bus stop, or in a market.

Convenience samples are easy for collecting data on a particular issue. However, it is not possible to evaluate its representativeness of the population and hence precautions should be taken in interpreting the results of convenient samples that are used to make inferences about a population.

(4) **Judgment Sampling.** Judgment sampling involves the selection of respondents who are in the best position to provie the desired information. The judgment sampling is used when a limited number of respondents have the information that is needed. In such cases, any type of probability sampling across a cross section of respondents is purposeless and not useful. This sampling method may curtail the generalizability of the findings due to the fact that we are using a sample of respondents who are conveniently available to us. It is the only viable sampling method for obtaining the type of information that is required from very specific section of respondents who possess the knowledge and can give the desired information.

However, the validity of the sample results depend on the proper judgment of the investigator in choosing the sample. Great precaution is needed in drawing conclusions based on judgment samples to make inferences about a population.

3.7 WHICH METHOD OF SAMPLING SHOULD BE ADOPTED?

The basic idea behind sampling is that the units selected will, on the whole, represent fairly correctly the characteristics of the universe or

population from which they have been derived. The problem in which method is superior. The answer is that no single method is appropriate in all the situations. However, the following fundamental factors are to be considered while adopting a sampling method:

(i) Purpose of the investigation

(ii) Type of Universe or Nature of the Universe

(iii) Type of the sampling units

(iv) Sources of information

(v) Size of the sample

3.8 HOW TO DECIDE THE SAMPLE SIZE?

Size of sample means the number of sampling units selected from the population for investigation. Thus, size of sample is directly related to standard of accuracy, unit cost and administration of the investigation.

The larger the size of sample, the greater will be representation of the units of the population in it; but is brings difficulties of managing it. In large-sized samples intensive study of the units becomes difficult. If the size of the samples is kept small, it will not represent the Universe ant the degree of accuracy in the results will be limited.

Hence, the size of a sample should be neither too big nor too small. It should be optimum. According to Parten, "An optimum sample survey size is one which fulfils the requirements of efficiency, representativeness, reliability and flexibility." The size of the sample should ensure minimum of cost and minimum of standard error. A small sample has a large standard error compared with a large sample of the same characteristics, while the cost will be less in the first as compared to latter. It should also be borne in mind that "Mere size, of course, does not ensure representativeness in a sample." According to **Croxton** and **Cowden**, *"A small random sample or stratified sample is apt to be much superior to a large but badly selected sample."*

The size of the sample which is required in order to make a valid inference about a population is dependent upon the following factors:

(i) **Homogeneity or heterogeneity of the Universe –** In a universe consisting of homogeneous units a small sample is suitable, while in a universe consisting of heterogeneous units, large sample is inevitable.

(ii) **Number of Class proposed –** If the proposed number of classes is large, the sample size should be large.

(iii) **Nature of Study –** For an intensive and continuous study small samples will be suitable. For general survey the size of the sample

should be large but for technical studies the size of the sample may be kept small.

(iv) **Practical Consideration –** The availability of finance, time and trained personnel are some practical considerations which affect the size of the sample.

(v) **Standard of accuracy –** Though it is viewed that larger the size of the sample, greater will be the degree of accuracy, yet this is not always so.

(vi) **Type of Sampling –** In random sampling, greater accuracy in results will be achieved only in a large sample. In a properly drawn stratified sampling, a small sample can give better results.

(vii) **Nature of Units –** When it is expected that a large number of units will not respond, then a large sample should be taken.

(viii) **Size of Questionnaire –** If the size of questionnaire is large and it contains difficult questions, the size of the sample should be kept small.

(ix) **Size of the Universe –** The larger the size of the universe, the larger should be the size of the sample.

3.9 LAW OF STATISTICAL REGULARITY

This law has its origin in the mathematical theory of probability. In the words of **L.R. Conner**, **"The law of statistical regularity lays down that a group of objects chose at random from a larger group tends to possess the characteristics of that large group (universe)."** According to **King"the law of statistical regularity lays down that the moderately large number of items chosen at random from a large group are almost sure on the average to possess the characteristics of the large group."**

The principle of statistical regularity impress upon the following two points :

(i) **Large sample size.** Logically, it seems that as the sample size increases, the sample is more likely to reveal the true characteristics of the population and thus provide better estimates of the parameters. It is known that the reliability of the sample statistic as an estimate of the population parameter is proportional to the square root of the sample size n. But due to certain limitations in terms of time, money and manpower, it is not always possible to take very large samples. Moreover, the effort and cost of drawing large samples might outlive the utility of sample study as against the complete enumeration (census).

(ii) **Random selection.** The sample should be selected at random from the population. By random selection we mean a selection in which each and every unit in the population has an equal chance of being selected in the sample.

If a sample is selected such that the above two conditions are satisfied, then it will depict the true characteristics of the population fairly accurately and can be used for drawing valid inferences about the population. For example, if we are interested in studying the average height of the students in Delhi University, then it is not desirable to resort to 100% enumeration of the students in the university. A fairly adequate sample of the students from each college may be selected at random and the average height of the students selected in the samples may be computed. Since the sample is random, it would be representative of the population and the average so obtained will not differ much from the true value (i.e., the average computed by the complete enumeration). This difference is attributed to fluctuations of sampling.

3.10 PRINCIPLE OF INERTIA OF LARGE NUMBERS

An immediate deduction from the Principle of Statistical Regularity is the *Principle of Inertia of Large Number* which states, *"Other things being equal, as the sample size increases, the results tend to be more reliable and accurate".* This is based on the fact that the behaviour of a phenomenon *en masse*. i.e. on a large scale is generally stable. By this we mean that if individual events are observed, their behaviour may be erratic and unpredictable but when a large number of events are considered they tend to behave in a stable pattern. This is because a number of forces operate on the given phenomenon and if the units are large, then the typical odd variations in one part of the universe in one direction will get neutralised by the variations in equally bigger part of the universe in the other direction. According to **A.L. Bowley**, *"Great numbers and averages resulting from them, such as we always obtain in measuring social phenomena have a great inertia."* Thus in dealing with large numbers, the variations in the component parts tend to balance each other and consequently the variation in the aggregate result is likely to be insignificant. However, it should not be inferred that in case of large numbers, there is no variation at all. Large numbers are relatively more stable in their characteristics than the small numbers. They (large numbers) also exhibit variations but they are of very small magnitude and intensity and are not violent in nature. For example, if a coin is tossed, say, 20 times then nothing can be said with certainty about the proportion of heads. We may 0, 1, 2...., or even all the 20 heads. But it is thrown at random a very large number of times, say, 5,000 times. then we may expected on the average 50% heads and 50% tails. As another illustration let us consider the production of a particular commodity, say, rice in two

districts in state for a number of years. The figures will show great variations due to favourable or unfavourable conditions in that particular region. However, the figures for the production of rice of the whole state over number of years will show relatively lesser variations because lower production in some of district will be compensated by the excessive production in some other districts of the state. Arguing similarly we find that the production figures for the whole of India will sow still lesser variation and for the entire world in would be more or less stable.

3.11 SELF HELP QUESTIONS

(a) Short answer questions :

1. What is statistical sampling?
2. Distinguish between a census survey and sample survey.
3. What is law of statistical regularity?
4. What do understand by random sampling?
5. What is law of inertia of large number?
6. What is meant by non-random sampling?
7. Write note on the following –
    a. Systematic sampling
    b. Judgment sampling
    c. Quota sampling
    d. Cluster sampling
    e. Multistage sampling

(b) Long Answer Questions –

1. Discuss the relative advantage and disadvantages of the census and sampling methods of calculations of statistics.
2. What is 'Statistical Sampling'? Describe the various methods of Statistical sampling.
3. "A good sample must be based on random selection". Elucidate.
4. "Sampling is necessary under certain condition?" Illustrate this by suitable example.
5. State and explain the law of statistical regularity and law of inertia of large numbers. How do they help the investigator in his work.

UNIT-4 ACCURACY OF DATA

Objectives

After studying this unit, you should be able to

- Know the statistical units are to be used in investigation for collection and as well as comparison.
- Understand the significance of reasonable accuracy and degree of accuracy required.

Structure

4.1 Statistical units

4.2 Feature of a good statistical units

4.3 Types of units

4.4 Accuracy

4.5 Perfect Accuracy

4.6 Significance of reasonable accuracy

4.7 Degree of accuracy

4.8 Self help questions

4.1 STATISTICAL UNITS

A well defined and identifiable object or a group of objects with which the measurements or counts in any statistical investigation are associated is called a *statistical unit*. For example, in a socio-economic survey the unit may be an individual person, a family, a household or a block of locality A very important step before the collection of data begins is to define clearly the statistical units in which the data are to be collected. In a number of situations the units are conventionally fixed like the **physical units** of measurement such as metres, kilometres, kilograms, quintals, hours, days, weeks etc., which are well defined and do not need any elaboration or explanation. However in many statistical investigations, particularly relating to socio-economic studies, *arbitrary units* are used which must be clearly defined. This is imperative since in the absence of a clear cut and precise definition of the statistical units, serious errors in the data collection may be committed in the sense that we may collect irrelevant data on the items, which should have, in fact, been excluded and

omit data on certain items which should have been included. This will ultimately lead to fallacious conclusions.

4.2 FEATURES OF GOOD STATISTICAL UNIT

The following points might serve as guidelines for deciding about the unit in the any statistical enquiry.

1. **It should be unambiguous :** A statistical unit should be rigidly defined so that it does not lead to any ambiguity in its interpretation. The units must cover the entire population and they should be distinct and non-overlapping in the sense that every element of the population belongs to one and only one statistical unit.

2. **It should be specific :** The statistical unit must be precise and specific leaving no chance to the investigators. Quite often, in most of the socio-economic surveys the various concepts/characteristics can be interpreted in different variant forms and accordingly the variable used to measure it may be defined in several different ways. For example, in an enquiry relating to the wage level of workers in an industrial concern the wages might be weekly wages, monthly wages or might refer to those of skilled labour only or of day workers only or might include bonus payments also. Similarly prices in an enquiry might refer to cost prices, selling prices, retail prices, whole-sale prices or contract prices. Thus in a statistical enquiry it is important to distinguish between the **conventional** and the **arbitrary** definitions of the characteristics/ variables, the former being the one prevalent in common use and shall always remain same (fixed) for every enquiry while the latter is the one which is used in a specific sense and refers to the working or operational definition which will keep on changing from one enquiry to another enquiry.

3. **It should be stable :** The unit selected should be stable over a long period of time and also w.r.t. places i.e., there should not be significant fluctuations in the value of a unit at different intervals of time or at different places became in the contrary case, the data collected at different times or places will not be comparable and this would mar their utility to a great extent. The fluctuations in the value of money different times (due to inflation) or in the measurement of weights at different places (due to height above sea level) might tender the comparisons useless. Thus, the unit selected should imply, far as possible, the same characteristics at different times or at different places.

4. **It should be appropriate to the enquiry :** As already pointed out earlier, the concept and definition of arbitrary statistical units keep on changing from enquiry to enquiry. The unit selected must be

relevant to the given enquiry. Thus, for studying the changes in the general price level, the appropriate units is the whole-sale prices while for constructing the cost of living indices (or consumer price indices) the appropriate unit is the retail prices.

5. **It should be uniform :** It is essential that the unit adopted should be homogeneous (uniform) throughout the investigation so that the measurements obtained are comparable. For example, in measuring length if we use a yard on some occasions and metre on other occasions in an investigation, the observations obtained would be confusing and misleading.

4.3 TYPES OF STATISTICAL UNITS

The statistical units may be broadly classified as follows:

(i) Units of collection.

(ii) Units of analysis and interpretation.

4.3.1 UNITS OF COLLECTION

The units of collection may further be sub-divided into the following two classes:

(a) Unit of Enumeration : In any statistical enquiry, whether it is conducted by 'sample' method or 'census' method, unit of enumeration is the basic unit on which the observations are to be made and this unit is to be decided in advance before conducting the enquiry keeping in view the objectives of the enquiry. The unit of enumeration may be a person, a household, a family, a farm (in land experiments), a shop, a live stock, a firm etc. As has been pointed out earlier, this unit should be very clearly defined in terms of shape, size etc. For instance, for the construction of cost of living index number, the proper unit of enumeration is household. It should be explained in clear terms whether a household consists of a family comprising blood relations only or people taking food in a common kitchen or all the persons living in the house or the persons enlisted in the ration card only. The concept of the household (to be used in the enquiry) is to be decided in advance and explained clearly to the enumerators so that there are no essential omissions or irrelevant inclusions.

(b) Units of Recording : The units of recording are the units in terms of which the data are recorded or in other words they are the units of quantification. For instance, in the construction of cost of living index number (consumer price index) the data to be collected from each household, among other things, include the retail prices of various commodities together with the quantities consumed by the class of people for whom the index is meant. The units of

recording for quantity may be weight (in case of food-grains), say, in kilograms, quintals, tons, etc., in case of clothing the unit of recording may be metres; the price may by recorded in terms of rupees and so on.

Units of measurement (recording) may be *simple* or *composite*. The units which represent only one condition without any qualification (adjective) are called *simple units* such as metre, rupee, ton, kilogram, pound, bale of cloth, hour, week, year, etc. Such units are generally *conventional* and not all difficult to define. However, sometimes care has to be taken in their actual usage. For example, the bale of cloth must be defined in terms of length, say, 20 metres, 50 metres or 100 metres. Similarly, in case of weight it should be clearly specified whether it is net weight or gross weight.

A simple unit with some qualifying words in called a *composite unit*. A simple unit with only one qualifying word is called a *compound unit*. Examples of such units are skilled worker, employed person, ton – kilometre, kilowatt hour, man hours, retail prices, monthly wages, passenger kilometres. For instance ton – kilometre means the number of tons multiplied by the number of kilometres carried ; man hours implies the total number of workers multiplied by the number of hours that each worker has put in and so on. If two or more qualifying words are added to a simple unit, it is called a *complex unit* such as production per machine hour, output per man hour and so on. Thus as compared to simple units, composite (compound and complex) units are much more restrictive in scope and difficult to define. Such units should be defined properly and clearly as they need explanation about the unit used and also about the qualifying words.

4.3.2 UNITS OF ANALYSIS AND INTERPRETATION

As the name implies, the units of analysis and interpretation are those units in the form of which the statistical data are ultimately analysed and interpreted. Is should be decided whether the results would be expressed in absolute figures or relative figures. The units of analysis and interpretation facilities comparisons between different sets of data with respect to time, place or environment (conditions). Generally the units of analysis are rates, ratios and percentages, and coefficients.

Rates involve the comparison between two heterogeneous quantities i.e., when the numerator and denominator are not of the same kind e.g., the mortality (death) rates, the fertility (birth) rates and so on. Rates are usually expressed per thousand. For instance, the Crude Birth Rate (C.B.R.) is the ratio of total number of live births in the given region or locality during a given period to the total population of that region or locality during the same period, multiplied by 1,000. Rate per unit is called coefficient. However, ratios and percentages are used for comparing homogeneous quantities e.g., when the numerator and denominator are of

the same kind. For example, "the ratio of smokers to non-smokers in a particular locality is 1:3" implies that 25% of the population are smokers.

From practical point of view of comparing data relating to different series, usually the unit of analysis is one which gives relative figures which are pure numbers independent of units of measurement. For instance, if we want to compare two series for variability (dispersion) the appropriate unit of analysis Coefficient of Variation and for comparing symmetry of two distributions, we study the Coefficient of Skewness.

4.4 ACCURACY

Perfect accuracy is hardly present in statistical investigation. According to **King**, *"Attempt to obtain the greatest possible degree of a accuracy are, frequently, merely waste of time."*

Perfect accuracy is also not possible is scientific measurement also, but to a very extent they are accurate in comparison to statistical measurements.

4.5 PERFECT ACCURACY

Perfect accuracy means that facts are presented in purely true form. In any branch of applied human knowledge perfect accuracy is not found. Even in the field of science, from the study of microbiology to sending rocket in space, exact measurements are not existing. In social sciences human beings or their activities are measured in which exact measurements are not possible. Statistics is regarded as the science of estimates and probabilities. In estimates, perfect accuracy is unattainable. In statistical investigations, accuracy in mathematical or accounting sense is neither possible nor desirable.

4.6 SIGNIFICANCE OF REASONABLE ACCURACY

The reasons for not attaining perfect accuracy are (1) Imperfections of the investigator, and (2) Imperfection of the instruments and tools of measurement. Man is not perfect; moreover, he has his own feelings of pride, bias, etc. When he collects facts, naturally, his mind also works according to his feelings. Acts have a very high degree of correlation with our feelings of heart and mind. Secondly, the statistical tools and instruments are not as perfect as the tools of a scientist's laboratory. Experimentation is not possible in statistical studies. Therefore, we should not expect perfect accuracy in statistical investigations. It is not desired also. In statistical investigations, there will be prejudices may-be deliberate or subconscious. Under such circumstances we can only arrive at a reasonable standard or accuracy only. But nonetheless, the science of statistics helps us in understanding

the factual world with all its inaccuracy and imperfections. We have to be content with a reasonable degree of accuracy which depends upon its practical value in relation to its costs.

4.7 DEGREE OF ACCURACY

A decision regarding the degree of accuracy or precision desired by the investigator in his estimates or results is essential before starting any statistical enquiry. An idea about the precision aimed at is extremely helpful in deciding about the method of data collection and the size of the sample (if the enquiry is to be on the basis of a sample study). The information gained from any previous completed sample study on the subject in the form of precision achieved for a given sample size may serve as a useful guide in this matter provided there is no fundamental reason to change this empirical basis. In any statistical enquiry perfect accuracy in final results in practically impossible to achieve because of the errors in measurement, collection of data, its analysis and interpretation of the results. However, even if it were attainable, it is not generally desirable in terms of time and money likely to be spent in attaining it and a reasonable degree of precision is enough to draw valid inferences.

A decision regarding the precision of the results very much depends upon the objectives and scope of the enquiry. For example, if we are measuring the length of cloth for shirting or pant, a difference of centimetres is going to make substantial impact. But if we are measuring the distances between two places, say Delhi and Bombay, a difference of few metres may be immaterial and while measuring the distance between two distant places, say Delhi and New York (U.S.A.) a difference of few kilometres may be immaterial. Likewise in measuring cereals (rice, wheat etc.) a difference of few grams may not matter at all whereas in measuring gold even $1/15^{th}$ of $1/20^{th}$ part of a gram is going to make lot of difference. In the words of **Riggleman** and **Frishbee**, *"the necessary degree of accuracy in counting or measuring depends upon the practical value of accuracy in relation to its cost."* However it should not be misunderstood to imply that one should sacrifice accuracy to conduct enquiry at low cost.

4.8 SELF HELP QUESTIONS

(a) Short Answer Questions –

1. What do you know by statistical unit?
2. What is reasonable accuracy
3. Write short notes on the following –
   (a) Accuracy
   (b) Degree of Accuracy

(c) Units of Analysis

(b) Long Answer Questions –

1. Define a statistical unit and explain what should be the essential requirement of a good statistical unit.

2. What is statistical unit? What do you mean by units of collection and units of analysis.

3. Explain the concept of accuracy is statistics.

4. What standard of accuracy is needed in statistical investigation?

Uttar Pradesh Rajarshi Tandon
Open University

Bachelor of Business Administration

BBA-107

Business Statistics

BLOCK

2

COLLECTION AND CLASSIFICATION OF DATA

UNIT-5 **61**

Approximation and Errors

UNIT-6 **73**

Collection of Data-Primary Source

UNIT-7 **91**

Collection of Data-Secondary Source

UNIT-8 **113**

Classification of Data

Curriculum Design Committee

Prof. Omji Gupta **Director**

School of Management Studies, UPRTOU, Prayagraj

Dr Gyan Prakash Yadav

Assistant Professor

School of Management Studies, UPRTOU, Prayagraj

Dr Devesh Ranjan Tripathi

Assistant Professor

School of Management Studies, UPRTOU, Prayagraj

Course Preparation Committee

Dr T.M. Raini **Author**

Associate professor, Firoj Gandhi PGCollege, Raibarely

Dr Shiv Bushan Gupta

Associate professor, MMM PG college, Kalakankar, Pratapgarh

Dr Gaurav Sankalp

Academic consultant, School of Management Studies, UPRTOU, Prayagraj

Dr Ajay Tiwari

Assistant Professor

MONIRBA, Department of Commerce, University of Allahabad

Prof J.N. Mishra **Editor**

Professor, MONIRBA, Department of Commerce, University of Allahabad

Dr Gyan Prakash Yadav **Coordinator**

Assistant Professor

School of Management Studies, UPRTOU, Prayagraj

©UPRTOU, Prayagraj-2021

ISBN :

©All Rights are reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the **Uttar Pradesh Rajarshi Tondon Open University, Prayagraj.**

BLOCK INTRODUCTION

This block comprises four units. The first unit of this block is concerned withApproximation & Errors. This unit presents an over view of methods ofapproximation, errors in statistics bised and unbaised errors their estimation & also sampling and non sampling errors. The next unit deals with collectionof primary data & its merits and demerits. The third one unit named Collectionof Secondary data, explains regarding sources of Secondary data, likePublished & unpublished data. The last unit of this block is related withclassification of data, its objectives, classification and construction of frequencydistrubution.

UNIT-5 APPROXIMATION & ERRORS

Objectives

After going through this unit you should be able to know about :

- Approximation
- Errors of approximation
- Estimation of Errors

Structure

5.1 Introduction

5.2 Measurement of Errors

5.2.1 Absolute Error

5.2.2 Relative Error

5.2.3 Percentage Error

5.2.4 Standard Errors of Estimate

5.2.5 Farward& Backward Error

5.3 Approximation in statistics

5.4 Truncation Error

5.5 Conclusion

5.6 Further Study

5.1 INTRODUCTION

The word ‘error’ is used in statistics in a special sense. In statistics the word ‘error’ means the difference between the true value and the estimated or approximated value of an item the word mistake in statistics a wrong calculation or use of inappropriate method in the collection or analysis of data. Another difference is that where error may be caused due to defect in measuring instruments, mistake is done because of ommission of commission.

Example : If we have to measure a piece of cloth which is 20 meters in length and as we proceed to measure it we omit to count 2 metres. Consequently a piece of 20 meters is actually measured as 18 meters. It is a mistake. On the other hand, if we do not omit to count 2 metres but

because of measuring instrument being one centimetre or two longer, we measure the cloth a bit less than 20 metres, it will be an error.

5.2 MEASUREMENT OF ERRORS

Where estimates are involved, statistical errors are bound to arise. It is, therefore, necessary to measure their magnitude and find out the limits within which our results can vary. From this point of view, there are two classes of errors :

(1) Absolute errors

(2) Relative errors

5.2.1 ABSOLUTE ERROR

Estimates are matters of subjective consideration. So when errors are measured absolutely, it is known as absolute error. It is the difference between the true value and the estimated value or approximated value of an item. It arises in two forms. It may be either a positive error or a negative error depending upon the case whether the actual value or estimated value is more.

Suppose, many students are sitting in a class. The teacher estimated their number as 60 but when they were actually counted it was found that the number of students actually present in the class was 62. So, absolute error is 2 (62-60). It is a positive error. It can be expressed in the form of a formula as follows :

A.E. = A.V.-E.V.

Where : A.E. stands for absolute Error

A.V. stands for Actual Value

E.V. stands for Estimated Value

But if the actual number of students is less than the estimated number, viz actual number is 58, then it will be negative error (58-60 = -2).

5.2.2 RELATIVE ERROR

Relative error is that error which is measured relatively. It is the ratio which the absolute error bears to the estimated or approximated value. Relative error can be expressed in the form of a formula as follows :

$$AE = \frac{\text{AE}}{\text{E. V.}}$$

Where : R.E. stands for Relative Error

A.E. stands for Absolute Error

E.V. stands for Estimated Value

Taking the figures from earlier example where absolute error is 2 and the estimated value is 60, the relative error will be :

$$R.E. = \frac{2}{60} = 0.033$$

5.2.3 PERCENTAGE ERROR

It a relative error is expressed in percentage, it is known as percentage error. Relative error multiplied by 100 gives us percentage error. It can be expressed in the shape of a formula as follows :

$$\text{P.E.} = \frac{A.E.}{E.V.} \times 100$$

where P.E. stands for Percentage Error.

Taking the figures from above example, percentage error will be :

$$\text{P.E.} = \frac{2}{60} \times 100 = 3.33\%$$

The relative error is more expressive of the actual magnitude of error than the absolute error. It can be easily shown by means of a simple example. The above illustration shows an absolute error of +2. In a public meeting addressed by Prime Minister, it was estimated that there were 20,000 people. When they were actually counted, the numbers came to 20,002 giving an error of +2. This shows that there is really no difference between the absolute errors in the two examples taken above. But it is obvious that an error of +2 on an actual strength of 62 is much more than a similar error on an actual strength of 20,002. The relative error brings out the position clearly. The relative error in the first case is 0.033 and in the second case it is 2 20000 or 0.0001. Thus, a relative error is a better guide of the degree of error than the absolute error.

5.2.4 STANDARD ERROR OF ESTAIMATE

The regression equations enable us to estimate or predict the values of Y for values of X or the values of X for values of Y. However, these are only predictions and cannot be treated as precise values. For example, if on the basis of regression studies, we find that when price of a commodity (X) is 40, its likely demand (Y) would be 130. This figure of 130 which we have got from the regression equation of Y on X is a mere estimate. The actual demand may be more or less than 130. Under such a situation, it is always better to find the likely error in the estimated values of Y or X.

This is done by calculating the Standard Error of the Estimate. This is analogous to standard deviation. In fact, if we find out the deviation of the original values of Y from its estimated values, square these deviations, total the squares and divide the total by the number of items and find out the square root of the figure we will get the Standard Error of the Estimate

of Y. In other words, if Sy is the standard error of the estimate of Y values, then,

$$S_y\sqrt{\frac{\Sigma(Y-Y_e)^2}{N}}$$

where Y represent the original values and Y_e the estimated value of Y. Similarly, the standard error of the estimate of X values, if represented by S_x would

be :

$$S_x=\sqrt{\frac{\Sigma(X-X_e)^2}{N}}$$

where X represents the original and Xe the estimated values of X.

If the sample size is small, it is better to take $S_y=\sqrt{\frac{\Sigma(Y-Y_e)^2}{N-1}}$ and $S_x=\sqrt{\frac{\Sigma(X-X_e)^2}{N-1}}$ This is on the basis of degrees of freedom which is discussed later in Chapters on Sampling and Tests of Significance.

In fact, $\Sigma(Y-Y_e)^2$ and $\Sigma(X-X_e)^2$ represent the 'Unexplained Variation' in Y and X, respectively. That is why, sometimes, the standard error of the estimates is represented as follows :

$$S_y=\sqrt{\frac{\text{Unexplained Variation in Y}}{N}} \quad \text{and}$$

$$S_x=\sqrt{\frac{\text{Unexplained Variation in X}}{N}}$$

However, the calculation of S_y and S_x by the above methods is tedious because we have to find out all the estimated values of Y for given values of X for the calculation of S_y. Similarly, we have to find out all the estimated values of X for given values of Y to calculate S_x. Therefore, a simpler way to calculate S_y and S_x is

$$S_y=\sqrt{\frac{\Sigma Y^2-a\Sigma Y-b\Sigma XY}{N}} \text{ and}$$

$$S_x=\sqrt{\frac{\Sigma X^2-a\Sigma X-b\Sigma XY}{N}}$$

[For small samples, we use (N – 1) instead of N]

In the above formulae we have directly introduced the technique of the method

of least squares in the calculation of S_y and S_x.

The value of S_y is interpreted in the same way as the standard deviation about the mean. If about the line of relationship, there is a normal distribution, 68

per cent of all the cases would lie within a range of ± S, 95.45% within ± 2 S and 99.7% within ± 3 S. If the dispersion about the line of equation is less, the value of S would also be less; if it is more, the value of S would also be more. The value of S is, therefore, an indicator of the significance of the computed values.

If the average deviaiton of the original data of Y series is calculated, it also gives a measure of significance of the arithmetic average from which deviations are calculated. The probable values of Y can be obtained either by the equation of the best fit or arithmetic average of Y-series and its standard deviation. The question that arises at this stage is which of the two estimates is better. Usually, a relationship is established between the two measures which is also helpful in the calculation of coefficient of correlation. Such a relationship is obtained by dividing the standard error of the estimates by the standard deviation of the Y series, or, in other words, such a relationship is expressed by finding out the value of $\frac{S_y}{\sigma_y}$. This is usually called measure of correlation. A better measure of correlation is obtained by finding out the Coefficient of correlation which is obtained by the following formula :

Coefficient of correlation, or $r = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}} = \sqrt{\frac{\text{Explained Variance}}{\text{Total Variance}}}$

In this formula, if there is no dispersion about the line of relationship, the value of S_y would be 0 and the value of the coefficient of correlation would be

1. The maximum value of y σ is that which is equal to S_y. In such cases, $\frac{S_y^2}{\sigma_y^2}$ would be equal to 1 and the coefficient of correlation would be equal to 0. It will indicate that there is no correlation between values of X and Y. Thus, the two limits of this correlation are 0 and 1. The higher the coefficient of correlation, the greater the confidence that may be placed on the computed values of Y and, as such, the greater the degree of correlation between the two series. The following examples would illustrate the calculation of the Standard Error of the Estimates.

Example 27. From the data given below, find the two regression equation and calculate the Standard Errors of the estimate of $Y(S_y)$ and $Y(S_x)$

x	2	4	6	8	10
y	5	7	9	8	11

Solution : The problem has been solved earlier (example 1 of this chapter) where we found out the two regression equations. These equations are :

$$Y_e = 0.65X + 4.1 \qquad X_e = 1.3Y - 4.4$$

From the above regressions we can find out the values of S_y and S_x by finding out the computed values of Y and X.

		Computed		Computed	
X	Y	values of Y		values of X	
		Y_e	$(Y - Y_e)^2$	X_e	$(X - X_e)^2$
2	5	5.4**	0.16	2.1***	0.01
4	7	6.7	0.09	4.7	0.49
6	6	8.0	1.00	7.3	1.69
8	8	9.3	1.69	6.9	4.00
10	11	10.6	0.16	9.9	0.01
$\Sigma X = 30$	$\Sigma Y = 40$	$\Sigma Y_e = 40$	$\Sigma (Y - Y_e)^2 = 3.10$	$\Sigma X_e = 30$	$\Sigma (X - X_e)^2 = 6.20$

$$S_y = \sqrt{\frac{\Sigma(Y - Y_e)^2}{5}} = \sqrt{\frac{3.10}{5}} = \sqrt{0.62} = 0.7874$$

$$S_x = \sqrt{\frac{\Sigma(X - X_e)^2}{5}} = \sqrt{\frac{6.20}{5}} = 1.1136$$

Now, if we calculate the standard deviation of the original series of Y, we get its value as 2. The coefficient of correlation between the original value of X and Y is – 0.919. We can also get the value by :

$$r = \sqrt{1 - \frac{s_y^2}{\sigma_y^2}} = \sqrt{1 - \frac{0.62}{4}} = \sqrt{0.848} = -0.919$$ **Ans.**

5.2.5 FORWARD AND BACKWARD ERROR

Suppose we want to compute y = f (x), where f is a real value function, but obtain approximate value y_{calc}

Forward error fwdcalc exact $E_{fwd} = y_{calc} - y_{exact}$

Backward error backwcalc $E_{backw} = x_{calc} - x$

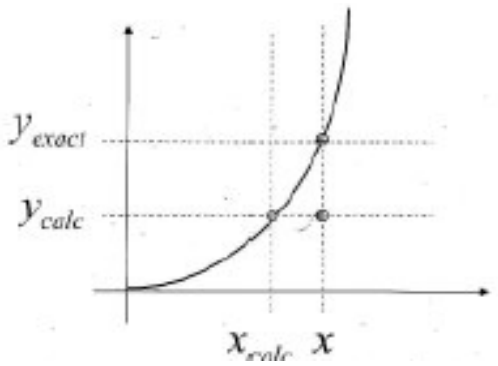


Fig.

$y_{exact} = f(x)$

$x_{calc} = f^{-1}(y_{calc})$

Suppose we want to compute y = f (x), where f is a real value function, but obtain approximate value ˆy

Forward error $\Delta y = \hat{y} - y$

Backword error $\Delta x = \hat{x} - x$

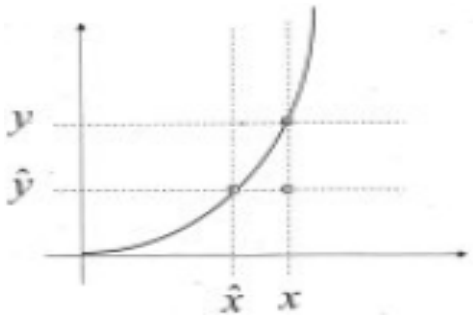


Fig.

y = f (x)

$\hat{x} = f^{-1}(\hat{y})$

5.3 APPROXIMATION IN STATISTICS

Before proceeding to collect data,it is necessary that the investigator decides about the degree of accuracy that he desires in his data collected. It should be kept in consideration that the absolute degree of accuracy is impossible because statistics is based on estimates and probabilities. Hence, it is very necessary to lay down the standard of accuracy desired. In this respect the most important thing is to maintain uniformity. For example, it should be clearly decided before handed that prices are to be quoted correct to nearest paisa, or fifty paisa, or rupee or ten rupees, or measurements are to be made correct to nearest centimetre, or metre or kilometre.

The most important point in this connection is that the degree of accuracy desired is a relative term depending upon he object under consideration and the nature of enquiry. It can be best explained with the help of an example.

Suppose, we are measuring distances between two cities, we may measure correct to the nearest kilometre, but in measuring the lengths of rooms we may measure correct to the nearest metre and in measuring the height of persons we will have to be more accurate and measure correct the nearest centimetre.

Sometimes the calculation of the statistical phenomena are not given in all the digits are calculated, rather they are rounded up to some nearest figures, *e.g.*, population of a city is counted as 14,58,639 in a population census, it may be rounded up as 14,58,600. This is known as approximation. In statistics whenever calculations are made it should be decided before hand that calculations are to be made nearest to the number or nearest to one or two decimal places. In this connection, the rule is that the actual calculation should be made upto one decimal place more than the place of which the results are required. The last decimal place will then be changed according to if the last calculated digit is less than 5, last decimal place digit will remain unaltered, if it is more than 5, the last decimal place digit will advance by 1 and if isis just 5, then it may remain the same or 1 may be added to it. Let use explain it with the help of an example :

Approximated figure correct to two decimal places

8.723 = 8.72

8.727 = 8.73

8.725 = 8.73 or 8.72

The only purpose in laying down these simple rules of approximation is to maintain uniformity in data collected by different investigators so that an uniform practice is adopted by all the investigators and there should not be any variation in their results.

Approximation as a process : Approximation is a process in which a statistical research finds exact appropriate value, well accepted by all the concerned authories. In approximation process, the person remains catious in each moment the cannot express carelessly taken value.

Approximate value : Approximate value is a vlaue which has constant and consistancy nature. To understand it some examples are considerable :

1. **Approximate value on the form of integer :**

   (1) As 1.1 to 1.4 is treated as 1, while 1..5 to 1.9 is treated as 2.

   (2) As 23.1 to 23.4 is treated as 23, while 23.5 to 23.9 is treated as 23.

2. **Approximate value in the form of one decimal point :**

   (1) As 1.51 to 1.54 is treated as 1.5 while 1.55 to 1.59 is treated as 1.6.

   (2) As 243.51 to 243.54 is treated as 243.5 while 243.55 to 243.59 is treated as 243.6.

3. **Approximate value in the form of two decimal point :**

   (1) As 1.421 to 2.424 is treated as 1.42 but 1.425 to 1.429 is treated as 1.43.

   (2) As 2.531 to 2.534 is treated as 2.53 but 2.535 to 2.539 is treated as 2.54.

4. **Approximate value in the form of three decimal point :**

   (1) As 4.5511 to 4.5514 is treated as 4.551 but 4.555 to 4.559 is treated as 4.552.

   (2) 25.441 to 25.444 is treated as 25.44 while 25.445 to 25.449 is treated as 25.45.

5. **Approximate value in the form of four decimal point :**

   (1) As 4.55551 to 4.55554 is treated as 4.5555 and 4.55555 to 4.55559 is 4.5556 and so on we can dicuss ahead.

5.4 TRUNCATION ERROR

Difference between true result and result produced by given algorithm using exact arithmetic.

Due to approximations such as truncating infinite series or terminating iterative sequence before convergence

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \sum_{n=4}^{\infty} \frac{x^n}{n!}$$

$$e^x \approx 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!}$$

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \sum_{n=4}^{\infty} \frac{x^n}{n!}$$

$$e^x \approx 1 \qquad e^{0.5} \approx 1$$

$$e^x \approx 1 + \frac{x}{1!} \qquad e^{0.5} \approx 1.5$$

$$\varepsilon = \frac{\text{current approximation} - \text{previous approximation}}{\text{current approximation}}$$

$$\varepsilon = \frac{1.5 - 1}{1.5} 100\% = 33.3\%$$

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \sum_{n=4}^{\infty} \frac{x^n}{n!}$$

$$e^x \approx 1 \qquad e^{0.5} \approx 1 \qquad \varepsilon = 33.3\%$$

$$e^x \approx 1 + \frac{x}{1!} \qquad e^{0.5} \approx 1.5$$

$$e^x \approx 1 + \frac{x}{1!} + \frac{x^2}{2!} \; e^{0.5} \approx 1.625$$

$$\varepsilon = \frac{1.625 - 1}{1.625} 100\% = 7.69\%$$

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \sum_{n=4}^{\infty} \frac{x^n}{n!}$$

$$e^x \approx 1 \qquad e^{0.5} \approx 1$$

$$e^x \approx 1 + \frac{x}{1!} \qquad e^{0.5} \approx 1.5 \qquad \varepsilon = 33.3\%$$

$$e^x \approx 1 + \frac{x}{1!} + \frac{x^2}{2!} \; e^{0.5} \approx 1.625 \qquad \varepsilon = 7.69\%$$

$$e^x \approx 1+\frac{x}{1!}+\frac{x^2}{2!}+\frac{x^3}{3!}e^{0.5} \approx 1.645833333 \qquad \varepsilon = 1.27\%$$

Conclusion

We have presented the concept of error, measurement of error and approximation in statistics but actually these theories require additional research.

UNIT-6 COLLECTION OF DATA-PRIMARY SOURCES

Objectives

- After reading this unit, you should be able to know about:
- Primary data
- Methods of primary data collection
- Problems in primary data collection and its merits & demerits

Structure

6.1 Introduction.

6.2 Factors Affecting choice of Method.

6.3 Methods of Collecting Primary data.

6.4 Direct Personalinvestigation.

6.5 Indirect oral investigation.

6.6 By Local Reports.

6.7 By Schedules & Questionnaires.

6.8 Drafting a Questionnaire.

6.9 Pre testing A Questionnaire.

6.10 Specimen Questionnaire.

6.11 Representation data.

6.12 Summary.

6.13 Further Study.

6.1 INTRODUCTION

Statistical data as we have already seen can be either primary or secondary. Primary data are those which are collected for the first time and are, thus original in character. Primary data are in shape of raw material to which statistical methods are applied for the purpose of

analysis and interpretation. After statistical treatment the primary data looses their original shape and became secondary data. It will be found that distimction between the Primary & Secondary data in many cases in one of degree only.

6.2 FACTORS EFFECTING CHOICE OF METHOD

It is obvious that the methods of collecting primary data and secondary data would not be exactly identical because in one case the data have to be original collected while in other the work is of the nature of compilation. There are various methods of collecting primary data and choice of method depends on number of factors. Nature, objects scope of the inquiry are the most important thing on which the selection of method depends on a number of factors. The method selected should be such that it suites the type of enquiry is being conducted.

Availability of finance is another factor which influences the selection of the method of the collection of data. When financial resources at the disposal of the investigater are scanty he shall have to leave aside expensive methods even though they are better than others which are comparatively cheap.

Availability of time has also to be taken into account. Some methods involve a long duration of enquiry while the others, the enquiry can be conducted in comparatively shorter duration. The time at the disposal of the investigator, this affects the selection of the techniques by which data are to be collected.

6.3 METHODS OF COLLECTING PRIMARY DATA

The following methods of collecting of primary data are in commone use :

- Direct personal investigation
- Indirect oral method
- By local reports
- By schedules & questionnaires

We shall briefly discuss each of them in turn.

6.3.1 DIRECT PERSONAL INVESTIGATION

In direct personal investigation, as the name suggests, the investigator has to collect the information personally from the sources concerned. He has to be on the spot for conducting the enquiry and has to meet people from whom data have to be collected. It is necessary that in

such cases, the investigator has a keen sense of observation and he is very polite and courteous. He should further acquaint himself with local conditions, customs and traditions so that he is in a position to identify himself fully with the persons from whom the information is sought. The investigator has to be very tactful and cautious in such cases. He should put easy and simple questions which are capable of being answered precisely and in a language which is not vague and fully understandable by the source. In some cases, it may not be possible or worthwhile to contact directly the persons concerned and in such cases, the investigator has to cross-examine other persons who are closely in touch with the sources of data. The information elicited in such a manner should be carefully used and the investigator should make sure that the persons from whom data are being collected actually know the facts fully and can deliver him the goods.

The method of direct personal investigation is suitable only for intensive investigations. It involves enormous cost and usually required a long time. It is naturally not suitable for extensive enquiries where the scope of investigation is wide. Further, in this method, the bias or prejudice of the investigator can do a lot of damage as he is in sole charge of the collection of data. This method, however, gives very satisfactory results if the scope of the enquiry is narrow and if the investigator is fully dependable and is completely unbiased.

6.3.2 INDIRECT ORAL INVESTIGATION

What the above mentioned method cannot be used either on account of the reluctance of persons to part with information when approached directly, or on account of the extensive scope of the enquiry or on account of some other reason, an indirect oral examination can be conducted. In this method, data are not collected directly from the persons concerned but through indirect sources. Persons who are supposed to have knowledge about the problem under investigation are interrogated and the desired information is collected.

Usually, in such enquiries, a small list of questions relating to the investigation is prepared and these questions are put to different persons (known as witnesses) and their answers are recorded. Most of the commissions and committees appointed by the Government to collect statistical data or to carry on such investigations in which factual data have to be compiled, make use of this method. They request different types of people to come and give evidences and on the basis of these records, facts about different problems are ascertained.

In such enquiries the evidence of one person should not be relied uponand the views of a number of persons should be ascertained to find out the real position.

In this method, the accuracy of data collected would largely depend on the type of persons whose evidences are being recorded. It is,

therefore, necessary to be very cautious in the selection of these persons. Invariably it should be seen that the person who is being questioned :

(a) knows full facts of the problem under investigation;

(b) is not prejudiced;

(c) is capable of expressing himself correctly and can give a true account; and

(d) is not motivated to give colour to the facts.

Proper allowance made for the inherent optimism or pessimism of the informants i. e, their inherent psychology. For example if there are two drunkards (one optimist and the other pessimist), each of whom was left with half a glass of wine illustrates the point very clearly. The optimist said, “What do I care for the world, I have yet half the glass with me” and the pessimist remarked, “What can I do in this world, I have only half the glass with me.”

Both of them were stating facts correctly and yet the two statements give entirely different impressions.

6.3.3 LOCAL REPORTS

The last method of collection of primary data is through local reports. In this method, data are not formally collected by enumerators but the local correspondents or agents in their own fashion and to their own likings.

Obviously, such data cannot be very reliable and, as such, this method is used in those cases where the purpose of investigation can be served with rough estimates only and where a high degree of precision is not necessary. This method has the advantage of being least expensive and it also saves the botheration usually associated with statistical investigation of other types.

6.3.4 SCHEDULES AND QUESTIONNAIRES

This is an important method for the collection of data which is followed usually by private individuals, research workers, non-official institutions and, sometimes, the Government also. In this method, a list of questions relating the problem under investigation is prepared and printed and information is collected from various sources in any of the following ways :

(a) By sending the questionnaire to the persons concerned and requesting them to answer the questions and return the questionnaire.

The main advantage of this method is that it is least expensive, and with it information can be collected from a wide area in a

comparatively short period of time. If the investigation is properly conducted, the method can easily ensure a reasonale standard of accuracy. Success in this method depends on the co-operation and the attitude of the informants. If the informants are indifferent in their attitude and do not take interest in answering the questions, the method may give any fruit for results.

However, this method cannot be used if the informants are illiterate. If they are literate but adopt an indifferent attitude, then also the method should be used with utmost caution as, in such cases, likelihood of error is very great.

(b) By sending the questionnaires through enumerators to help the informants in filling the answers.

In this method, the enumerators go the informants along with the questionnaires and help them in recording their answers. The enumerators explain the aims and objects of the investigation to the informants and also emphasise the nesessity and usefulness of correct answers. They also remove the difficulties which any informant may feel in understanding the implications of a particular question or the definition or concept of difficult terms. This method is very useful in extensive enquiries and with it, fairly dependable results can be expedted. It is, however, very expensive and usuallysuchenquiries can be conducted only by the Government. Population census all over the world is conducted bythis method. In such enquiries, it is necessary that not only the questions are simple and few in number but theenumeratorsarealso courteous and polite and have proper training.

The selection of enumerators is very important and should be carefully done. The enumerators should be explained the nature,scopeandsubject of the investigation thoroughly andthey should properly understand the implications of the differentquestions put and the definitions of the various terms used.

The enumerators should have intelligence and capacity of cross-examination for the purpose of finding out the truth and they should be persons who are hard-working and should have patience and perseverance.

6.3.5 DRAFTING A QUESTIONAIRE

The success of the questionnaire method depends to a large extent on proper drafting of the questionnaire. It is a highly specialised job and needs lot of skill and experience. Though there is no hard and fast rule for preparing a questionnaire, yet some broad principles followed in this regard are given below :

1. (i) **A very polite covering letter should be sent to the respondents with the questionnaire,** emphasising the need and usefulness of the information that is being collected and

requesting them to give their co-operation by sending correct replies.A self-addressed stamped envelope should also be enclosed for the respondent's convenience in sending the reply.

(ii) An assurance must be given to the respondents that if they wish the information given by them would be treated as confidential and not disclosed to anybody, it will be taken care of.

(iii) If possible, some inducement could also be given to the respondent in the form of small gofts so that there is greater chance of gettig a response.

(iv) The respondents could also be given a promise that if they so desire a copy of the results of the survey would be sent to them.

2. Number of questions should not be very large : Unnecessary details should be avoided and only relevant questions should be framed. There is no hard and fast rule with regard to the number of questions as it would depend on the nature of the enquiry.

3. The questions should be short and clear : There should be no ambiguity in the questions. The questions should not be confuisng and should be capable of a straight answer. As far as possible, the questions should be capable of objective answers. A set of possible answeresamy be given against each question and the respondent may be asked to tick the appropriate answer. For example if we have to ask a respondent about the percentage of his income spent by him on his children's education, we can frame the question as follows:

What percentage of your total income is spent on your children's education?

Tick the appropriate square.

(a) Less than 5%

(b) 5% to 10%

(c) 10% to 15%

(d) 15%n to 20%

(e) more than 20%

Note : If you spend exactly 10% of your income, please tick 10% to 15% and, similarly, if it is exactly 15%, then tick 15 to 20 percent.

Another example could be when you wish to find out the mode of cooking in a household. The question should be :

Which of the following modes of cooking do you use?

(a) Wood

(b) Coal

(c) Kerosine

(d) Gas

(e) Electricity

Please tick the relevant square, If you use more than one method, please tick the one which is mostly used.

4. **If the opinion of a respondent is sought on a particular point, if possible** : the question should be so framed that he can ansere it in yes or no. For example, if a respondent's opinion is sought about some government policies, the answer may not be clear-cut yes or no. Such questions should, as far as possible, be avoided.

5. **Questions of a personal nature should not be asked :** because the respondent would either not give information or give incorrect information. Questions relating to income tax returns or perks given to certain employees may ot be liked by the respondents. In fact, no question which elicits soe sort of confidential information should be included in the questionnair.

6. **No such question which hurts the sentiments of the respondent should be asked:** For example questions on indebtedness, private life, litigation, etc., should be avoided.

7. **There should be such questions which are corroboratory in nature** : Such questions are meant for cross-checs to the anseres given by the respondent.If the answers given are correct, there would be no contradiction in answers.

8. **Questions whose answers require calculation should not be generally asked** : For example, if the annual income of a salaried class person is asked or if monthly profits of a concern are called for, or if the percentage of parttime employees to total employees is asked, a lot of calculation work is involved. The respondent may not wish to undrtake this arduous task.

9. **The questionnair should look attractive and impressive**: It always helps if it is so. There should be sufficient space for answering the questions and the quality of paper and printing whould also be good.

6.3.6 PRE-TESTING A QUESTIONAIRE

Before the questionnaire is finalised, it is always worthwhile to get it pretested. A small sample from the relevant universe may be picked up and the quetionnnaire tested on it. This has many advantages. We are in a position to know the type of response that we may ultimately get from the

respondents. Some questions may be found to be inappropriate and many need a chage. Some concepts may need clarification as the sample respondents might have displayid confusion while answering questions.

By pre-testing a questionnaire, it is also possible to find out the ways in which greater co-operation of the respondents may be achieved.

After the questionnaire has been pretested, it should be modified in the light of the experience gathered and then used for the purpose of collecting information.

6.3.7 SPECIMEN QUESTIONAIRE

Given below is specimen questionnaire issued by the Hodhpur Management Programme to collect information about the influence of textile advertisements on consumers. The questionnaire is accompanied by a polite letter to the respondents as given below :

LUCKNOW MANANGEMENT PROGRAMME

Faculty of Commerce, University of Lucknow, Lucknow

Mr/Mrs/Miss.....

Date.....

.....

.....

Dear Sir, Madam,

We are collecting information about the effect of textile advertisements on the consumers of textile goods. We hope your are conscious of the fact that millions of rupees are spent by textile companies on advertisemnt of their products. This money forms a part of their cost production and is ultimately borne by you. We wish to find out whether it is worthwhile to spend so much money on advertisements whcih the consumer has to pay.

We would be extremely obliged to you if you fill up the enclosed questionnaire and let us have your valued opinion on the various issues involved in the problem. The enquiry will help both the consumbers as well as the producers of the textile goods.

Kindly fill up the enclosed quesionnaire and return the same to us at your earliest. Aself addressed stamped envelope is enclosed herewith for your convenience.

The information supplied by you will be treated as confidential, and copy of the findings wil be sent to you, if you so desire.

Thanking you, Yours truely

.....

Enc. (i) Questionnaire

(ii) Envelope

LUCKNOW MANAGEMENT PROGRAMME

FACULTY OF COMMERCE, UNIVERSITY OF LUCKNOW

MARKET SURVEY

TEXTILE ADs AND CONSUMERS

A. Profile of the Respondent:

1. Name :
2. Age :
3. Address :
4. Education :
5. Annual Income :
6. No. of members in the family : Male Female
7. Occupation :

B. Patterns of purchase decisions :

1. Do you have any predetermined idea about the clothes to be purchased when you enter the shop, in terms of :

   Mostly Sometimes Never

   1. Price range
   2. Quality
   3. Textile of a particular company
   4. Textile expossed to you through advertisement.

   If yes, list the factors that helped you most in forming this idea:

   (i) Suggestions offered by friends, relatives, etc.

   (ii) Textile advertisement seen by you.

   (iii) Window display in the shop.

   (iv) Seeing someone else wearing it.

2. When you enter the shop your enquiry begings with :

   Mostly Sometimes Never

   1. Textile fo a particular company
   2. Textile in certain price range
   3. Textile which has bneem seen in Ads.
   4. Textile based on quality alone.

C. Factory influencing purchae decisions:

1. General influence :

What factors influence you most while taking a decision about the textiles to be purchased :

(Rank in order of pereference)

(a) Price

(b) Colour and design

(c) Quality and textue

(d) Impact of Advertisement Company image

(e) Sellers advice

(f) Friends and relatives advice

(g) Displays in the shop bearing Company's name

2. Influence of the Advertisements:

(i) Which type of textiles do you buy on the basis of Ads only :

(Tick)

Types of Textiles	**Textile for**
(a) Cotton textile.	(a) Domestic purposes
(b) Synthetic fibre textiles	(b) Childre's wear
(c) Woollen clothes	(c) For wife
(d) Any textile offering incentives	(d) For self
(e) Giving as gifts.	

(ii) Do you think the textile ads. have an impact on you :

Mostly Sometimes Never

1. Only while looking at it
2. Sometime after having seen it
3. When future purchases are planned
4. It is retained in your memory when you enter the shop.
5. While making the purchase in the shop.
6. Only when incentives are offered through the advertisements.

(iii) Do you think textile ads. seen by you are really helpful in our decision-making:

Types of advertisement	Verey helpful	Helfpful not at all
1. Printed ads. (Newspaper etc.).		
2. Radio jingles		
3. Visual ads. (Cinema, T.V., etc.)		

If yes, how?

(iv) If you think the ads. are not helpful, the reasons could be:

(a) ads. highlight only the Company name.

(b) Ads. do not provide details in terms of:

1. Price range
2. Quality
3. Colours and designs available.
4. Product range available.
5. Places where advertised cloth is available.
6. Uses to which advertised cloth could be put to
7. Advertised cloth is generally not available
8. Clothes when seen actually do not apperar so attractive.
9. All companies do not advertise.
10. The whole product range of the comany is not advertised.

D. Source of information about the Company's types of textiles, etc.

1. How do you learn about the companies' name and their products:

	Mostly	Sometimes	Never
(i) By visiting the shop			
(ii) Through associates who patronise a particular company			
(iii) Through advertisements			
(iv) Old established companies, the image of which is retained.			

2. How do you learn about new textiles that come to the market:

	Mostly	Sometimes	Never
(i) Through friends			
(ii) Through advertisements			
(iii) Window displays in the shop			
(iv) Through sellers.			

3. Among the textiles ads. to which you are exposed, mention those which you remember most (the co's name)

   (i)

   (ii)

   (iii)

   (iv)

   (v)

   (vi)

4. You remember those ads. because (Tick)

   (a) they appear very often

   (b) the slogan is very catchy

   (c) ads. are of the most established companies.

   (d) ads. are very attractive

   (e) they furnish the desired details.

   (f) you try to base your decisions on these ads. since they inform you of the lates.

E. Your suggestions :

1. What factor do you think should be incorporated in the ads. so as to make them really helpful in decision-making (rank the suggestion in order to preference)

   (i) They must furnish details in terms of:

      (a) price range

      (b) product range

      (c) quality

      (d) colur range

      (e) place of availability

(f) uses of the textile

(ii) should appear often

(iii) advertisement clothes should be made available in local markets

(iv) should be made more attractive

(v) should be technicolour

(vi) should be only in black and white

(vii) should be only good taste

(viii) any other :

(a)

(b)

(c)

(d)

Thank you

6.4 REPRESENTATIVE DATA

As has been pointed out earlier, a statistical investigation can be either of census type or of sample type. In acensus enquiry all the untis associated with a particular problem are taken in to account where as in sample enquire only a few selected units are stdied and on the baiss of such studies attempts are made to dreaw generalisations whih may be applicable to the whole data. If, for example, we have to find out the average monthly expenditure of the 2,000 students residing in the hostels of the Allahabad University and if we hold a census investigation, we shall have to study the monthly expenditure of each one of these 2,000 students. If, however, we hold sample ivestigation, we shall select, say, 200 students out of these 2,000 and then study their expenditure.On the basis of the study of these 200 units (technically called a “sample”), we can draw conclusions which will hold good about the expenditure of all the 2,000 students (tecnically called a “universe” or “population”). The sample is considered to be a representative of universe and if the sample has been properly selected and if its size is proper, whatever holds good for the sample should also hold good for the universe. If the scope of the enquiry is very wide, a census inverstigation would not only be very expensive but highly cumbersome also. Moreover, it will take a very long time and require a large number of enumerators. In such cases, a sample ivestigation is very suitable. A sample usually gives representative dada and the gerneralisations made on the basis of such data usually hold good for the universe.

The most important point, however, is the Selection of the Sample. A sample study would give dependable conclusion only if the sample is a true representative of the universe. Broadly speaking, there are tow methods by which samples can be selected and they are:

1. Deliberate or purposive sampling.
2. Random or chance sampling.

Deliberate Selection or Purposive Sampling :

In deliberate selection or purposive sampling, the iverstigator himself chooses from the universe a few such units which, according to his estimates, are best representatives of the population. His selection is deliberate and is based on his own ideas about the represntatives of the population. His selection is deliberate and is based on his own ideas about the representativeness of the sampled units. These selected units are intensively studied and certain conclusions are arrived at. It is supposed that these conclusions would hold good for the whole population.

This technique of selection has many drawbacks. The first and the foremost fo them is that the bias or prejudice of the investigator has enough scope to work and influence the selection.If the investigator is biased, it is but natural that he could select such a saple which would give conclusions which suit his requirements and views. It, for example, an investigator wants to show that the expenses of students residing in the hostels of th university are very high, he can select such a sample which consists of those students only who are aristocratic and who spend much more than others.Another defect of purpose sampling is that it is not possible to have any idea about the degree of accuracy achieved in any statistical investigation conducted by this method. If the scope of enquiry is very wide, the selection of the sample by this method can never be recommended. Hover, if the investigator is unbiased and has the capacity of keen observation and sound judgement, evern purposive selection can give fiarly dependable results.

Chance Seleciton or Random Sampling :

In random sampling, the selection of the units is done in such a manner that the chance of selection of each unit of the universe is the same.In other words, the selection of the units depends entirely on chance and one does not know beforehand which units will actually consistitue the sample.It is for this reason that this method is ablso known as the methods of chance selection, It is, in fact, a lottery method of selection. How can such a selection be made is a question not easy to answer. Methods, which on first thought, appear to be perfectly random may actually prove to be otherwise. If we have to select a samploe of 200 students out of 2000 hostellers, we can write their names on small chits of papers and after folding them and mixing them together can draw blindfold 200 chits.This is the lottery method. it appears to be random but, in actual practice, it may not be so. It is quite possible that some chits were

folded and pressed less than others and so their size was slightly bigger thant the size of other chits and if it was so, the chance of the selection of bigger chits cannot be said to be the same as that of the smaller ones. Thus, we see that it is not easy to have a purely random sample. However, various methods are in vogue and amog them the technique of Random Nibers is most popular. In random sampling, attempts are made to eliminate human bias in all forms and that is why selections are usually made with the help of machines. Each unit of the universe is assigned a number and then certain numbers are mechnically selected to constitute a sample.

Chance selection or random samplig has many advantages over purposive selection. The most signinficant merit of this system is that by theory of probability it is possible to have an idea about the errors of estimation, and we can always find out whether the results are significant or not.It is possible to assign limits within which the true value of a measure of universe must invariably lie.Another point in favour of this method is that the selection is not affected by the prejuidce or bias of the investigator. As we have noted above, the seletion in most cases under this system is made by mechanical devices and naturally human bias has hardly any scope here. We should prefer to select the sample on random basis becuase the results are likely to be put for than purposive sampling.

Accuracy and size of sample : It should further be kept in mind that the size of the sample has a relation to the degree of accuracy that it is expected to achieve. Ordinarily, the bigger the size of a sample the greater would be the accuracy, but a avery big sized sample is likely to become unmanageable and is very offten unnecessary. No hard and fast rule can be laid down with regard to size of a sample.An ideal size would depend on the type of the sereies and the size of the universe.If the series is comparatively more variable (have more dispersion), the sample should be big to cover up all types of variatons.Again, ordinarily, the bigger the universe the greater should be the size of the sample. The accuracy in a random sample has more or less a fixed relation to its size.

$$SE = \frac{\sigma}{\sqrt{n}}$$

Random sampling and the theory of probability : The technique of random samplig is based on the Theory of Probability is a mathematical concept and indicates the likelihood or the chance of the happening or not happeining of a particular event. If, for example a coin is tossed it can fall is tow ways— either with head up or tail up. Eeach of these wyas is euqlly likely and so the probablity of the coin falling head or tail up is equal. It is 1/2. If, however, a dice is thrown, there are six possible ways in which it can fall.The probability of its failling with No.6 up is 1/6 because the chance of its faliing with any of the six numbers upward is equal. The chance of the dice not falling with 6 upward would be 5/6 as there are five ways in which it can fall without No.6 being upward. Thus, if an event can happen in a wyas and fail to happen in b ways in and if each of them is

equally likely, the chance of its happening would be $\frac{a}{a+b}$ and of its not happening $\frac{b}{a+b}$.

If from a pack of 52 card one card is drawn at random the chance of its being any king is clearly 4/52 and the chance of its being any card of spade is 13/ 52.This clearly indicates that if the chances of selection of all the units in a universe are equal, and if from it, selections are made at random, then the possibility is, that in the sample so selected the various type of units would be in the same proportion in which they are in the universe. On this basis it is said that random sampling gives a representative sample which contains the characteristic of the population. So if the accuracy is required to be doubled i.e., SE to be halved, the sample size should be four times. The bigger the size of the sample the greater is the chance of accuracy.

Law of statistical regularity : Thus, accoridng to the rule of the theory of probablity, if from the universe a moderately large sized sample is chosen at **random,** it is almost certain that, on an average, the sample so chosen will have the same characteristics as the universe.It is on this basis that games of chance are played successfully by a large number of persons and the insurance comapanies are able to insure people against various types of calamities. In statistics, this law is known as the "Law of Statistical Regularity". It is a corollary to the main theory of probability. The theory of probability tells us of the mathematical expectations of the success or failure of an event and, on this basis, the law of statistical regularity tells us that random selection from the universe is very likely to give a representative sample.

Law of inertia of large numbers : We have mentioned above that there is a relationship between the size of sample and its accuracy. The larger the sample, the greater would be the accuracy. The reason for this lies in the fact that in large numbers, the chances of **compensatory** action are greater. If in the first ten tosses of a coin there are seven heads and threee tails, it is quite likely that in the next ten tosses, the situation might be reversed and there may be seven tails and three heads. The larger the number of such experiments the greater are the chance of one irregularity compensating the other. It is said on this basis that large numbers have got an inertia or that they are more constant. The production of wheat in the district of Allahabad might show great variations year after year but the production figures of the state of U.P. would not vary much, because if in some districts the crop is above normal it is very likely that in others it might be below normal. Similarlity, the production figures of wheat for the whole of India would show still less variations and the figures of world production would show hardly any significant change. This phenomenon is characterised as the "Law of Inettia of Large Numbers" which states that large numbers are relativerly more constnt and stable than small ones.

It is on the basis of this law that we say that larger the size f the sample, the greater would be its accuracy.

6.5 SUMMARY

It should not be concluded from the above discussion that the law of inertia of large numbers does not allow any change in figures with the passage of time. Al that it means is that large numbers are more constant and stable than small ones. There are no violent fluctuations in large numbers.After all, the figures of world production of wheat do change from time to time but these changes are not violent and sudden. Thet are slo and gradual. Long-period trend is indicated by large numbers; they simply ignore the short-period regular and irregular fluctuations.

6.6 FURTHER STUDY

1. Agrawal, B.N., Elhance, D.N., ELhance, Veena : Fundamental of Statistics
2. AnandVinod, Chand Mahesh : Economic Theory - A Mathematical Approach
3. Monga, G.S. : Elementary Statistics
4. Gupta, S.B. : Principal of Statistics

UNIT-07 COLLECTION OF DATA-SECONDARY DATA

Structure

7.1 Introduction

7.2 Sources of Secondary Data

7.3 Distinction Between Primary and Secondary Data

7.4 Methods for Evaluating Secondary Data Sources

7.5 Advantages and Disadvantages of Secondary Data Analysis

7.6 Locating Appropriate Secondary Data

7.7 Questions to ask about any Secondary Data Set

7.8 What Evidence is there for A Lack of Quantitative Skills?

7.9 Summary

7.10 Further Study

7.1 INTRODUCTION

This topic introduces the field of secondary data analysis. It begins by considering what it is that we mean by secondary data analysis, before describing the type of data that might lend itself to secondary analysis and the ways in which the approach has developed as a research tool in social and educational research.

7.1.1 NATURE OF SECONDARY DATA

Secondary data is data which has been collected by individuals or agencies for purposes other than those of our particular research study. For example, if a government department has conducted a survey of, say, family food expenditures, then a food manufacturer might use this data in the organisation's evaluations of the total potential market for a new product. Similarly, statistics prepared by a ministry on agricultural production will prove useful to a whole host of people and organisations, including those marketing agricultural supplies.

No marketing research study should be undertaken without a prior search of secondary sources (also termed desk research). There are several grounds for making such a bold statement.

- Secondary data may be available which is entirely appropriate and wholly adequate to draw conclusions and answer the question or solve the problem. Sometimes primary data collection simply is not necessary.

- It is far cheaper to collect secondary data than to obtain primary data. For the same level of research budget a thorough examination of secondary sources can yield a great deal more information than can be had through a primary data collection exercise.

- The time involved in searching secondary sources is much less than that needed to complete primary data collection.

- Secondary sources of information can yield more accurate data than that obtained through primary research. This is not always true but where a government or international agency has undertaken a large scale survey, or even a census, this is likely to yield far more accurate results than custom designed and executed surveys when these are based on relatively small sample sizes.

- It should not be forgotten that secondary data can play a substantial role in the exploratory phase of the research when the task at hand is to define the research problem and to generate hypotheses. The assembly and analysis of secondary data almost invariably improves the researcher's understanding of the marketing problem, the various lines of inquiry that could or should be followed and the alternative courses of action which might be pursued.

- Secondary sources help define the population. Secondary data can be extremely useful both in defining the population and in structuring the sample to be taken. For instance, government statistics on a country's agriculture will help decide how to stratify a sample and, once sample estimates have been calculated, these can be used to project those estimates to the population.

Secondary data can embrace a whole spectrum of empirical forms; they can include data generated through systematic reviews, through documentary analysis as well as the results from large-scale datasets such as the National Census or NFHS. Secondary data can be numeric or non-numeric. Nonnumeric, or qualitative secondary data, can include data retrieved second hand from interviews, ethnographic accounts, documents, photographs or conversations. In the UK, an excellent source of archived qualitative data with huge potential for secondary analysis is available through the Economic and Social Data Services (ESDS) Qualidata facility

based at the University of Essex. Data available through Qualidata includes in-depth and semi-structured interviews, field notes and observations, as well as personal documents. The service provides support and training, as well as access to contemporary and classic studies of British society.

7.2 SOURCES OF SECONDARY DATA

The secondary sources can be classified into two categories via. Published and unpublished sources. Hence, the secondary data can be obtained through

1. Internal Sources - These are within the organization
2. External Sources - These are outside the organization

7.2.1 INTERNAL SOURCES OF DATA

If available, internal secondary data may be obtained with less time, effort and money than the external secondary data. In addition, they may also be more pertinent to the situation at hand since they are from within the organization. The internal sources include

1. **Accounting resources-** This gives so much information which can be used by the marketing researcher. They give information about internal factors.
2. **Sales Force Report-** It gives information about the sale of a product. The information provided is of outside the organization.
3. **Internal Experts-** These are people who are heading the various departments. They can give an idea of how a particular thing is working
4. **Miscellaneous Reports-** These are what information you are getting from operational reports.

If the data available within the organization are unsuitable or inadequate, the marketer should extend the search to external secondary data sources.

7.2.2 EXTERNAL SOURCES OF DATA

External Sources are sources which are outside the company in a larger environment. Collection of external data is more difficult because the data have much greater variety and the sources are much more numerous.

External data can be divided into following classes.

a. **Government Publications** - Government sources provide an extremely rich pool of data for the researchers. In addition, many of these data are available free of cost on internet websites. There are number of government agencies generating data. These are :

i. **Registrar General of India -** It is an office which generate demographic data. It includes details of gender, age, occupation etc.

ii. **Central Statistical Organization -** This organization publishes the national accounts statistics. It contains estimates of national income for several years, growth rate, and rate of major economic activities. Annual survey of Industries is also published by the CSO. It gives information about the total number of workers employed, production units, material used and value added by the manufacturer.

iii. **Director General of Commercial Intelligence -** This office operates from Kolkata. It gives information about foreign trade i.e. import and export. These figures are provided region-wise and country-wise.

iv. **Ministry of Commerce and Industries -** This ministry through the office of economic advisor provides information on wholesale price index. These indices may be related to a number of sectors like food, fuel, power, food grains etc. It also generates All India Consumer Price Index numbers for industrial workers, urban, non manual employees and cultural labourers.

v. **Planning Commission -** It provides the basic statistics of Indian Economy.

vi. **Reserve Bank of India -** This provides information on Banking Savings and investment. RBI also prepares currency and finance reports.

vii. **Labour Bureau -** It provides information on skilled, unskilled, white collared jobs etc.

viii. **National Sample Survey -** This is done by the Ministry of Planning and it provides social, economic, demographic, industrial and agricultural statistics.

ix. **Department of Economic Affairs -** It conducts economic survey and it also generates information on income, consumption, expenditure, investment, savings and foreign trade.

x. **State Statistical Abstract -** This gives information on various types of activities related to the state like - commercial activities, education, occupation etc.

b. Non Government Publications - These includes publications of various industrial and trade associations, such as

i. The Indian Cotton Mill Association

ii. Various chambers of commerce

iii. The Bombay Stock Exchange (it publishes a directory containing financial accounts, key profitability and other relevant matter)

iv. Various Associations of Press Media.

v. Export Promotion Council.

vi. Confederation of Indian Industries (CII)

vii. Small Industries Development Board of India

viii. Different Mills like - Woolen mills, Textile mills etc. The only disadvantage of the above sources is that the data may be biased. They are likely to colour their negative points.

c. Syndicate Services - These services are provided by certain organizations which collect and tabulate the marketing information on a regular basis for a number of clients who are the subscribers to these services. So the services are designed in such a way that the information suits the subscriber. These services are useful in television viewing, movement of consumer goods etc. These syndicate services provide information data from both household as well as institution.

In collecting data from household they use three approaches

i. Survey - They conduct surveys regarding - lifestyle, sociographic, general topics.

ii. Mail Diary Panel - It may be related to 2 fields - Purchase and Media.

iii. Electronic Scanner Services - These are used to generate data on volume. They collect data for Institutions from

iv. Whole sellers

v. Retailers, and

vi. Industrial Firms

Various syndicate services are Operations Research Group (ORG) and The Indian Marketing Research Bureau (IMRB).

Importance of Syndicate Services :

Syndicate services are becoming popular since the constraints of decision making are changing and we need more of specific decisionmaking in the light of changing environment. Also Syndicate services are able to provide information to the industries at a low unit cost.

Disadvantages of Syndicate Services :

The information provided is not exclusive. A number of research agencies provide customized services which suits the requirement of each individual organization.

d. International Organization- These includes

i. The International Labour Organization (ILO)- It publishes data on the total and active population, employment, unemployment, wages and consumer prices

ii. The Organization for Economic Co-operation and development (OECD) - It publishes data on foreign trade, industry, food, transport, and science and technology.

iii. The International Monetary Fund (IMA)- It publishes reports on national and international foreign exchange regulations.

7.3 DISTINCTION BETWEEN PRIMARY AND SECONDARY DATA

The distinction between *primary* and *secondary* data depends on the relationship between the person or research team who collected a data set and the person who is analyzing it. This is an important concept because the same data set could be primary data in one analysis and secondary data in another. If the data set in question was collected by the researcher (or a team of which the researcher is a part) for the specific purpose or analysis under consideration, it is *primary data*. If it was collected by someone else for some other purpose, it is *secondary data*. Of course, there will always be cases in which this distinction is less clear, but it may be useful to conceptualize primary and secondary data by considering two extreme cases. In the first, which is an example of *primary data*, a research team conceives of and develops a research project, collects data designed to address specific questions posed by the project, and performs and publishes their own analyses of the data they have collected.

As an example of the same data set serving as both primary and secondary data, consider the increasingly common practice of one researcher performing an analysis of data collected by a research team with whom he or she has no connection. This type of analysis is facilitated by the ease of sharing data stored electronically and the concomitant creation of electronic data archives that allow access to secondary users. Such analyses may serve a variety of purposes, such as addressing questions not considered in the original analysis or examining how a different analytic approach might change the conclusions reached from the first analysis.

In either case, the same data set serves as *primary data* for the original research team and *secondary data* for the researcher performing

the later analysis. There is another type of secondary data, again not mutually exclusive with the first, meaning statistical information about some geographic region or other entity. This type of information is often useful to researchers: when you place your research project in context by describing the racial makeup or median house value in the metropolitan area where you conduct your research, the data used to compute those statistics were probably secondary data. Most of the data sets discussed in this volume contain either data collected through surveys or censuses, such as the National Family Health Survey and the Indian Census, or administrative records such as the medical claims records submitted to the Medicare system. There are other types of secondary data, including diaries, video-recordings, and transcripts of interviews and focus groups.

7.4 METHODS FOR EVALUATING SECONDARY DATA SOURCES

The sources that are used to acquire the data for research can be primary or secondary. The sources are known as secondary data sources when the information occurring therein has already appeared somewhere. Examples of secondary data sources are – online archives, museums, libraries, databases, repositories etc. The secondary data sources can further be classified into – published and unpublished sources.

Researchers need to choose from two or more than two sources of data. Before choosing appropriate data sources, one should consider some methods that are available for evaluating secondary data sources. Researchers should make sure that they verify the accuracy of the information. For this, they may need to get the transcripts of court hearing in case they are using the information from legal proceeding. It is not easy to evaluate the secondary data in the research. Hence, one should put all the necessary efforts to prove the credibility of the sources that are used in the research.

The researcher should use the secondary data sources, if and only if he/she finds no primary source of data. The researcher should consider the scope, purpose, authority, source of financing and validity for the purpose of evaluating the secondary source. All these factors help the researcher in evaluating the value of the secondary source. Using the secondary data as the only resource may lead to the problem of credibility and verification. As a result of which the researcher's job becomes a little harder and can sometimes bring extra work.

In order to evaluate the secondary data sources, researchers should consider a few things like who collected the data and for what purpose it was collected. The researcher should examine the source of the collected the data and the purpose for which the data was published. The success of the source (organization) depends on the relationship with the client and proves that the data provided is accurate.

The motivation and capabilities of the individuals who collected the data should be assessed. For this, the qualification, reputation and experience of the individual are considered. Individuals who are working for an independent agency are more genuine compared to the individuals who are working for an organization that is biased towards one side of the research.

The researcher should discover the purpose for which the data is published. The data that are published by a group of interest or promotion can be under suspicion. There should be some evidence that the data is collected and carefully processed, where the conclusions should be supported by the data. In this way, researchers can evaluate secondary data sources.

7.5 ADVANTAGES AND DISADVANTAGES OF SECONDARY DATA ANALYSIS

7.5.1 ADVANTAGES

The choice of primary or secondary data need not be an either/or question. A more useful approach to this question is to focus on selecting data that are appropriate to the research question being studied and the resources available to the researcher; the latter include time, money, and personal expertise. In this spirit, we offer a summary of the major advantages and disadvantages of working with secondary, as opposed to primary, data.

The first major advantage of working with secondary data is economy: because someone else has already collected the data, the researcher does not have to devote resources to this phase of research. Even if the secondary data set must be purchased, the cost is almost certainly lower than the expense of salaries, transportation, and so forth that would be required to collect and process a similar data set from scratch. There is also a savings of time. Because the data are already collected, and frequently also cleaned and stored in electronic format, the researcher can spend the bulk of his or her time analyzing the data. There is also the influence of preference: secondary data analysis is an ideal focus for researchers who prefer to spend their working hours thinking of and testing hypotheses using existing data sets, rather than writing grants to finance the data collection process and supervising student interviewers and data entry clerks.

The second major advantage of using secondary data is the breadth of data available. Few individual researchers would have the resources to collect data from a representative sample in the India.

The third advantage in using secondary data is that often the data collection process is in formed by expertise and professionalism that may not available to smaller research projects.

7.5.2 DISADVANTAGES

One major disadvantage to using secondary data is inherent in its nature : because the data were not collected to answer your specific research questions, particular information that you would like to have may not have been collected. Or it may not have been collected in the ggeographic region you want to study, in the years you would have geographic region you want to study, in the years you would have chosen, or on the specific population that is the focus of your interest. In any case, you can only work with the data that exist, not what you wish had been collected. A related problem is that variables may have been defined or categorized differently than you would have chosen: for instance, a data set may have collected age information in categories rather than as a continuous variable, or race may have been defined as only White/ Other. A third difficulty is that data may have been collected but are not available to the secondary researcher: for instance, address and phone number information for survey respondents may have been recorded by the original research team but will not be released to secondary researchers for confidentiality reasons. If an analysis incorporating geographic information was planned, such a restriction might make the data set unusable. For these reasons, a secondary data set should be examined carefully to confirm that it includes the necessary data, that the data are defined and coded in a manner that allows for the desired analysis, and that the researcher will be allowed to access the data required.

A second major disadvantage of using secondary data is that because the analyst did not participate in the planning and execution of the data collection process, he or she does not know exactly how it was done. More to the point, the analyst does not know how well it was done and therefore how seriously the data are affected by problems such as low response rate or respondent misunderstanding of specific survey questions. Every data collection effort has its “dirty little secrets” that may not invalidate the data but should be taken into account by the analyst. If the analyst was not present during the data collection process, he or she has to try to find this information through other means. Sometimes it is readily available; for instance, many of the federaldata sets have extensive documentation of their data collection procedures, refusal rates, and other technical information available on their websites or in published reports. However, many other secondary data sets are not accompanied by this type of information, and the analyst must learn to “read between the lines” and consider what problems might have been encountered in the data collection process.

Whilst the benefits of secondary sources are considerable, their shortcomings have to be acknowledged. There is a need to evaluate the quality of both the source of the data and the data itself. The main problems may be categorized as follows:

Definitions—The researcher has to be careful, when making use of secondary data, of the definitions used by those responsible for its

preparation. Suppose, for example, researchers are interested in rural communities and their average family size. If published statistics are consulted then a check must be done on how terms such as “family size” have been defined. They may refer only to the nucleus family or include the extended family. Even apparently simple terms such as ‘farm size’ need careful handling. Such figures may refer to any one of the following: the land an individual owns, the land an individual owns plus any additional land he/she rents, the land an individual owns minus any land he/she rents out, all of his land or only that part of it which he actually cultivates. It should be noted that definitions may change over time and where this is not recognised erroneous conclusions may be drawn. Geographical areas may have their boundaries redefined, units of measurement and grades may change and imported goods can be reclassified from time to time for purposes of levying customs and excise duties.

Measurement error—When a researcher conducts fieldwork she/he is possibly able to estimate inaccuracies in measurement through the standard deviation and standard error, but these are sometimes not published in secondary sources. The only solution is to try to speak to the individuals involved in the collection of the data to obtain some guidance on the level of accuracy of the data. The problem is sometimes not so much ‘error’ but differences in levels of accuracy required by decision makers. When the research has to do with large investments in, say, food manufacturing, management will want to set very tight margins of error in making market demand estimates. In other cases, having a high level of accuracy is not so critical. For instance, if a food manufacturer is merely assessing the prospects for one more flavour for a snack food already produced by the company then there is no need for highly accurate estimates in order to make the investment decision.

Source bias—Researchers have to be aware of vested interests when they consult secondary sources. Those responsible for their compilation may have reasons for wishing to present a more optimistic or pessimistic set of results for their organisation. It is not unknown, for example, for officials responsible for estimating food shortages to exaggerate figures before sending aid requests to potential donors. Similarly, and with equal frequency, commercial organisations have been known to inflate estimates of their market shares.

Reliability—The reliability of published statistics may vary over time. It is not uncommon, for example, for the systems of collecting data to have changed over time but without any indication of this to the reader of published statistics. Geographical or administrative boundaries may be changed by government, or the basis for stratifying a sample may have altered. Other aspects of research methodology that affect the reliability of secondary data is the sample size, response rate, questionnaire design and modes of analysis.

Time scale—Most censuses take place at 10 year intervals, so data from this and other published sources may be out-of-date at the time the researcher wants to make use of the statistics.The time period during which secondary data was first compiled may have a substantial effect upon the nature of the data. For instance, the significantincrease in the price obtained for Ugandan coffee in the mid-90's could be interpreted as evidence of the effectiveness of the rehabilitation programme that set out to restore coffee estates which had fallen into a state of disrepair. However, more knowledgeable coffee market experts would interpret the rise in Ugandan coffee prices in the context of large scale destruction of the Brazilian coffee crop, due to heavy frosts, in 1994, Brazil being the largest coffee producer in the world. Whenever possible, marketing researchers ought to use multiple sources of secondary data. In this way, these different sources can be cross-checked as confirmation of one another. Where differences occur an explanation for these must be found or the data should be set aside.

7.6 LOCATING APPROPRIATE SECONDARY DATA

There is a vast quantity of secondary data are available to the individual researcher. However, the sheer quantity of data available, and the fact that the data are collected and archived by many different governmental and private entities, means that the process of locating appropriate secondary data is not always straightforward. There is no single process to be followed in every case, but we offer two examples of the process of locating and analyzing secondary data to address a specific research question or problem.

The most typical way to use secondary data for research is to begin with a research question and seek a data set that will allow analysis of that question. An alternative method is to begin by selecting from among the available secondary data sets, and then formulating a research question that may be answered using the data chosen. Although the first method conforms more to standard beliefs about how research is done, the second approach is particularly useful in classroom instruction, and both methods can produce quality research.

If the researcher begins with a question and then seeks out an appropriate data set, the following generalized sequence of procedures may be useful :

1. Define the question you want to study; for instance, "How does the experience of racism affect an individual's health?"

2. Specify the population you want to study. Are you interested in children, adults, or people of all ages? What races or ethnicities do you want to study? Do you want to analyze a national sample or one confined to a smaller area? What is the range of years you

would consider (e.g., you may only be interested in data collected over the last 5 years)?

3. Specify what other variables you want to include in your analysis. In this example, you might believe that it was important to have information about the respondents' race, Hispanic ethnicity, age, gender, income, and educational level in order to include those factors in your analysis. If so, you must confirm that the data you desire are contained in the data set that you choose and that they are recorded in a manner that is useful to you. If you are interested in comparing the experiences of Hispanic Blacks and non-Hispanic Blacks, information about Hispanic ethnicity would need to be recorded in the data set independently of information about race.

4. Specify what kind of data is most appropriate for your research question: for instance, can it best be addressed through a national survey, examination of hospital claims records, or transcriptions of interviews? Also, specify if there are any specific data collection techniques you believe are particularly appropriate or inappropriate for your question. For instance, if you do not believe people would answer questions about racism honestly in a personal interview, you would not consider any data sets collected using that technique. However, if you believe that a telephone survey would be the best way to collect this information, you might begin your search by looking at surveys that used this data collection method.

5. Create a list of data sets that include information related to your research question and examine them to see if they meet your other requirements (age range included, year of collection, etc.). This is where the interactive process begins because you may have to revise either your question or your data requirements, depending on the data that are available to you.

6. Once you have chosen your data set, examine the variables you intend to use for the analysis of problems such as missing data or out-of range values. Also, read whatever information you can find about the data collection process, data cleaning procedures, and so on in order to evaluate whether the data quality is sufficient to meet your needs. If so, continue with the analysis; if not, either devise a way to work around it (e.g., by imputing values for the missing data) or choose another data set.

If you take the approach of beginning with a data set and crafting a research question that can be addressed using it, the process is similar, but the order of events is different. In this case, you would begin by looking at the variables contained in the data set and considering how you might combine them to create an interesting question. The process can begin with a germ of an idea, which may reflect your personal interests or a question that has arisen in your work. For instance, you might be

interested in how disability affects the amount of physical activity in which a person engages. You then need to operationalize this question so it may be tested using the variables available in the data set :

how will you define disability, and how will you define physical activity?

At this point, a Medline search for related articles would be in order, to see how others have addressed similar questions and whether they have done so with the data set you will be using. This step will help keep you from reinventing the wheel and will place your research in context.

7.7 QUESTIONS TO ASK ABOUT ANY SECONDARY DATA SET

Once you have located a secondary data set that you think is appropriate for your analysis, you need to learn as much as you can about why and how it was collected. In particular, you will want to answer the following three questions:

1. What was the original purpose for which the data were collected?

2. What kind of data is it, and when and how were the data collected?

3. What cleaning and/or recoding procedures have been applied to the data?

Sources for this information include the website of the agency or other entity responsible for collecting and/or making the data available, published reports, research articles based on the data, and personal communications with relevant individuals. For instance, many of the federal agency websites include one or more contact people who are available to answer questions about the data collected by that agency, and a Medline search will often produce citations to reports and articles discussing the procedures used to collect particular data sets.

The question of determining the original purpose of the project that produced the data is important because its influence may be present in other characteristics of the data, from the population targeted to the specific wording of questions included in a survey. Because you were not involved in planning phases for the project whose data you will analyze, you need this information in order to place the data in context.

To take an extreme case, you would certainly want to know if a research project on the health effects of smoking was sponsored by a tobacco company or by a nonprofit dedicated to smoking prevention. You would also like to know if there was any particular philosophy or model of health behavior that shaped the project: for instance, was a smoking cessation program structured using the Trans theoretical Model? Knowledge of the core philosophical beliefs behind a research project can illuminate the reasons for many choices made in the planning and

execution of the research and will be reflected in the end product, the data you are proposing to analyze.

It is almost impossible to know too much about the data collection process because it can influence the quality of the data in many ways, some of them not obvious. To start with, you need to know when the data were collected. Second, you want to know the process by which the data were collected: was it via telephone interviews, in-person interviews, abstraction of hospital records, or some other technique? Third, you want to know the details of the data collection process. Questions in this regard include who actually did the data collection, how extensive was their training, and how carefully were they supervised. If the data were collected through chart review, what specific instructions were given to the reviewers? If the data were collected through a survey, what was the response rate? How many efforts were made to collect data from non responders? If data were collected through a telephone survey, how were numbers selected? Was there any attempt to correct for the bias introduced because households without a telephone are not a random sample of all households? The issues of survey data quality are the same whether the data set is primary or secondary.

The third major question in working with any secondary data set is what was done to the data after they were collected. For instance, almost all data sets include some missing data. Were these data left as missing, or were values imputed, and if so, how was the imputation done? Was any data cleaning done to remove out-of-range values, and were those cases assigned missing values or was some other procedure followed? Were certain combinations of answers considered invalid, and if so, how were they treated?

7.7.1 DEFINING SECONDARY DATA ANANLYSIS

Numerous definitions of secondary data analysis appear in the literature, many with subtle differences which together suggest a lack of consensus about what is meant by the term. For example, one relatively straightforward definition of the secondary analysis of survey data was suggested by Hyman (1972: 1), as ‘the extraction of knowledge on topics other than those which were the focus of the original survey’.

Other definitions of secondary analysis have emphasised its usefulness for exploring new research questions: ‘the study of specific problems through analysis of existing data which were originally collected for another purpose’ (Glaser 1963: 11); or: ‘the further analysis of an existing dataset with the aim of addressing a research question distinct from that for which the dataset was originally collected and generating novel interpretations and conclusions’ (Hewson 2006: 274). However, such definitions appear to disregard the potential of secondary analysis in re-analysing existing datasets with novel statistical or theoretical approaches in such a way that: ‘secondary analysis is the re-analysis of data for the purpose of answering the original research questions with

better statistical techniques, or answering new research questions with old data' (Glass 1976: 3). One apparent area of consensus among those looking for a definition of secondary analysis is that it should involve the analysis of someone else's data: 'a collection of data obtained by another researcher which is available for re-analysis' (Sobal 1981: 149). However, this has been disputed as: 'even re-analysis of one's own data is secondary data analysis if it has a new purpose or is in response to a methodological critique' (Schutt 2007: 4127). Other researchers point to the ubiquitous credentials of secondary analysis as 'neither a specific regime of analytic procedures nor a statistical technique, [but] . . . a set of research endeavours that use existing materials' (Kiecolt and Nathan 1985: 10). And emphasise its difference from primary analysis: 'which involves both data collection and analysis, while secondary analysis requires the application of creative analytical techniques to data that have been amassed by others' (Kiecolt and Nathan 1985: 10).

Given the rather subtle differences in the definition and interpretation of secondary analysis that we see here, it seems likely that neat distinctions between primary and secondary data will not always be possible (Dale et al. 1988). Such lack of consensus might leave one wishing to adopt a very general definition of secondary analysis such as that offered by Jary and Jary (2000): as 'any inquiry based on the re-analysis of previously analysed research data' (p. 540) or one such as Hakim's:

> [S]econdary data analysis is any further analysis of an existing dataset which presents interpretations, conclusions or knowledge additional to, or different from, those produced in the first report on the inquiry as a whole and its main results. (Hakim 1982a: 1)

Whichever definition one favours, secondary analysis should be 'an empirical exercise carried out on data that has already been gathered or compiled in some way' (Dale et al. 1988: 3). This may involve using the original, or novel, research questions, statistical approaches and theoretical frameworks; and may be undertaken by the original researcher or by someone new.

In education, recent methodological preoccupations in both the UK and the USA have focused on the quality and relevance of research in the field. Educational research is widely viewed as having an 'awful' reputation (Kaestle 1993) of being 'not very influential, useful or well funded' (Burkhardt and Schoenfeld 2003: 3), of following fads (Slavin 1989) and of being of indifferent quality (Hargreaves 1996; Tooley with Darby 1998). While the call in both countries is for a greater unity between research and practice (Burkhardt and Schoenfeld 2003; Hargreaves 1996), there is some divergence in how this might actually be achieved.

In the USA, legislation introduced in 2001 stipulates that all federally funded research must adopt scientifically based research methods (Eisenhart and Towne 2003; Olson and Viadero 2002). For some this is

seen as an opportunity to elevate educational research to the status of medicine and agriculture (Slavin 2002) and for 'nurturing and reinforcing' a scientific research culture in the field (Feuer et al. 2002: 4). For others, it exemplifies the privileging of certain research methods: namely, experiments and randomised control trials, a failure to understand the complexity of the field and a lack of commitment by the US federal government to promoting true evidence-based practice (Berliner 2002).

In the UK, general methodological concerns centre on a perceived imbalance in the types of research method adopted by educational and other social science researchers (ESRC 2006). Much of this concern is centred on the 'dubious dichotomy' (Payne et al. 2004: 153) that exists between 'quantitative' and 'qualitative' methods. For example, according to the Economic and Social Research Council (ESRC), 'the lack of quantitative skills is endemic in many areas of Social Science and . . . there is an urgent need to enhance research quality' (ESRC 2006: 12).

The ESRC has demonstrated its commitment to building research capacity in the field of quantitative methods through sponsorship of centres such as the Social Science Research Methods hub at Southampton University, the development of data management infrastructures such as UK Data Archive and its sponsorship of the European Social Survey and other internationally renowned longitudinal studies such as the National Child Development Study and the British Household Panel Survey.

In the field of education, a lack of quantitative skills among researchers is seen as being one of its most significant 'defects' (Gorard et al. 2003a: 19) and, according to key stakeholders, there is a significant lack of quantitative skills in the field:

There is a widely acknowledged absence of quantitative research of particular kinds, especially, there's a weakness, there's a relative absence and there's no mechanism for addressing that currently. (HE researcher and ESRC Teaching and Learning Research Programme team leader, cited in Gorard et al. 2003a: 13)

In addition to concerns about a lack of quantitative work in education, there are also concerns about the quality of such work: I think you can get terrible quantitative work, there are people who just think there's a kind of nonproblematic general linear model reality out there and you just tag variables and start with race and that's it . . . switch on SPSS, press the buttons with all the defaults on is garbage in garbage out. (chief executive of research funding body, cited in Gorard et al. 2003a: 17)

Thus it can be argued that building research capacity in the use of quantitative methods is not just about increasing the volume of such research but focusing on more basic quantitative techniques. The concern appears to be not only about the relative shortage of quantitative research 'but also a shortage of the quantitative research skills required to be able to understand, and critically review, quantitative research' (Gorard et al. 2003a: 19). The next section considers these concerns in light of a review

of publications submitted to the 2001 Research Assessment Exercise (RAE) and a more recent examination of the published output of eight social science journals.

7.8 WHAT EVIDENCE IS THERE FOR A LACK OF QUANTITATIVE SKILLS?

In the UK, the funding councils for higher education institutions (HEIs) have highlighted the importance of comparative measures of research excellence. This has resulted in the Research Assessment Exercise (the RAE), the purpose of which is to produce quality profiles for all research activity in UK HEIs. Indeed, the RAE is the principal means by which institutions assure themselves of the quality of research undertaken in the HE sector (RAE 2008). The first RAE was carried out in 1986 with four subsequent exercises prior to the latest RAE in 2008. In RAE 2001 research quality ratings were based on a range from 1–5*, according to how much of the work was judged to reach national or international levels of excellence. In this way, the quality ratings derived from the RAE provide a single measure of research strength in UK higher education. In this section, therefore, we report the findings from an analysis of the results for the most recently available RAE, RAE 2001, and examine the evidence for a lack of quantitative skills in the social sciences and in the field of education, in particular (Gorard et al. 2003b). We begin by looking at the overview reports for the four main RAE social science panels – education, social policy, social work and sociology. Their comments suggest that in some areas of the social sciences, a concern over the paucity of quantitative research does remain:

> The strengths of qualitative, and relatively small scale, research have been complemented by some increase in work that has also drawn on quantitative data. However, there is room for more approaches that use advanced quantitative methodologies and for education to play an active and significant role in methodological innovation. (Education Panel RAE 2001a)

> There was a paucity of quantitative analysis, particularly that combined with qualitative approaches. Quantitative work, while sometimes of very high quality, is rare and not always good when found. (Social Work Panel RAE 2001b)

Note that similar concerns about quantitative work in sociology and social policy and administration were not raised in the RAE 2001 overview documents for these two panels (RAE 2001b). This, of course, does not necessarily mean that such concerns do not exist (and, indeed, some of the evidence presented in the following section argues that, at least in sociology, they well might) rather it may simply be the case that they were just not reported by the RAE 2001 panels.

Because of the variety of submissions in the field of education, the RAE panel for 2001 required that institutions also provide additional

information about the methods undertaken in each submission. This theory/ method field was to be used to classify 'any conceptual framework within which the research may be set, together with a brief indication of the design and/or research method' (RAE 2001c). This means that for education, in addition to providing indicators of quality, the RAE is also able to provide a much more detailed audit of the research methods that were being used (Gorard et al. 2003b).

While the use of the theory/method field to classify research methods is not itself unproblematic, an analysis of the 8700 publications submitted to the education panel in RAE 2001 does reveal an interesting balance of methodological approaches. Table 1.1 shows the number of times a particular method was reported as being the first or main method used in submissions to the education panel in RAE 2001. According to the entries reported in the theory/method field, a large number of submissions (32%) were not clearly empirical: for example, they involved literature reviews or thought pieces. Of the remainder around 14% were largely or wholly quantitative and 28% were largely or wholly qualitative – a balance of around 2:1 (Gorard et al. 2003b). Notice also the relatively lowly position of secondary data analysis, which was listed as the main method for fewer than 1% of the pieces submitted to RAE 2001.

The number of actual methods used in the sample of work submitted for RAE 2001 can be compared with a self-reported audit of methods that was undertaken by Gorard et al. (2003b) as part of the same study and reproduced in Table 1.2. This self-audit was completed by 521 researchers who were actively engaged in empirical research in the UK in 2002– 2003. They were asked to summarise their knowledge and use of a range of methods taken from a specified list of nearly 300 approaches. The approaches were collapsed into similar categories as used for the theory/ method analysis described in Table 1.1. The responses indicate a considerable capacity to undertake quantitative work in the field of education: around three-quarters of the research community reported having conducted a survey and around 65% report having used a secondary numeric source of data.

If we consider the number of times researchers report using quantitative and qualitative methods in their everyday research and the types of method actually adopted in publications that were submitted to RAE 2001, we see that while similar proportions of researchers report using quantitative and qualitative techniques in submissions to RAE 2001, twice as many publications were qualitative in nature than were quantitative. While this gap is perhaps smaller than expected given the stakeholders' concerns that were explored earlier, it does suggest that perhaps researchers are not quite as eclectic in their use of methods as they might see themselves to be. The differences between those reporting using secondary data analysis and those actually using this method are more apparent than for the broad area of quantitative methods.

7.8.1 DEVELOPMENT OF SECONDARY ANALYSIS

Secondary analysis has a long pedigree. In 1790 the first national population census was undertaken in the USA, followed in Great Britain in 1801. The potential of these data for secondary analysis and their contribution to the social sciences is exemplified by Booth's work on occupation patterns that were derived from secondary analysis of the 1801–1881 UK Censuses (Booth 1886). By the end of the nineteenth century, the large-scale studies of urban poverty that were pioneered by Joseph Rowntree in York and Charles Booth in London marked the start of the social survey movement and the wealth of opportunities it afforded for secondary analysis. Official records were also put to use for secondary analysis, most notably in Durkheim's important research into the sociology of suicide (Durkheim 1952) which is considered in more detail in the next chapter. In the United States, secondary analysis as a research strategy coincides with the rapid increase in the number of attitudinal surveys before the Second World War. The 'first notable effort' at secondary analysis from a theoretical and methodological perspective was The American Soldier (Glaser 1963: 11). The American Soldier is part of a collection of around 260 studies of several hundred thousand army personnel that was undertaken during the Second World War. Its four volumes are the accumulation of several years of investigation and include surveys and observations of soldiers' lives before, during and after combat (Lazarsfeld 1949). The work provides a fascinating insight into the lives of service personnel: for example, their relationships with their fellow soldiers and officers, their changing attitudes towards war as a result of experiencing combat and their adaptation to life as civilians. In addition, the collection is seen as a 'fine model indeed for the secondary analysis of a rich archive of data collected for other purposes' (Smith 1984: 196, cited in Williams 1989: 157). The legacy of The American Soldier has not just been its contribution to the field of military history. Its theoretical and conceptual developments have found influence in developing theories of class position, of race relations, of 'relative deprivation' and social adjustment; and its methodological applications have led the field in the development of attitude scaling and latent structural analysis (Glaser 1963; Lazarsfeld 1949; Williams 1989). Other classic examples of secondary analyses include the re-analysis of the Coleman Report during the early 1970s (Mosteller and Moynihan 1972). This re-analysis was in a large part prompted by the huge interest generated by the study Equality of Educational Opportunity (Coleman et al. 1966). The original study found that the variation in learner outcomes in schools was strongly related to factors external to the school, such as family background and that the school itself had a relatively small part to play in mediating between different academic outcomes. Perhaps unsurprisingly, given the reception that greeted the original study, a reanalysis of the findings soon followed. Undertaken by faculty members at Harvard University, the secondary analysis identified some discrepancies, errors and issues with the primary study but largely 'affirmed and strengthened' (Smith 1972: 311) the original findings. Another example of important secondary analyses of

influential studies is Elashoff and Snow's (1971) re-analysis of Rosenthal and Jacobsen's (1968) study Pygmalion in the Classroom.

In the UK, one important use of secondary analysis in the sociology of education is the 1980 re-analysis of the 1972 Oxford Mobility Project dataset by Halsey et al. This study is not only important for its empirical and theoretical contributions, as the subsequent book Origins and Destinations: Family, Class and Education in Modern Britain is one of the classics in the sociology of education.

Works such as The American Soldier, Origins and Destinations and Suicide set the standard for secondary analysis in the social sciences, exemplified by their methodological, theoretical and substantive contributions to the development of the field. However, arguably the potential for secondary analysis as an important social science method has never fully been realized in many branches of the discipline, as many of the objections to its use attest. While secondary analysis may be more established as a research method in the United States (Hakim 1982), it has remained relatively underused in many areas of the social sciences in the UK. Indeed, the 1988 edition of The Penguin Dictionary of Sociology contains no entry for secondary analysis. In the entry for 'official statistics' sociologists were advised to 'approach such data with caution', its use was described as 'perilous' and data collected by government for its own use considered 'meaningless' (Abercrombie et al. 1988: 170). Perhaps some minor rehabilitation of the field has occurred in the intervening years as a definition for 'secondary analysis' was suggested in the 2000 edition and the authors' objections to the use of official statistics have been somewhat mollified in these later editions (Abercrombie et al. 2000).

7.9 SUMMARY

A search of secondary data sources should precede any primary research activity. Secondary data may be sufficient to solve the problem, or at least it helps the reader better understand the problem under study. Secondary data is cheaper and quicker to collect than primary data and can be more accurate.

Before making use of secondary data there is need to evaluate both the data itself and its source. Particular attention should be paid to definitions used, measurement error, source bias, reliability and the time span of the secondary data. Where possible, multiple data sources should be used so that one source can be cross-checked for consistency with another.

A great deal of potentially useful secondary information already exists within enterprises. Typically useful information would be that relating to sales, finance, production, storage and transportation.

Where a serious search of secondary sources is undertaken then the marketing researcher often finds an abundance of relevant material. Searches of printed secondary data should begin with a consultation of referral sources such as directories, handbooks, indexes, and the like.

It will almost certainly become the case, in all parts of the world, that electronic information sources will eventually supersede traditional printed sources. With the advent of Internet and CD-ROM, searches of secondary sources are becoming more efficient and more effective. Computer-based information systems give access to four different types of database bibliographic, numeric, directories and full-text.

It is a field with a long pedigree that extends back to the pioneers of the Victorian social surveys through to the opinion polls, censuses, administrative records and international surveys of today. Despite the huge potential of secondary data analysis methods, there is limited consensus on a definition for the term, although commentators tend to agree that it involves some form of re-analysis or reporting of existing data. Such flexibility in its definition perhaps reinforces its suitability to involve any analysis involving the reinterpretation of existing data which bring new methodological and theoretical perspectives or which adopt the original or novel research questions and which are undertaken by the original researcher or individuals new to the data. It is also important to emphasise that the analysis of secondary data need not involve the manipulation of complex and large datasets. Indeed, much of the data that are available for secondary analysis are in the form of aggregate data – that is, data that have already been analysed and are presented in summary form.

It took as its starting point contemporary concerns that quantitative methods are underused in the field and, by focusing specifically on the field of education, examined the disparity between the type of methods researchers claim to use and their actual published output, which indicates what methods they actually do use; an examination that points, in particular, to a lack of methodological pluralism in the use of secondary data in education research. One of the reasons why secondary data analysis is relatively underused in social research can perhaps be attributed to the wide range of criticisms and concerns that the field attracts.

7.10 FURTHER STUDY

1. Mathematical Economics : G.S. Monga
2. Fundamental of Statistics : D.N. Elhance
3. Principal of Statistics : S.B. Gupta

UNIT-08 CLASSIFICATION OF DATA

Objectives

After reading this unit you should be able to know about

- Introduction, meaning & characteristics of an ideal classification.
- Types of classification &seriation of data
- Rules of classification.

Structure

8.1 Introduction

8.2 Meaning & characteristics of an ideal classification

8.3 types of classification

8.4 Seriation of data

8.5 Formation of discrate frequency distribution

8.6 Formulation of continuous frequency distribution

8.7 Rules of classification

8.8 Conclusion/Summary

8.9 Further Study

8.1 INTRODUCTION

In this chapter we are discussing classification of data. The data contained in schedules or questionnnaire are in a form which does not give an idea about the salient features of the problem under study. They are not directly fit for analysis and interpretation. If, for example, the figures of the monthly expenses of 2,000 students residing in Allahabad University hostels are before us, as collected, it would not be possible to draw any inference from them. For purposes of comparison, analysis and interpretation, it is essential that the data are in a proper form. Further, it is also essential that the likes are separated from the unlikes. All the 2, 000 students, no doubt, are alike in the sense that all of them belong to a particular university and live in hostels but they differ in other respects. Some may be living in single-seated rooms and others in double or treble-seated rooms; some may be living in costlier hostels, and others in comparatively cheaper ones; some may be having their private messing arragements while others may have joined the common mess. Thus, even though the data collected relate to one set of persons, yet there may be many types of dissmilarities even within this group. For the purpose of

analysis and interpretation, data have to be divided in homogenous groups. In order to remove these defects--of volume and heterogeneity--statistical data are tabulated with a view to present an condensed and homogeneous picture. But before the tabulation of data, it is necessary to arrange them in homogeneous groups or classes so that there may be no difficulty in tabulation.

8.2 MEANING AND CHARACTERISTICS OF AN IDEAL CLASSIFICATION

The process of arranging data in groups or classes according to resemblances and similarities is technically called classification. Thus, by classification we try to strike a note of homogeneity in the heterogeneous elements of the collected information. Classificatio gives expresssion to the similarilities which may be found in the diversity of individual units. In classification of data, units having a common characteristic are placed in one class and, in this fashion the whole data are divided into a number of classses. Even after classificatin, the statistical data are not fit for comparison and interpretation and need proper tabulation.After tabulation of data, statistical analysis and interpretation are possible. **Classification is a preliminary to tabulation and it prepares the ground for proper presentation of statistical facts.**

Despite the fact that classification is a very important preliminary to statistical analysis, no hard and fast rules can be laid douwn for it. Technically, the classification of data in each investigation has to be decided after taking into account the nature, scope and purpose of the enquiry. However, an ideal classification should prossess the following **characterisitics** :

1. **It should be unambiguous :** If there is ambiguity in classification, the very purpose for which it is meant is not served. Classificaiton is meant for removing ambiguity. It is necessary that the various classes should be so defined that there is no room for doubt or confusion. It is by no means an easy task. If we have to divide the population of India into two classes, say, litrates and illitrates, exahustive definition of the terms used, would be essential. Some criterion has to be laid down.In the last census of pupulation of India, a literate was defined as one who could read and write a simple letter. This is technically not a very satisfactory definition. After all, what is meant by a simple letter, is a point on which there can be difference of opinion. But, for practical purposes, the definition can be said to be fairly satisfactory.

2. **It should be stable :** The ideal classification should have the merit of stability. If a classification is not stable and if each time an enquiry is conducted, it has to be changed, the data would not be fit for comarison. The occupational classification in the Indian pupulation census suffers from this defect. Various occupations

have been defined in different ways in successive censuses and, as such, these figures are not strictly comparable.

3. **It should be flexible:** A good classificaiton should be flexible and should have the capacity of adjustment to new situations and circumstances. When we talk of stability of classification, we do not mean rigidity of classes. The term is used in a relative sense.Noclassificaiton can be stable for ever, With change in time, some classes become obsolete and have to be dropped, while fresh classes have also to be added. An ideal classificaiton should be such that it can adjust itself to these changes and yet retain its stability. The data should be divided into a few major classes.Only small sub-classes may need a change and the classificaiton can, thus, retain the merit of stabiltiy and yet possess flexibility.

8.3 TYPES OF CLASSIFICATION

Statistical data are clasified on the basis of the charactersistics possessed by different groups of units of a universe.As has been pointed out earlier, these charactersitics give expression to the unity of attributes which may be traced in a diversity of individual units.Unity in certain grups of units may be on the basis of geopgraphical distribution or on the basis of chronological arrangement or on the basis of attributes or certain quantitative characteristics. Thus, broadly speaking, data can be classified on the following four bases :

1. Geographical, i.e., in relation to place
2. Chronological, i.e., on the baisi of time
3. Qualitative, i.e., according to some attributes
4. Quantitative, i.e., in terms of magnitude.

1. **Geographical Classification :** In geographical classification, data are classified on the basis of place. If, for example, we write down the population of the Indian Union on the basis of various States or if we write down the number of students in different universities of the country, or production of wheat in different geographical areas of the country, the series that we would get would need classification on the basis of geographical distribution. Sereies which are arranged on the basis of place are called spatial series.

The following is an example of a geopgraphical distribution :

Country	National income in U.S. dollars
Canada	7930
U.S.A.	7880
West Germany	7510
France	6730
U.K.	4180
U.S.S.R.	2800
India	140

Geographical classificaitons are generally listed in alphabetical order or listed by the frequency size to emphasise the importance of various geographical regions as has been done in the table given above.

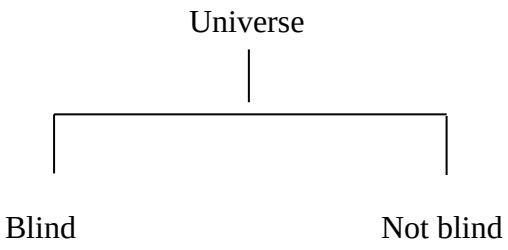
2. **Chronological classification :** When the data are classified on the basis of time, then it is known as chronological classification. Such series are also known as time series because one of the variables in them is time. If the population of India during the last eight censuses is classified, it will result in a time series or chronological classification.The following table would give an idea of chronological classification:

Production of Refrigerators by Company 'X'

1995	12,800
1996	13,500
1997	14,200
1998	15,800
1999	19,300
2000	17,000
2001	17,500
2002	18,300

3. **Qualitative classification** : In qualitative analysis, data are classified on the basis fo descriptive characteristics or on the basis of attributes like sex, literacy, religion, education, etc., which cannot be measured. We can only find out the presence or absence of a particular attribute in an individual.If we are studying the problem of blindness or deafness, we can only find out whether an individual is blind or not blind, whether he is deaf or not deaf. We cannot measure blindness or deafness.In such cases, the attribute divides the universe only into two parts: one in which the attribute in question is present and the other in which it is not present.In case of blindness, there are only two classes, i.e., those who are blind and those who are not blind. These classes are mutually exclusive so that those who are blind cannot come into the category of those who are not blind. Such a classificatin is called classification by dichotomy.

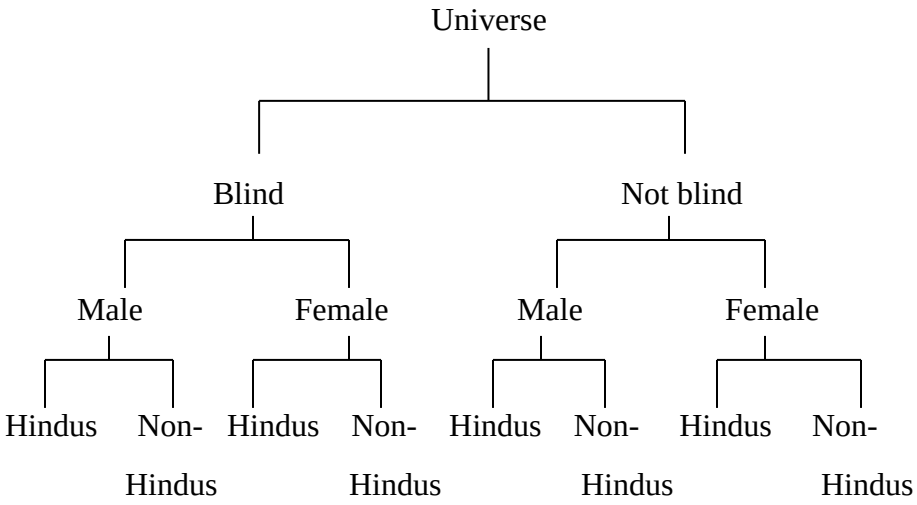
This form of simple classification would be as follows:



If, however, more than one attribute is being studied simultaneiously, the data would be divided into a number of classes. If the problem of blindness is studied sex-wise, there are two attributes under study, namely, blindness and sex. A person can be either blind or not blind; further, a person can be either a male or a female. Eeach of the two attributes is capable of division in two classes. The data would, thus, be divided in four classes, (1) males who are blind, (2) males who are not blind, (3) feamales who are blind, (4) females who are not blind. The study can be further extended if we have a third attribute, say, religion. Now, each of the above four classes is capable of further subdivision on

the basis of religion.Such classification in which more than one attribute is taken into account is called manifold classificaiton.

Data in such a case would be classified as follows:



Arbitrary nature of classification : In the various groups which are formed in the above mentioned manner the differences are not always natural or very well defined. Ordinarily, such classsification is of an arbitrary nature. If the universe is divided in two groups--tall men and short men--we shall have to give arbitrary definitions of the two classes. It can be said tht those who are 160 centimetres or above are tall and those who are less than 160 centimetres are short. The classification is obviously arbitrary. In these cases where a particular attribute is decided on the basis of quantitiative study, as in the above case of tall and short men, the classification is comparatively more definite and precise. But, this is not always possible.Many attributes cannot be studied with the help of figures. The difference between literacy and illiteracy is an example. Here, one attribute gradually changes into another attribute and ther is no clear-cut line of demarcation. The difference between literate and an illitrate is always a matter of opinion.There may be many persons, whom it would be difficult to classify as either literates or as illitrates. Whenever data are classified according to attributes, this point should be kept in mind and attempts must be made to define the attributes, in such a manner that there is the least possibility of doubt and ambiguity.

4. **Quantitative classification :** As the term suggests, quantitiative classification refers to classification that is based on figures or, in other words, it is a classification which is based on such characteristics which are capable of quantitative measurement.

Such characteristics can be height, weight, income, expenditure, numebr of marks obtained by students of a class, etc. For example, marks obtained by 60 students of a class can be classified in the following manner :

Marks	No. of students
0-20	5
20-40	15
40-60	25
60-80	10
80-100	5

In the above frequency distribution, there are 2 elements : (i) a variable, and (ii) a frequency. In the above example, "marks" is a variable and the number of students indicate "frequencies" Variable refers to a characteristic which varies. It can be either continuous or discrete. A continuous variable is one which is capable of assuming any fractional value within a range of numbers. Thus, if we talk of the height of individuals or their weight, the variables can have any value within conceivable limits. The height of a man can be 160 centimetres or 160.1 centimetres or 160.12 centimetres or any other value. *In a continuous variable, the data are obtained by measurement, rather than counting.* A discrete variable, on the other hand, cannot assume a fractional value. For example, the number of students in a class-room or the number of class rooms in a school or college cannot assume fractional values.

Discrete distributions are obtained by counting and not by measuremet. Thus, in continuous series the statistical unit is capable of division and can be measured in fractions of any size, no matter how small.A tonne of coal can be divided in a 100, 1,000, 10,000 or even more parts. Theoretically, a tonne of coal can be divided into a limiless number of sub-divisions. In discrete series, **statistical unit is either not divisible or is not divided.** We can imagine half a tonne, one-fourth of a gallon or one-tenth of a pound, but it would be absurd to talk about half a son, one-fourth of a student and one-tenth of a wife. Here, the unit is complete and indivisible. We can, however, have discrete series, even from divisible units whre they are not conventionally divided. For

example, marks are given to the students in whole numbers. It is possible in such a case to have a discrete series of marks.

Following are examples of continuous and discrete variables along with related frequencies. These are respectively known as continuous series and discrete series.

Height (cms.)	(a) Continous series	(b) Discrete series	
	No. of students	No. of children per couple	No.of couples
145-150	10	0	20
150-155	20	1	30
155-160	18	2	50
160-165	10	3	70
165-170	12	4	30
Total	70	Total	200

8.4 SERIATION OF DATA

The process of seriation is closely associated with classification of data. According to L.R.Connor *"If two variable quantities can be arranged side by side so that the measurable differnces in the one correspond to the measurable differences in the other the result is said to form a statistical series."*

As discussed in the previous section, seriation can be done on the basis of time, space, or some condition.So far as time series and spatial series are concerned, there is no problem in their formulation. Frequencies can be noted down on the basis of time or space. However, when series are formed on the basis of changes in some condition (like age, weight, marks, produciton, etc.) as mentioned earlier, the series can be either discrete or continous.We shall discuss below the methods of forming discrete and continous series.

8.5 FORMATION OF DISCRETE FREQUENCY DISTRIBUTION

Formation of discrete frequency distribution is a very simple task. The number of times a particular variable has repeated is noted down and the total is the frequency of that class. For example, if we have to form a

discrete series of the number of rooms in the houses occupied by 25 families, we will note down the number of families having houses containing 1,2,3, or more number of rooms. Suppose, the data collected about 25 families is as given below :

No. of rooms in the houses kept by 25 families

1	2	4	3	4
2	5	3	2	2
4	1	2	3	5
1	3	5	1	3
3	1	3	1	1

Now, if we have to form discrete series out of th above data, we will first note down the values from the lowest to the highest. After this we will put a vertical line (bar) opposite the value to which the item refers. To facilitate counting generally blocks of five bars are prepared (4 bars are kept vertical and the 5th bar diagonally crossing fur vertical bars). This facilitates counting. In the above example, values are 1,2,3,4,5 which we wil put in one column, and in the other column, the number of times each value has occurred would be noted with the help of bars.After this, the frequencey would be found out. The data would, thus, emerge as follows :

Number of rooms	**Tallies**	**Frequency**
1	𝍸 II	7
2	𝍸	5
3	𝍸 II	7
4	III	3
5	III	3
		25

Thus, from the above, it is clear that out of 25 families, 7 were living in 1 room each, 5 in 2 rooms each 7 in 3 rooms each, 3 in 4 rooms each and 3 in 5 rooms each.

8.6 FORMATION OF A CONTINUOUS FREQUENCY DISTRIBUTION

Formulation of continuous frequency distribution is done in a number of ways. Incontinous frequency distribution, data are divided into class intervals instead of individual values, as is done in case of discrete frequency distribution. Thus, if the heights of students are to be classified, they would be grouped in class intervals like 150-152 centimeters, 152-154 centimetres and so on. Difference between the upper limit and the lower limit of a class intervals is called magnitude of the class interval. Thus, in the above illustration the magnitude of the class interval is 2 centimetres, i.e., 152 minus 150 centimetres. The upper and lower limits of a class are called *class boundaries or class limits*.In the above illustration, the class boundaries are 150 cms. And 152 cms. The lower boundary of the class interval is 150 cms. and the upper bouday 152 cms.

Claassification according to class intervals involves four things :

1. Determination of the number of classes.
2. Determination of the magnitude of the class intervals.
3. Choice of the class limits or class boundaries.
4. Determination of the class frequency.

1. Number of classes : Ordinarily, a frequency distribution should not contain more than 20 to 25 and not less than 6 to 8 classes, depending on the total number of items of the series.If the number of items in asries is large, it can have a large number of class intervals also, because in such cases, all class intervals would have a fairly good frequency. If on the other hand, the number of items is small, the number of classes should also be small as otherwise there would be no frequency in some classes and very small frequency in others. The idea contained in the data can be eaasily and radily grasped when the number of classes is few, but in such a case there is the danger of obscuring some important characteristics of the data.If the number of classes is large, all the charactersitics of the data are contained in them, but on account of too many classes, it becomes difficult to ascertain them. In fact, a balancy should be struck between these two situations.*An ideal number of classes for any frequency distribution would be that which gives the maximum information in the clearest fashion.*

Sturges has given a formula to detrmine the number of classes. According to this formula.

$$k = 1+3.322 \log N$$

Where k is the number of classes and log N is the logarithm of the total number of observations

Thus, if the total number of observation is 100, number of classes would be :

1+(3.322 x log 100)

or 1+ (3.322 x 2)

or 1+6.644 or 7.644

which means that the numbr of classes should be 8, as fractions are always rounded up.

Similarly, if the total number of observations is 1, 000, the number of class would be :

1+(3.322 x 3) or 1+9.966 or 10.966

The number of classes would be 11, after rounding up the fraction.

In this formula, the number of classes cannot be less than 4 and would not generally exceed 20.

2. Magnitude of class intervals : The magnitude of class intervals depends on the range of the data and the number of classes. If the range (difference between the maximum and the minimum values) of the heights of a group of persons is 50 centimetres, and if it is desired to have 10 classes, the magnitude of each class interval would be 50/10 or 5 centimetres. Besides these things, a few other points should also be kept in mind. The magnitude of the class intervals should be such that it does not distrort or obscure the important characteristics of the data. Bearing this fact in mind, the magnitude of the class interval should be 2, 5, 10, 25, 50, 1000, 500, 1,000, 5,000 and so on, rather thatn odd figures like, 1,3,7,11,23,57,92 and 472, etc. The multiples of 2,5 and 10 are in common use and human mind considers them almost as natural magnitudes.

Sturges has also given a formula for determining the magnitude of class intervals. It is :

$$i = \frac{\text{Range}}{1 + 3.322 \log N}$$

where *i* is the magnitude of the class interval and log N is the logarithm of the total number of observations.

Range is the difference between the highest and the lowest value of the variable in question. In other words, magnitude of the class interval would be equal to the **difference between highest and lowest value of the variable divided by the number of classes.**

For example, if the weights of 1,000 students are spread over 100 to 200 ibs. magnitude of the class intervals accoridng to Sturges formula would be :

$$\frac{200-100}{1+3.322\log 1000} \text{ or } \frac{100}{1+3.322\times 2} \text{ or } \frac{100}{10.966}$$

which is equal to 9 approximately

However, in accordance with the principles given earlier, the magnitude of the class intervals even in such a case should be 10.

3. Class Limits : The most important thing that should be kept in mind while choosing the class limits is that these should be chosen in such a manner that the mid-point of a class interval and the actual average of items of that class intervals, should be as close to each other as possible. If it is not so, the class limits would be obscure and distort the main characteristics of the data. Consistent with this point, wherever possible the classs limits should be located at multiples of 2,5,10,100 and such other figures. The class limits must be such that mid-points of class intervals are familiar and common figures ending with 0,2,10,15, etc. These are capable of easy and simple analysis. As far as possible, in frequency distributions there should be no indeterminate class like **under 10 or over 1,000**. Such classification may create difficulties in analysis and interpretation.

There are two ways in which data are classified on the basis of class intervals, viz., **(a) exclusive method, (b) inclusive method.**

(a) Exclusive method : When the data are classified in such a way that the upper limit of a class interval is the lower limit of the succeeding class interval, then it is said to be the exclusive method of classifying data. The following example illustrates the point :

Marks	**Number of students**
30-40	14
40-50	26
50-60	30
60-70	20
70-80	10
	100

Exclusive method of classification ensures continuity of data, as the classintervals are not broken since the upper limit of one class is the lower limit of the succeeding class. In the above example, 14 students are securing marks ranging from 30 to less than 40 and 26 students' marks ranging from 40 to less than 50. A student who is getting exactly 40 marks would be classified in the 40-50 group and not in the 30-40 group.However, this point is not always clear. Therefore, to avoid confusion, the data are sometimes displayed in a slightly different form as given below :

Marks	**Number of students**
30 but lest than 40	14
40 but less than 50	26
50 but less than 60	30
60 but less than 70	20
70 but less than 80	10
	100

(b) ***Inclusive Method :*** Unlike the exlusive method, under the inclusive method, the upper limit fo a class interval is included in the interval itself, and does not go to the next higher classs. The following example will illustrate the point:

Marks	**Number of students**
30-39	14
40-49	26
50-59	30
60-69	20
70-79	10
	100

In the above table, a student getting 39 marks will be indluded in 30 to 39 class interval and, similarly, a student getting 49 marks will be included in 40 to 49 class interval. **Thus, the confusion** which was observed in the exlusive

method is avoided as the upper limit of **a class here is not the lower limit** ofthe next class.

In order to decide whether the exlusive method or the inclusive method shouldbe sued to classify a set of data, we should find, out whether the variable is ofcontinuous type or of discrete type.For example, height is a continuous distribution and can have any value within conceivable limits. The height of aperson can be 160 cms. or 160.01 cms or 160.09 cms. In such cases, it isalwyas desirable to have classification done by exlusive method and we canhave class intervals like 160-162 cms., 162-164 cms and so on.If, however,the data to be classified gives a discrete distribution and relate to a variablelike the number of students in differnt sections of a class, inclusive methodwould be more appropriate as the observations woudl be in whole numbersonly and not in fraction. Class interval in such cases should be 40-44, 45-49,50-54 and so on.

Thus, in a continous series, the statistical unit is capable of division and can be measured in fractions of any size, no matter how small. A tonne of coal can be divided in a 100, 1,000, 10,000 or even more parts.

Theoretically, a ton of coal can be divided into a limitless number of subdivisions. In discrete series statistical unit is either not divisible units where they are not conventionally divided. For example, marks are given to the students in whole numbers. It is possible in such a case to have a discrete series of marks.

If a continous variable is classified according to inclusive method certain adjustmet in the class interval is needed to obtain continuity.Suppose, the data relating to weights of a group of 100 students are classified as follows :

Weight in pounds	**No. of students**
100-109	10
110-119	15
120-129	30
130-139	35
140-149	10
	100

Now, to ensure continuity, we will find out the **difference between upper limit of a class** and the lower limit of the next higher class and divided it by 2. This fraction would be subtracted from **all lower limits** and **added to all upper limits of various classes.** In the above example, the difference between the upper limit of a class and the lower limit of the next class is one and if divided by two, we get a correction factor as 0.5. This will be subtracted from the lower limit of all classes and added to the upper limit of all classes. The modified table would, thus, become continuous and would be as follows :

Weight in pounds	**No.of students**
109.5-119.5	10
10.5-119.5	15
119.5-129.5	30
129.5-139.5	35
139.5-149.5	10
	100

Another example would clarify this method of classification. Suppose, some data are classified by the inclusive method in the following manner :

Variable	**Frequency**
10-19.5	5
20-29.	10
30-39.	10
40-49.5	5

The correction factor would be 20 minus 19.5 divided by 2 or 2.5.The modified classification, thus. would be :

Variable	Frequency
9.75-19.75	5
19.75-29.75	10
29.75-39.75	10
39.75-49.75	5

As has been mentioned earlier, as far as possible, class intervals should be of equal magnitude. If the magnitude of the class interval is unequal, it may give a misleading impression and in such a case comparison of one class with another is not possible.

Further, as far as possible, data should not be classified under classes like less than 100 or more than 500. In such cases, if any statistical analysis is to be done, it would be impossible to arrive at a correct midpoint.More than 500 may mean values upto 1,000 or 10,000 or more and it would not be possible to bave a **precise mid-point of the class interval.**

Simple and Cummulative Series: Statistical series can be either simple or cummulative. In simple series, frequency against each class interval or value is shown separately and individually. In a cummulative series, the frequencies are progressively totalled and aggregates are shown. The following example would clearly sho the difference between simple and cummulative series both discrete as well as continuous.

(a) Example of simple and cumulative discrete series :

Simple discrete series		**Cumulatives discrete series**	
No. of children per couple	**No. of couples**	**No. of children per couple**	**No. of couples**
2	50	Upto 2	50
3	75	Upto 3	125
4	40	Upto 4	165
5	28	Upto 5	193

(b) Example of simple and cumulative continous series:

Simple cummulative series		Cummulativediscreat series	
Height (in cm)	**No. of person**	**Height (in cm)**	**No. of person**
160-162	12	Less than 162	12
162-164	15	Less than 164	27
164-166	24	Less than 166	51
166-168	13	Less than 168	64

8.7 RULES OF CLASSIFICATION

Rules of classificationhave been discussed in the foregoing paragraphs. Given below is a summary of the rules of classification :

(1) The number of classes should not be excessive :

They may range between 5 to 20 thogh there is no hard and fast rule in this respect. The actual number of classes would depend in the spread of the values of the variable under study and the total number of observations.

(2) Magnitude of the class intervals should be as far as possible in multiples of 5, like 10, 15, 20 etc. Values of the magnitude of the class interval like 3,5,7, etc. should be avoided.

(3) The lowest value of a variable should not necessarily be the lower value of the first class interval. If, for example, lowest value of a variable is 11, the first class interval need not start from 11. It may start from 10 or 5 or even zero.

(4) Class intervals shoul, as far as possible, be of equal size.

(5) Values of the variable under study or the class interval values should be written on the left-hand side, and opposite each class interval, the respective frequencies should be written on the right-hand side of the table.

Some examples of classification are given below:

Example 1. Prepare a statistical table from the following weekly wages of 100 workers (in Rs.) of Factory A :

88	23	27	28	86	96	94	93	86	99
82	24	24	55	83	99	95	86	82	36
96	39	26	54	87	100	56	84	83	86
102	48	27	26	29	100	59	83	84	48
104	46	30	27	40	101	60	89	46	49
106	33	36	30	40	103	70	90	49	50
104	36	37	40	46	106	72	94	50	60
24	39	49	46	66	107	76	96	46	67
26	78	50	44	43	46	79	99	36	68
29	67	56	99	93	48	80	102	32	51

Solution : The lowest value is 23 and the highest 106. The difference in the highest and the lowest value is 83. If we take a class interval of 10, nine classes would be formed. The first class should be taken as 20-30 instead of 23-33 as per the principles of classificaton.

Frequency distribution of the wages of 100 workers

Wages (Rs.)	**Tally bars**	**Frequency**
20-30	~~𝍸 𝍸~~ III	13
30-40	~~𝍸 𝍸~~ I	11
40-50	~~𝍸 𝍸 𝍸~~ III	18
50-60	~~𝍸 𝍸~~	10
60-70	~~𝍸~~ I	6
70-80	~~𝍸~~	5
80-90	~~𝍸 𝍸 𝍸~~	14
90-100	~~𝍸 𝍸~~ II	12
100-110	~~𝍸 𝍸~~ I	11
		Total 100

Example 2. Present the following data of the percentage marks of 60 students in the form of frequency table with ten classes of equal width one class being 40-49 :

41	17	33	63	54	92	60	58	70	06	67	82
33	44	57	49	34	13	54	63	36	52	32	75
60	33	09	79	28	30	42	93	43	80	03	32
57	67	24	64	63	11	35	82	10	23	00	41
60	32	72	53	92	88	62	55	60	33	40	57

Solution : Lowest value is 0; therefore, the first class interval would be 0-9. Frequency distrisbution of the marks of 60 students are :

Marks	**Tally bars**	**Frequency**
0-9	IIII	4
10-19	III	3
20-29	III	3
30-39	~~IIII~~ ~~IIII~~	10
40-49	~~IIII~~ II	7
50-59	~~IIII~~ ~~IIII~~	9
60-69	~~IIII~~ ~~IIII~~	11
70-79	~~IIII~~	5
80-89	~~IIII~~	5
90-99	III	3
		Total : 60

Example 3 : Using Struges rules determine the number of classes in which the following data should be divided. Classify the data in equal class intervals.

178	114	110	161	164	155	175	157	108	128
62	30	165	195	87	71	133	151	94	42
141	104	130	167	116	164	156	124	146	149
143	93	103	162	62	79	150	149	113	121
148	20	140	187	40	197	144	184	87	122

Solution. Number of classes or k would be

$k = 1 + 3.322 \log N$

$N = 50$ So $k = 1 + (3.322 \times 1.6990)$

or $1+5.63 = 6.63$ or 7

The range of the data is 204 - 30 or 174. If it is divided by the number or classes (174/7) the magnitude of the class interval would be 25 approximately.

The lowest class be 30-55 and the next 5.3-80 and so on.

The data would be classified as follows :

Class interval	**Tally bar**	**Frequency**
30-55	III	3
55-80	IIII	4
80-105	~~𝍸~~ I	6
105-130	~~𝍸~~ IIII	9
130-155	~~𝍸~~ ~~𝍸~~ II	12
155-180	~~𝍸~~ ~~𝍸~~ I	11
180-205	~~𝍸~~	5

Example 4 : Form a frequency distribution from the following data by Inclusive Method taking 4 as the magnitude of class intervals :

10, 17, 15, 22, 11, 16, 19, 24, 29, 18,

25, 26, 32, 14, 17, 20, 23, 27, 30, 12,

15, 18, 24, 36, 18 15, 21, 28, 33, 38,

34, 13, 10, 16, 20, 22, 29, 29, 23, 31,

Solution : Since the minimum value of the variable is 10 which is a very convenient figure for taking the lower limit of the first class and the magnitude of the class intervals is given to be 4, the classes for preparing frequency distibution by the "Inclusive Method' will be 10-13, 14-17, 18-21, 22-25,...., 34-37, 38-41, the last class being 38-41, because the maximum value in the distribution is 38. The final frequency distribution along with the tally marks is as follows :

Class interval	**Tally bar**	**Frequency**
10-13	𝍸	5
14-17	𝍸 III	8
18-21	𝍸 III	8
22-25	𝍸 II	7
26-29	𝍸	5
30-33	𝍸	4
34-37	II	2
38-41	I	1
Total :		40

Example 5 : If the class mid points in a frequency distribution of age of a group of persons are 25, 32, 39, 46, 53 and 60 find : (i) the size of the class interval, (ii) the class boundaries, and (iii) the class limits, assuming that the age quoted is the age completed on last birthday.

Solution :

(i) The size of the class interval

= Difference between the mid-value of any two consecutive classes

= 7 (Since 32 - 25 - 39 - 32 = ... = 60 - 53 = 7)

(ii) Since the magnitude of the class is 7 and the mid-values of the : classes are 25, 32,.....60, the corresponding class boundaries for different classes are obtained by adding (for upper class boundaries) and subtracting (for lower class boundaries) half the magnitude of the class interval, viz., 7/2=3.5 from the mid-values respectively. For example, the class boundaries for the first class will be (25 - 3.5, 25 + 3.5) i.e., (21.5, 28.5); for the second class will be (32 - 3.5, 32 + 3.5) i.e., (28.5, 35.5) and so on. Thus, the various classes (exclusive type) with class boundaries are as given in the following table :

Class	**Mid-Value**
21.5-28.5	25
28.5-35.5	32
35.5-42.5	39
42.5-49.5	46
49.5-56.5	53
56.5-63.5	60

(iii) Assuming that the age quoted (X) is the age completed on lastbirthday, then X will be a discrete variable which can take only integrat values. Hence, the given distribution can be expressed in an 'inclusive type' of classes with class interval of magnitude 7, as given in the following table ;

Age (on last birthday)	Mid-Value

22-28	25
29-35	32
36-42	39
43-49	46
50-56	53
56-63	60

8.8 CONCLUSION

The process of arranging data in groups or classes according to resemblance and similarities is technically called classification. It is a preliminary to tabulation and it prepares the ground for proper presentation of statistical facts.

8.9 FURTHER STUDY

1. Agarwal, B.N., Elhance, D.N. &Elhance, Veena : Fundamental of Statistics
2. Anand, Vinod & Chand, Mahesh : Economic Theory–A Mathematical Approach
3. Margo, G.S. : Elementary Statistics
4. Gupta, S.B. : Principal of Statistics
5. Sinha, V.C. : Principal of Statistics

Uttar Pradesh Rajarshi Tandon
Open University

Bachelor of Business Administration

BBA-107

Business Statistics

BLOCK

3

PRESENTATION OF DATA

UNIT-9 **141**

Tabular Presentation

UNIT-10 **149**

Diagrammatic Presentation

UNIT-11 **167**

Graphical Presentation

UNIT-12 **179**

Ratio, Percentae and Rates

Curriculum Design Committee

Prof. Omji Gupta **Director**

School of Management Studies, UPRTOU, Prayagraj

Dr Gyan Prakash Yadav

Assistant Professor

School of Management Studies, UPRTOU, Prayagraj

Dr Devesh Ranjan Tripathi

Assistant Professor

School of Management Studies, UPRTOU, Prayagraj

Course Preparation Committee

Dr T.M. Raini **Author**

Associate professor, Firoj Gandhi PGCollege, Raibarely

Dr Shiv Bushan Gupta

Associate professor, MMM PG college, Kalakankar, Pratapgarh

Dr Gaurav Sankalp

Academic consultant, School of Management Studies, UPRTOU, Prayagraj

Dr Ajay Tiwari

Assistant Professor

MONIRBA, Department of Commerce, University of Allahabad

Prof J.N. Mishra **Editor**

Professor, MONIRBA, Department of Commerce, University of Allahabad

Dr Gyan Prakash Yadav **Coordinator**

Assistant Professor

School of Management Studies, UPRTOU, Prayagraj

©UPRTOU, Prayagraj-2021

ISBN :

©All Rights are reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the **Uttar Pradesh Rajarshi Tondon Open University, Prayagraj.**

BLOCK INTRODUCTION

In the present scenario data presentation is looked as a vital event. Today all researched are based on data. The simpler one presents the data, the easier its is for reader to understand. The present block name presentation of data comprise of four unit.

Unit-9 Discusses about tabulation of data, its objectives, difference between tabulation and classification

Unit-10 Deads with diagramtie representation of data, its principal and various type of diagrams

Unit-11 Deals with graphical presentation of data with the help of various charts.

Unit-12 Deals with ratio, percentage and rates, how compute and its application.

UNIT-09 TABULAR PRESENTION

9.1 Objective

9.2 Meaning of Tabular Presentation

9.3 Objective of Tabular

9.4 Difference between classification and tabulation

9.5 Kinds of tabulation

9.5.1 On the basis of purpose

9.5.2 On the basis of Origin

9.5.3 On the basis of Construction

9.6 Questions

9.7 Summary

9.1 OBJECTIVE

After studying this unit you should understand about meaning and objective of tabular presentation.

- Difference between classification and tabulation.
- Kind of tabulation i.e. on the basis of purpose, origin and construction.

9.2 MEANING OF TABULATION

Tabulation may be defined as the orderly and systematic record of a numerical data in rows and columns. In tabulation it's very important to have proper presentation of data in appropriate manner because proper presentation of data is helpful for clear understand, comparison and drawing conclusion.

When a numerical data are collected for analysis they must be organized or grouped to facilitate the analysis. Tabulation is used for the purpose of analyzing some of characteristic of the data, such as the largest and the smallest value, and the values which appear most often and so on.

Tabulation is simple and easy method, because here we use the table for salutation of any data.

9.3 OBJECTIVE OF TABULATION

Main objective of tabulation are described as follows:-

1. **Methodical Presentation:-** Main objective of Tabulation is to provide the information in particular serial and in a orderly manner.

2. **To help in Reference:-** A tabulation data are provide reference of data, so it's easily indentified the result.

3. **Easily analysis and comparison:-** With the help's of tabulation can easily analyze and compare one data from another, so it's useful for analysis and comparison of data.

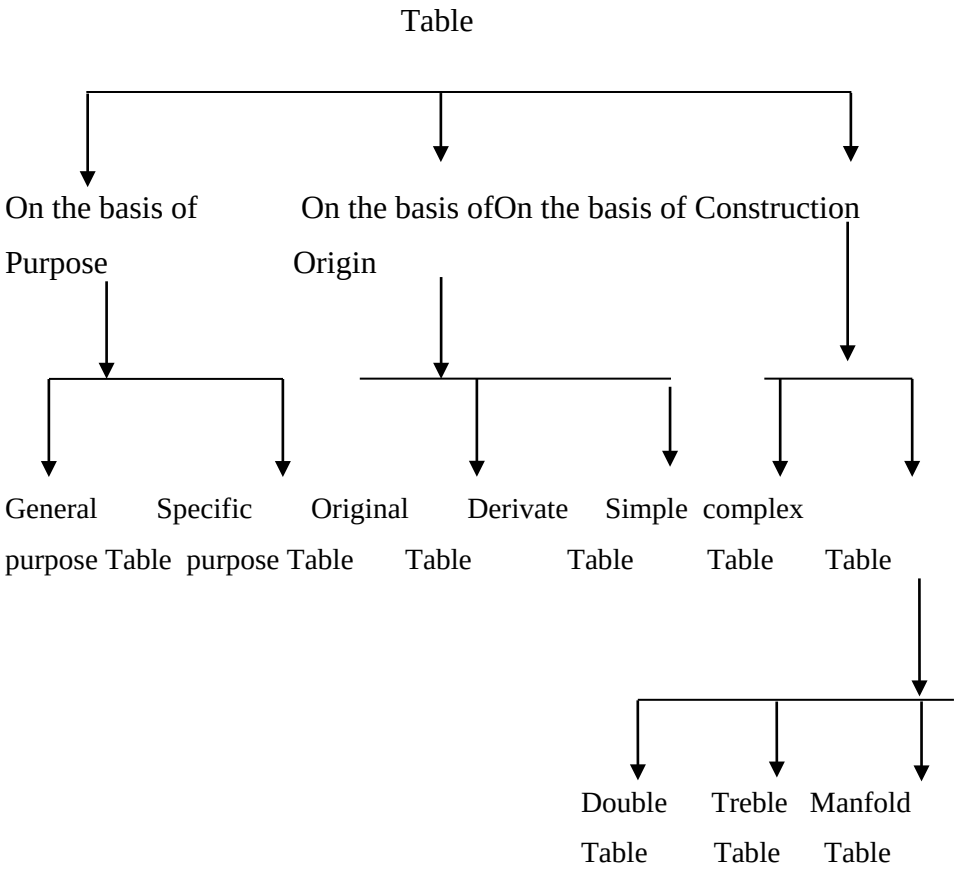
4. **Classify the Characteristic of data:-** It's clear a Characteristic and nature of data's significant feature. We can easily find the nature of any particular data.

5. **Detect error and omission in the data:-** A tabulation can easily detect the error and omission in the grouped, and correct the figure or data very easily.

9.4 DIFFERENCE BETWEEN CLASSIFICATION OF DATA AND TABULATION

Classification	Tabulation
• Classification is a process of arranging the data in a group or in a class. • Classifications is not based upon tabulation. • In classification a data are classified. • Classification is done on the basis of fact or on the basis of feature. • In Classification data are divided in a separate Class. • In Classification numerical data are not represented in row and columns.	• Tabulation is process of presentation of data in simple, clear, and orderly manner • Tabulation is based upon Classification. • In Tabulation firstly data are classified then presented in table. • Tabulation is done on the basis of classification. • In Tabulation a data are not divided in separate class. • In Tabulation numerical data are representation row and columns.

9.5 KIND OF TABULATION



9.5.1 ON THE BASIS OF PURPOSE

(i) General purpose or reference Table:

General purpose are also called repository table. General purpose table are very easy and simple method. With the help general purpose we can also make a summary or derivative table. General purpose table presents the basic row material.

(ii) Specific purpose table :

This table is also called summary table, test table and analytical table. Specific purpose table are maintain for some specific

purpose and event, with the help of general table. Specific table proforma are very small, and this table is use for median, percentage and ratio etc.

9.5.2 ON THE BASIS OF ORIGIN

(i) Original table: The table is also called as primary table or classification table. In original table a data is represented in original form or natural nature.

(ii) Derivation table: In derivation table data are presented such as- table, percentage, ratio, coefficient etc.

9.5.3 SIMPLE TABLE

A simple table is also called a "single attribute table" or "Table of the first order." In simple table here we present only way.

Example: Distribution of population of population in different age.

No. of student in different department in Lucknow University.

Department	No. of student
Art	25
Commerce	75
Agriculture	30
Science	80

(ii) Complex table: In complex table here we present more then one quality or specialty. Complex table are three types:-

(a) Double or two way table

(b) Treable or three – way table

(c) Manifold table

(a) Double or two way table: In this table we include two quality and these quality are related to each other.

Example:-

Distribution of population according to sex.

State / Population		Male	Female	Total
U.P.	50,000	30,000	20,000	50,000
M.P.	40,000	30,000	10,000	40,000
Bihar	45,000	25,000	20,000	45,000
Total		85,000	50,000	135,000

(b) Treable or three-way table : In three-way table we present three types of quality or attributers
Example :

Course	No. of student					
	Male		Females		Total	
	Resident	Non-resident	Resident	Non-resident	Resident	Non - resident
D.A.B						
B.Com						
M.B.A.						
M.Com.						
Total						

(c) Manifold table: In manifold table we include many attribution and qualities.

Course	Total No.of Income																	
	Male						Female						Total					
	Resident			Non-Resident			Resident			Non-Resident			Resident			Non-resident		
	Married	Unmarried	Total	Married	Unmarried	Total	Married	Unmarried	Total	Married	Unmarried	Total	Married	Unmarried	Total	Married	Unmarried	Total
Jan.																		
Feb.																		
Mar.																		
Apr.																		
May																		
Jun.																		
July.																		
Aug.																		
Sept.																		
Oct.																		
Nov.																		
Dec.																		

Main Parts of a table :

Form a data in row and columns. A table has following parts:-

1. **Table Number :** In table firstly we write the number of table. Its helpful for giving a particular number to any matter. Table number are written on top of the table.

2. **Title :** Title is a writers of the table and is given after its number. A title is written in bold and capital letter. The objective of title are prepared to find subject matter of table.

3. **Captions and stubs** : In table a vertical columns is called " Captions" and titles of horizontal. Column is known as "stubs".

4. **Ruling and Spacing :** In time of preparing table it's compulsory to follow the rules and regulation. To make table look attractive. Attractive of table depends to a large extent on proper spacing and ruling. A format of table are prepared before posting the data.

5. **Main body of a table :** A body of table is its main part. In this feature include row and column. In total the data are presented from top to bottom under captions and from left to right under stubs.

6. **Footnotes :** Footnotes is a remark a table when any important point are not written , include in table, then it's written in footnotes.

   A foot notes is always written below the table.

7. **Source :** A source is written after footnotes. The source represent the quality of table.

Structure of table

Table No…….

Title…….

Headline (if any) ……..

	Columns		
	Caption	Caption	Total
Rows	STUB	Body	
	STUB		
	Total		

Footnote: ……

Sources: …….

9.6 QUESTIONS

Theory Question:

Question : What is tabulation? What are it's users? Mention the items that a good statistical table should contain.

Question : Explain the kinds' of Tabulation?

Question : Explain the main part of a table and write the objective of tabulation?

Question : Write the difference between classification and tabulation?

9.7 SUMMARY

Orderly and systematic record of data is known as tabulation methodical presentation, to help in references, easy analysis and comparison are some of objective of tabulation. Difference between classification and tabulation.

Kinds of tabulation on the basis of purpose i.e. general purpose, specific purpose. On the basis of origin i.e, original table and derivative table and on the basis of construction simple construction, complex construction.

UNIT-10 DIGRAMMATIC PRESENATION

10.1 Objective

10.2 Introductions

10.3 Advantages of diagrammatic presentation

10.4 General rule for constructing diagram

10.5 Types of Diagram

10.6 One Dimension Diagram

10.7 Two Dimension Diagram

10.8 Three Dimension Diagram

10.9 Question

10.10 Summary

10.1 OBJECTIVE

After studying this unit one should understand about diagrammatic presentation its advantages, rules and types.

10.2 INTRODUCTION

In statistics it is very important to represent the data in a very easy and simplest from, and diagrammatic presentation is the best example for this.

In diagrammatic presentation the data are presented in form of diagram. This diagram are represented by figure, bar, circle, rectangle, and square etc. A general person who have no knowledge of statics he can easily understand statics by diagram. A diagram is plated in paper and can be explain easily any hardest statics data. A person can easily understand numerical data by the help of diagram.

10.3 ADVANTAGES OF DIAGRAMMATIC PRESENTATION

A diagrammatic presentation has many advantages:-

(1) Attractive and Impressive: A diagram look's beautiful as a diagram attracts person who is not interested in statics or think that statics is very hard. In present time we use many color for making a diagram, so that it's look very alterative and impressive.

(2) Simple and intelligible presentation: A diagram is very simple and intelligible as a diagram makes data simple and intelligible which are hard large and cult to understand. It's a human nature that we get's any data very easily with the help's of diagram and a diagram retains in one's mind for long a time.

(3) Facility in comparison: Comparison in to numerical data, diagram is easy for caparisoning the many year data.

(4) Informative Entertainment: A diagrammatic presentation entertains the reader. With the help's of diagram information can easily be done.

(5) Successful media of communication: A diagram when looks alterative and effective therefore it's a medium of exchange & communication. A company and firm can also use this diagram method for presenting the annual report.

(6) Universal applicability: A diagram method is not use only for representing statics numerical data. But it's also useful for economic and sciences field.

10.4 GENERAL RULE FOR CONSTRUCTING DIAGRAM

General rule for constructing diagram are follows:-

(1) Title: In time of making diagram it's compulsory to give a title and the title are very short and easy. We can write title on top or under the diagram.

(2) Proper size: Diagram is plotted in particular space. Here there is no rule for making a diagram in particular space, but a perfect ratio of length and breadth make a diagram effective.

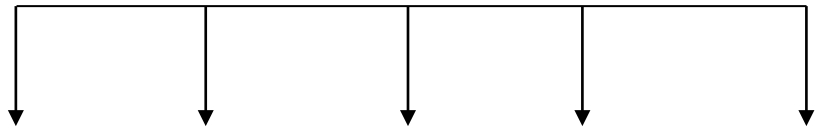
(3) Proper Size: Proper scale is very important to make a perfect diagram. When a scale of diagram looks same and a person may be totally confused.

(4) Arithmetic and Neatness: A proper shape, writing and clarity makes a diagram looks very attractive and neat here we use the color, shades and line for alternative diagram.

(5) Simplicity: A diagram should look very simple, it should be draw in mammy that it should be easily understand.

(6) Other rules: (i) Diagram am always plated left to right. (ii) When comparison of two or more diagram is done it should be in the same time (iii) there must be a draw line outside the diagram.

10.5 TYPES OF DIAGRAM



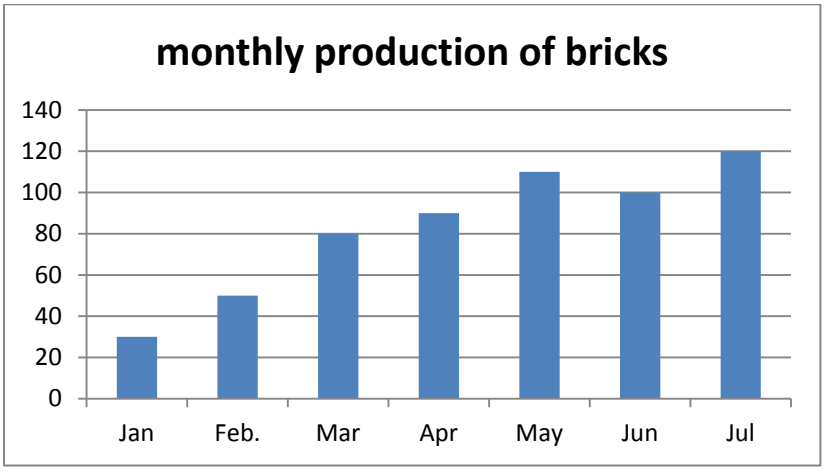
One Dimensional Diagram's	Two Dimensional Diagram's	Three Dimensional Diagram's	Pictogram	Maps and layouts
➢ Simple bar diagram ➢ Sun-divided bar diagram ➢ Multi bar diagram ➢ Sub –divided bar ➢ Diagram on percentage basis	➢ Rectangle ➢ Square ➢ Circula	➢ Cubes ➢ Cylinder, Cone and		

10.6 ONE DIMENSION DIAGRAM

(i) Simple bar Diagram: A simple bar diagram are use for comparison study of two or more than two variable. For example, it's use for finding number of labor, sales, capital etc. This diagram plated vertical or horizontal, but vertical bar diagram many useful and very easy.

Example: A monthly production of bricks in a factory are as follows :

Month	Jan	Feb.	Mar.	April.	May.	June.	July
Production	30	50	80	90	110	100	120



(ii) Sub-divided bar diagram: A sub-divided bar diagram are also known's as "Complete bar diagrams. The use of data and present the part, division and components of the data. For example: - the total no. of student in different department for different year uses the sub-divided bar diagrams.

In sub-divided we use many different color and different design. A main demerit's of this diagram's is it's difficult for comparison because here department and components are not in some basis.

Example:

Year			Men	Women	Children	Total
1985	2,00,000	1,50,000	4,00,000	7,50,000		
1995	5,00,000	4,00,000	3,00,000	12,00,000		
2005	8,00,000	6,00,000	5,00,000	19,00,000		

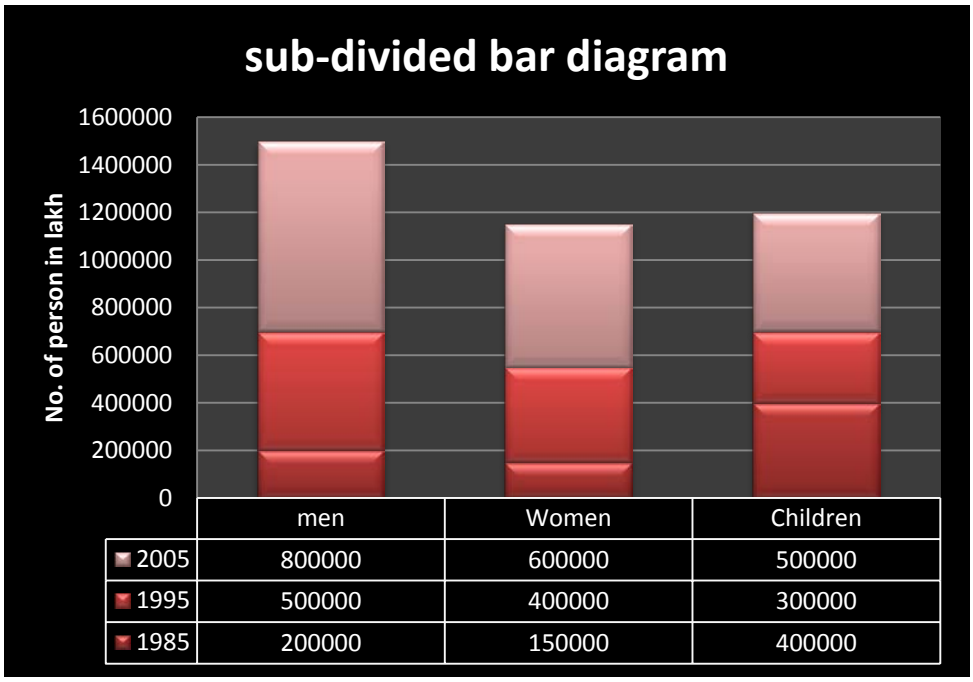
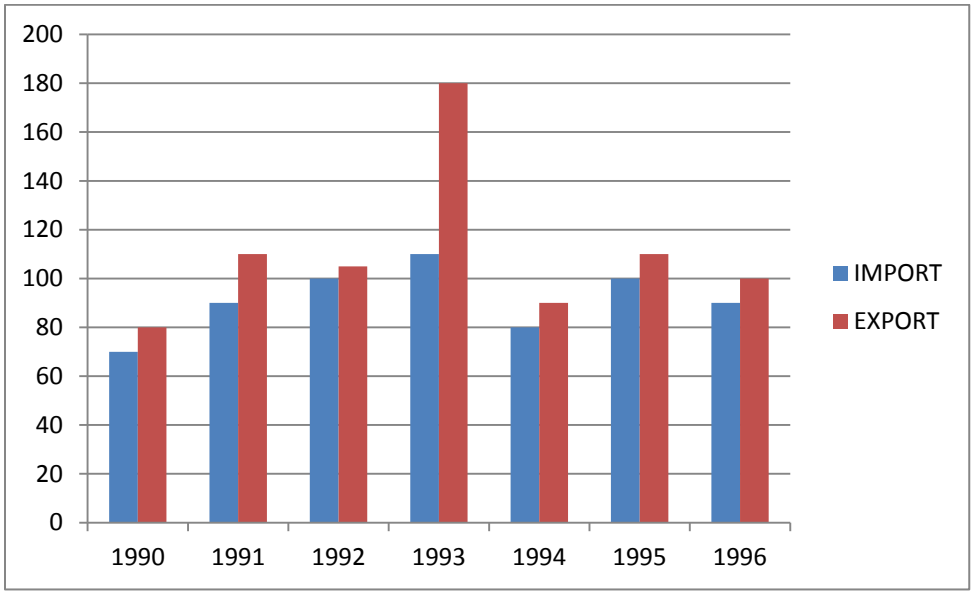


Diagram showing no. of person (men, women and children)

(iii) Multiple bar diagram: A multiple bar diagram are also known's a compound bar diagrams. This bar diagram is use when had to comparison the two or more than two element's and use the different color and shades are used for different data. For example: - Show the important and export of a country in different year.

In this example we take two different data (import export) in same year.

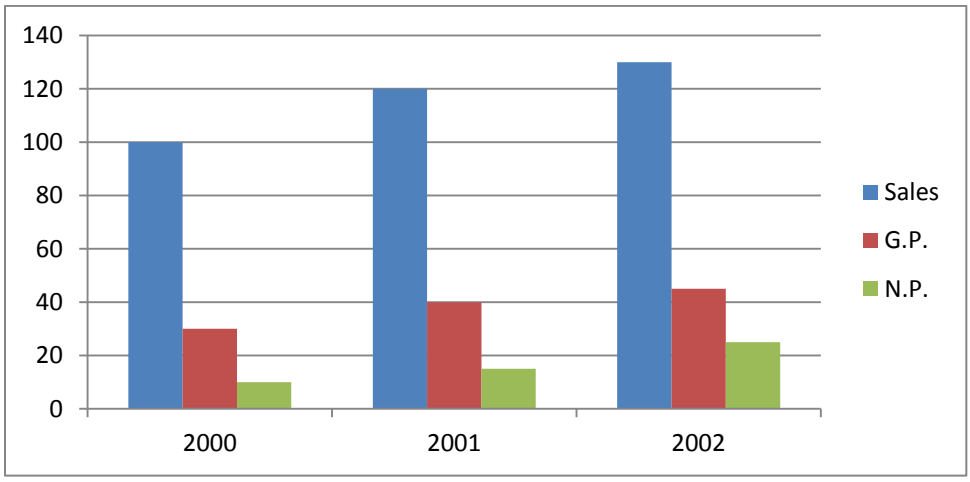
DIAGRAM SHOWING IMPORT AND EXPORT (1990-96)



Year	1990	1991	1992	1993	1994	1995	1996
Import	70	90	100	110	80	100	90
Export	80	110	105	180	90	110	100

Example: Rs. in thousands.

Year	2000	2001	2002
Sales	100	120	130
G.P.	30	40	45
N.P.	10	15	25



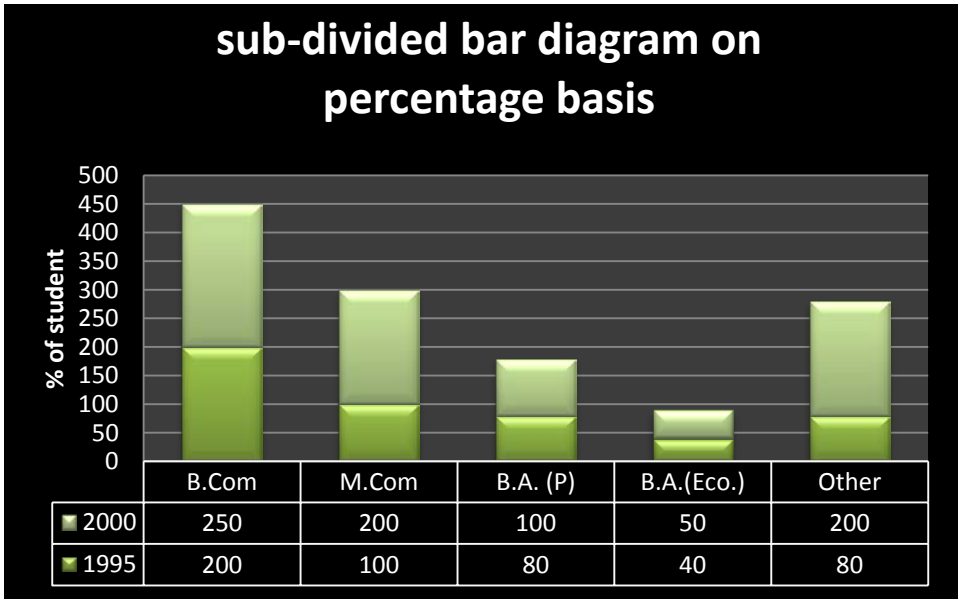
(iv) **Sub-divided bar diagram on percentage basis:** A Sub – divided bar diagram on percentage basis is use to denote the data in percentage basis, and it's a very helpful for comparing the figure in one year basis are not very simple to comparison with other diagram we firstly convert the data in percentage and then plot the diagram.

Example :

Year	B.Com(H)	B.Com(0)	B.Com(P)	B.Com (Eco.)	Other	Total
1995	200	100	80	40	80	500
2000	250	200	100	50	200	800

Solution:

	1995		2000	
Courses	No. of student	%	No. of student	%
B.Com(H)	200	40	250	31
B.Com(0)	100	20	200	25
B.Com(P)	80	16	100	13
B.Com (Eco.)	40	8	50	6
Other	80	16	200	25
Total	500	100	800	100



10.7 TWO DIMENSIONAL BAR DIAGRAMS

Two dimensional bar diagrams are also known as "surface diagrams' or 'area diagram'. In two dimensional bar diagrams length and breadth are based equally, so size of the observation are represented by the area of the diagram.

A two dimensional bar diagram are of four types:-

(i) Rectangular diagram:

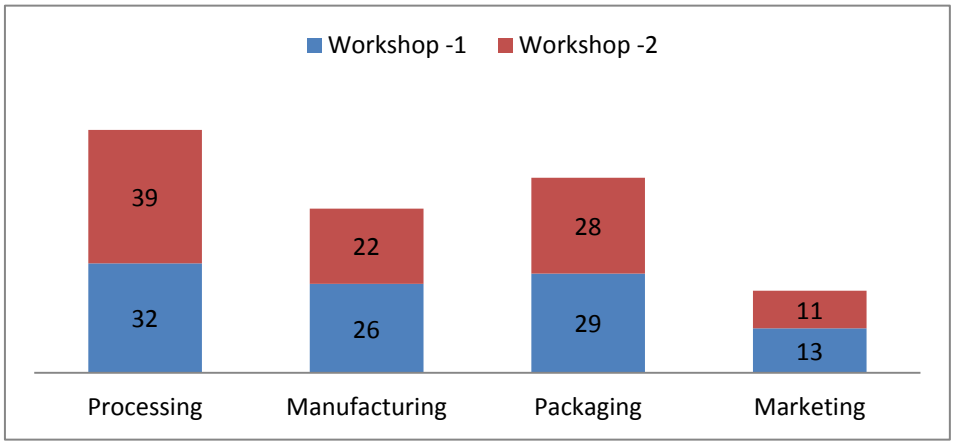
A rectangular diagram is similarly too sub-divided bar diagram on percentage basis. It is based on principal of area. Here we present the two different quality of one data in same time, so length and width are equally importance in case of two dimensional diagrams.

Example:

Department	Workshop -1	Workshop-2	
Processing		100	70
Manufacturing		80	50
Packaging		90	40
Marketing		40	20
Total		310	180

Solution:

Department	Workshop -1	%	Workshop-2	%
Processing	100	32	70	39
Manufacturing	80	26	50	22
Packaging	90	29	40	28
Marketing	40	13	20	11
Total	310	100	180	100



(ii) Square diagram: - when the values of range are so large then, it's not easy to make bar diagram. Suppose income of x in first year Rs. 500 and income second year Rs. 5000. In this case one diagram is very small and second diagrams in plated very large.

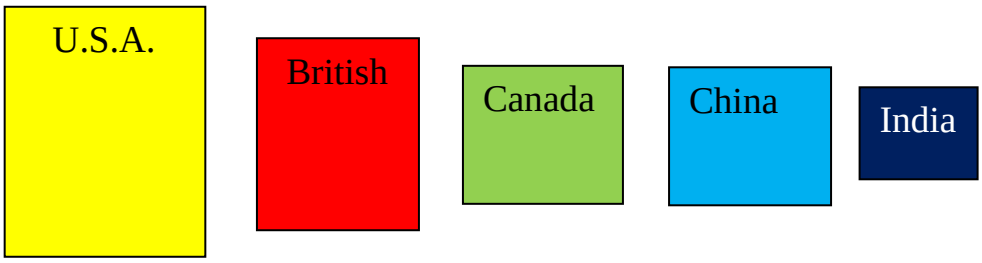
In this case a square diagrams very suitable and useful, in square diagram here large no. area comes to short number. For example:- a income x of first year Rs. 2500 and second year Rs. 25000 then first square the income of both year 50 and 500 and divided the same number and changing the minimum value. So here we divided by 10 then it comes 5 and 50 then we divided by 5 then it comes 1and 10 so an income have come 1 and 10, it's very easy to plotted a diagram.

Example:

Pre Capital income (assumed) of five countries:

Country	U.S.A.	British	Canada	China	India
Per-capital income (Rs.)	10,000	8100	6400	1500	160

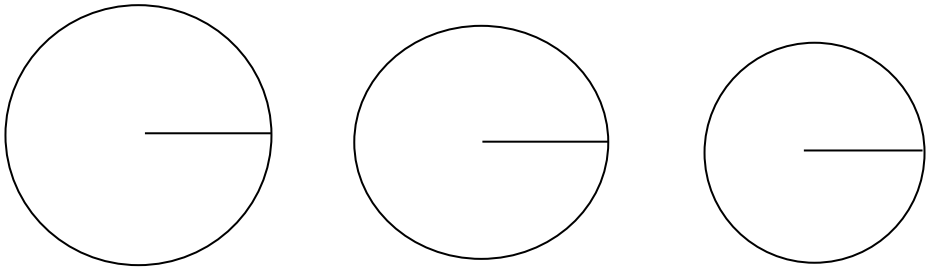
Country	Per-capital income (Rs.)	Square	Line of square (divided by w)
U.S.A.	10,000	100	2.5
British	8100	90	2.25
Canada	6400	80	2
China	1500	50	1.25
India	160	40	1



(iii) Circular Diagram: A Circular diagram used where a square diagram is not used. In circular diagram they area of circle is

calculated, for this formula πr^2, here the value of '∇' is 22/7 and 'r' is the radius of circle. First circle color should be totally different form second circle.

Example:



Income Tax Import duty Production duty

(iv) Pie-Diagram: In Pie diagram firstly we converted the data in percentage. A total area of circle is 360^0 and it's assumed whole to be equal to 360^0. In pie diagram whole area represent the aggregate and circle is divided into value too many parts. In pie diagram we divided the valued of total angle 360^0 and find the valued of one angle ($360^0/100$) = 3.6. If two more sets of data are to be proper tonal to the ratios square roots of their 'whole' or magnitude. When circle are plotted a radius is drawn touches the circumference and it will represented in proportion of first component:

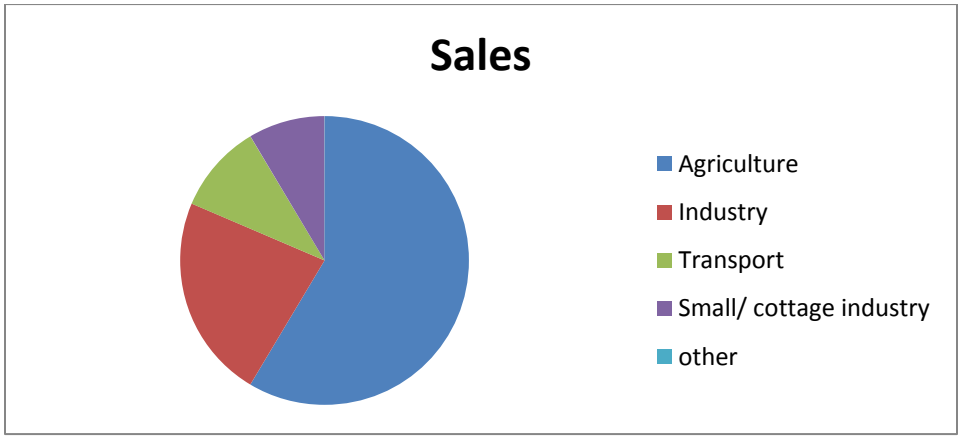
Example: Following tables gives expenditure incurred in various sectors of economy in the year 2001-2002.

Sector	Rs. in cores
Agriculture	200
Industry	180
Transport	150
Small/ cottage industry	100
other	80

Represent the figure through pie diagram.

Solution: Percentage are converted into degree by multiplying in with 3.6^0, 10% of 360^0.

Sector		% of total	Angles at centre
Agriculture	200	28	100
Industry	180	25	90
Transport	150	21	76
Small/ cottage industry	100	15	54
other	80	11	40
Total	710		360



10.8 THREE DIMENSIONAL DIAGRAMS

In three dimensional diagrams we take length, breath and height and are expressed in diagram. The three dimensional diagramed drawn are cubes, cylinder, blocks or spheres etc.

Three dimensional diagrams are also known as 'cubic diagram". When there is a lot of different in item then we use three dimensional diagrams.

(i) Cubes:

Cubes diagram are also called volume diagram. Cubes are three dimensional diagrams and here we take length, breadth and height. Supposed you have a ratio of data 1:125, and then firstly cubes the ratio and get smallest number 1:5 is the smallest ratio of 1:125 and it's easy for plotted.

Here are some steps for construction for cubes diagrams.

a. First take a square. ABCD

b. A perpendicular line PQ bisecting AB at –mid point is drawn is a manner that AB goes through the middle of PQ.

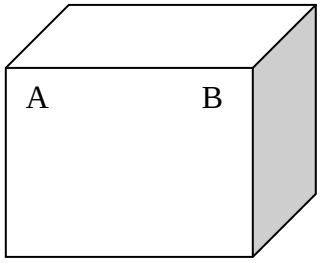
c. AP and CQ are joined.

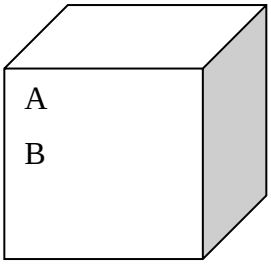
d. A line BR parallel and equal in length to AD is drawn from point B.

e. From R, a line Rs. parallel and equal to PQ is drawn.

f. DS and QS are joined.

g. To get a real cube, the line CQ PQ and QS are erased and we will get a clear cube CDSRPA.





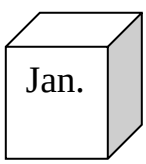
Example: The following table gives the amount spent by marketing division of an enterprise on advertisement for the first four month of a Year:

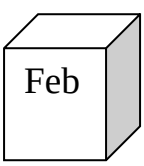
Month	Jan.	Feb.	Mar.	April
Rs. 000	60	100	250	400

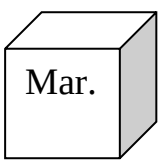
Solution: For drawing the above figures in cubes, firstly the cubic roots derived. The sides of cubes should be in proportion to the ratio of cubic roots. The cubic roots sometimes are divided by a common number to minimize to proportion.

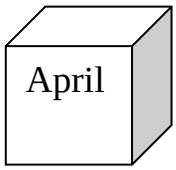
Calculation:

Month	Rs. 000	Cubic roots	Ratio of side of cube
Jan.	60	3.92	1.306
Feb.	100	4.46	1.546
Mar.	250	6.30	2.1
April	400	7.37	2.456



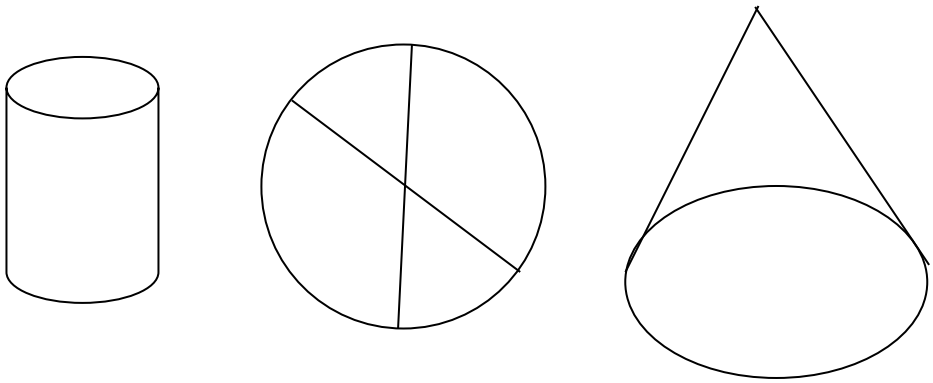






(ii) Cylinder, cone and sphere: - Cylinder is a solid body with straight sides and a circular section. Cone is also a solid figure with a circular plan base. Sphere is also a solid figure with every point.

They are also other form of three dimensional diagrams but it is not used in presentation of data in statistical analysis.



Pictograph :

A pictograph is generally used for presentation of data. A pictograph is very simple and easily understood. Pictograph is generally used by government institutions like social welfare department, health department, education department and much non-government social organization.

Example: Following tables show the number of fair price shops in a locality for 1990, 2000 and 2020.

Year	1990	2000	2010
No. of shops	7	12	16

1990	☺☺ ☺ ☺ ☺ ☺ ☺
2000	☺ ☺ ☺ ☺ ☺ ☺ ☺ ☺ ☺ ☺ ☺ ☺
2010	☺ ☺ ☺ ☺ ☺

(e) Maps and Layout: Maps and Layout's are also known as cartogram. They are useful for illiterates. It's use for quantitative data on geographical basis. Their scope is limited and suited in geographical studies of a phenomenon.

10.9 QUESTIONS [LINE DIAGRAM]

Question: The age of so students of a class is given below. Represent by line diagram.

Name	A,B,C,D,E,F,G,H,I,J
Age in year	20,19,22,17,25,18,21,26,27,16

Question: Draw a line diagram from the following rainfall data received in different years at Delhi.

Year	2000,2001,2002,2003,2004
Rainfall	418,621,520,444,771

Simple Bar Diagrams:

Question: Depict the following data of India's population by simple bar diagram:

Year	1998, 1999, 2000,2001,2002
Imports (Rs.)	308,312,321,300,306

Question: Present the following data of India's population by simple bar diagram :

Year	1941, 1951, 1961, 1971, 1981, 1991 2001
Imports (Rs.)	31.8,22.6,38.1,46.7,66.2,84.1,100.7

Question: The following table present birth rate per thousands for a certain period present these data by simple bar Diagram:

Country	Birth rate
India	38
Germany	18
England	80
China	20
New Zeeland	80
Sweden	45

Question: Draw Horizontal bar diagram form the following data of the distance covered by six bus routes in one day:

Route No.	A,B,C,D,E,F,G,H,I,J
Distance covered in Kms.	50,60,70,80,90,100

*Multiple Bar Diagram:

Question: The following table shows the results of M.A. (Economics) of a university represent the same by multiple bar diagram.

Result	No. of students			
	2000	2001	2002	2003
I Division	10	15	17	20
II Division	18	26	28	18
III Division	30	48	45	35
Fail	20	32	20	32

Question:He table given shows the results of B.A. student of Lucknow University represent the same by multiple bar diagram :

Result	No. of students in different years		
	2001	2002	2003
I Division	10	12	21
II Division	21	34	29
III Division	15	18	36
Fail	24	28	42

Question (a): Write the uses and importance of diagram representation?

(b) Representation the following data by multiple bar diagram.

Facilities	No. of students in different years	
	2000-01	2002- 03
Art	222	500
Science	310	480
Education	120	160
Commerce	200	340

Question : The data below give the yearly profits (in thousands of rupees) of two companies A and B. Represent the data by means of a suitable diagram.

Year	In thousands of rupees	
	Company (a)	Company (b)
1997-98	140	95
1998-99	180	100
1999-00	160	104
2000-01	185	132
2001-02	170	150

Question: by multiple bar diagram the number of students is four facilities in a college three years.

Facilities	No. of students in different years	
	2000-01	2002- 03
Art	222	500
Science	310	480
Education	120	160
Commerce	200	340

Question: Show the following data by a sub- divided bar diagram.

College	No. of students in different years				
A	Art	Science	Commerce	Agriculture	Total
B	1,400	640	800	720	3,560
	625	310	480	375	1,180

Question: Represent the following data by a sub-divided bar diagram.

Year	I Div.	II Div.	III Div.	Fail
2000	55	133	280	110
2001	60	160	310	85
2002	53	220	420	153

Question: Show the following data by a sub- divided bar diagram.

College	Art	Commerce	Science	Total
X	1,500	435	345	2,100
Y	1,300	650	666	2,500
Z	1,130	845	721	3,000

Question: Represent the following data by a sub-divided (in corer rupees) bar diagram.

Year	Import	Export	Balance
2000	680	602	-78
2001	642	640	-2
2002	663	701	+28
2003	720	590	-130

Question: Represent the following data by a sub-divided (in corer rupees) bar diagram.

Year	Public Co.	Private Co.	Total
1997	340	2100	2440
1998	400	1650	2050
1999	420	2300	2720
2000	545	2045	2590
2001	614	1480	2094

Question : Represent the following data diagrammatically:

Particulars	2000(Rs.)	2003(Rs.)
Sales price per heater	45	57
Cost per heater		
Wages	23	18
Material	15	21
Other Cost	18	28
Profit or loss	8	8

Directional Bar Diagram :

Question: Represent the following by means of Directional Bar Diagram.

Year	Net Profit	Expenses	Gross Profit
2001	3	1	3
2002	6	2	8
2003	5	2	7
2004	8	3	11
2005	9	3	12

Question: Represent the following by suitable diagram.

Year	Production	Complete goods	Incomplete goods
2003	400	300	100
2004	450	350	100
2005	600	400	200
2006	650	500	50

Rectangular Diagram:

The number of units produce by A and B were 1,000 and 700 respectively. Show also cost and profit per unit.

Question: The following data relate to the monthly expenditure by two families. Represent the data by means of suitable Two- dimensional diagram.

Items of expenses	Family (A) Income (1,000 Rs.)	Family (B) Income (6,000 Rs.)
Food and Clothing	530	680
House Rent	340	415
Education	95	120
Miscellaneous	180	100

Question: Construct a suitable diagram representing the following information :

Items of expenses	Family (A) Income (1,000 Rs.)	Family (B) Income (6,000 Rs.)
Food	2,340	3,500
Clothing	2,000	2,500
House Rent	430	805
Education	820	460
Miscellaneous	900	1,320

Square Diagram :

Question : Represent the following data by square diagram

Country:	A	B	C	D
:	2,400	3,000	1,520	2,220

Circular Diagram or Pie Diagram:

Question: Construct a Pie diagram represent the following data:

Cost	Labor	Bricks	Cement	Steel	Timber	Supervision
Present	28	18	15	10	12	16

Question : Represent in income of five families by circular diagram.

Family:	A	B	C	D	E
Income (Rs.):	3,000	2,500	1,000	800	58

Question : Represent the following data by means of a suitable diagram.

No. of employed in the factories				
Year	Man	Women	Children	Total
2003	1200	1000	650	2850
2004	2500	2000	1300	5800

10.10 SUMMARY

Data are presented in a diagrammatic from in a diagrammatic presentation. It makes data impressive attractive, simple, easy to compare. There are three types of diagram, one dimension, two dimensions and three dimension diagrams.

UNIT-11 GRAPHICAL PRESENTATION

11.1 Objective

11.2 Introduction

11.3 Important of a graphic presentation

11.4 Graph of time series – histogram

11.5 Graph of frequency distribution

11.6 Types of frequency distribution

11.7 Questions

11.8 Summary

11.1 OBJECTIVE

After studying this unit student should understand about importance of graphic presentation, frequency distribution and its types.

11.2 INTRODUCTION

Graphical presentation is more used for presentation of data in graphic position. A graphic presentation is very easy in making and analyzing.

11.3 IMPORTANT OF A GRAPHIC PRESENTATION

1. Calculation of mode and median: A graphical presentation is useful data of mode and median by the help's of graph.
2. Presentation of historical and time-based information:

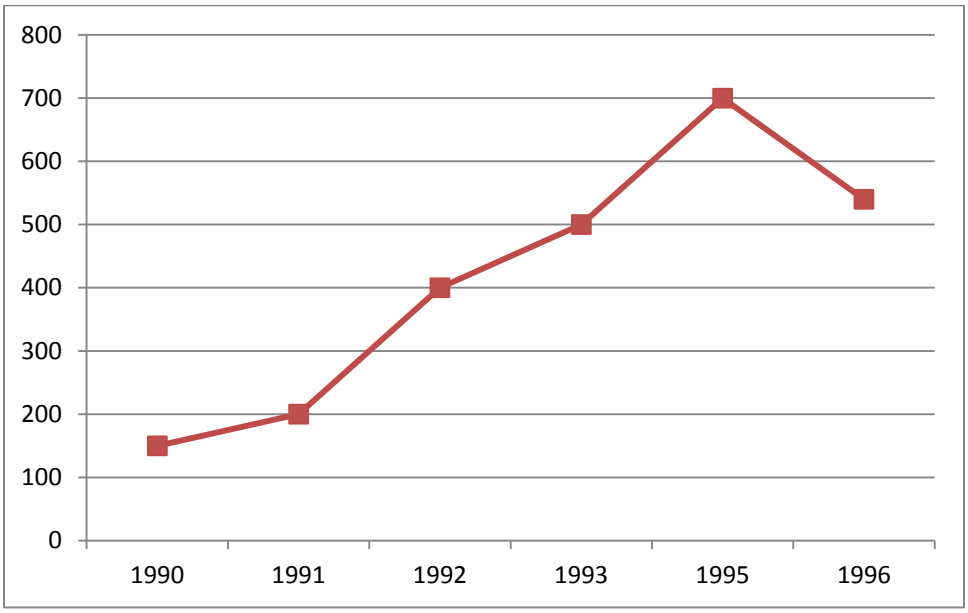
   A historical data is easily presented by historical and time based information in graphical presentation.
3. Easy to understand: Graph can be making of the largest and hardest data and can easily understand.
4. Suitable for Comparative study: When two graphs are presented on the base of different data, then it's suitable for comparison two different graphs.
5. Time Save : A graphical presentation takes less time and save the time as compression with table representation.

11.4 GRAPH OF TIME SERIES-HISTOGRAMS

Time series can be plotted by the help of graph, time series is a statically data and this statistical data are arranged in chronological order.

In time series we use the data related to economy and business. When an original data are plated in graph paper then it's called absolute histogram.

In histogram an independent variable on x-axis and dependent variable on y-axis.



Types of Histograph:

Histograph can be of following types

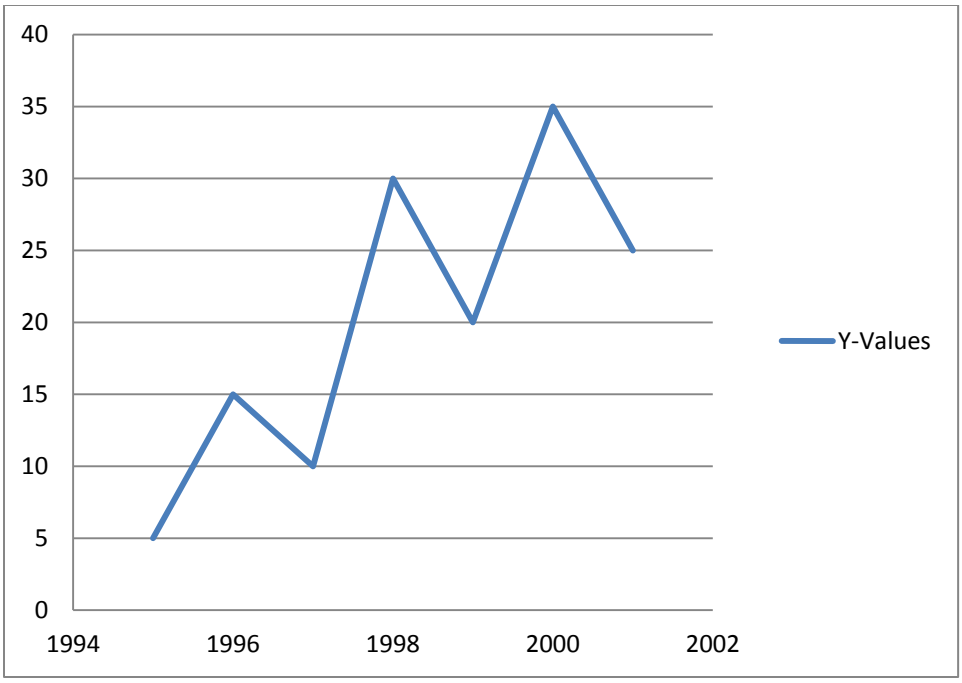
(I) Natural Scale histograms: A natural scale histograms are divided in five part:-

(a) On dependent variable Histogram: - In one dependent variable histograms, we represent only one variable. One variable are use for the salary of made in different year and other is no. of student passed in particular subject.

Example: Present the following data by a suitable graph:

Year	1995	1996	1997	1998	1999	2000	2001
Production	5	15	10	30	20	35	25

Solution : Graph showing production 1995-2001

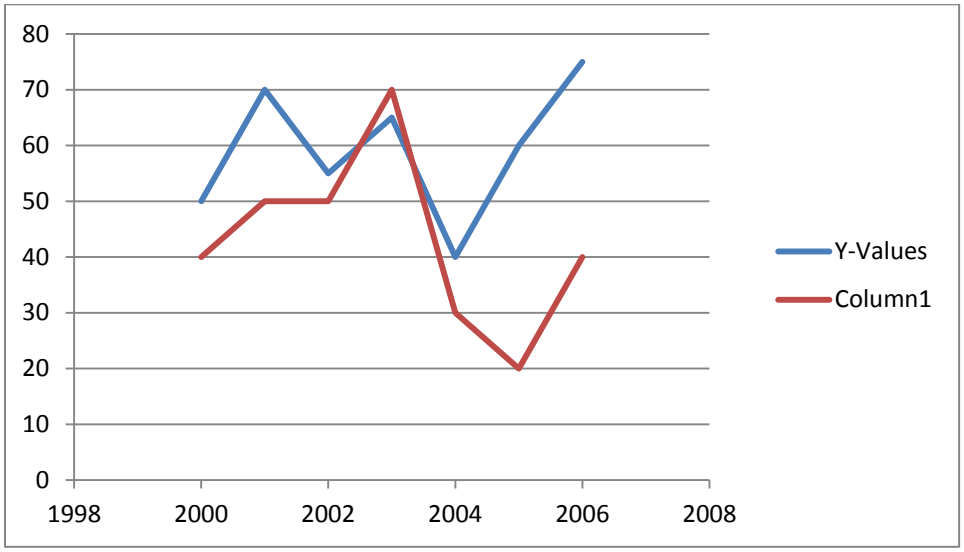


(b) More than one dependent variable: In this case here we use two variables in one graph. In plating one variable is dependent and expenses import and export and production of different product etc.

Example: The following data give the income and expenses of different year.

Year	2000	2001	2002	2003	2004	2005	2006
Income	50	70	55	65	40	60	75
Expenses	40	50	50	70	30	20	40

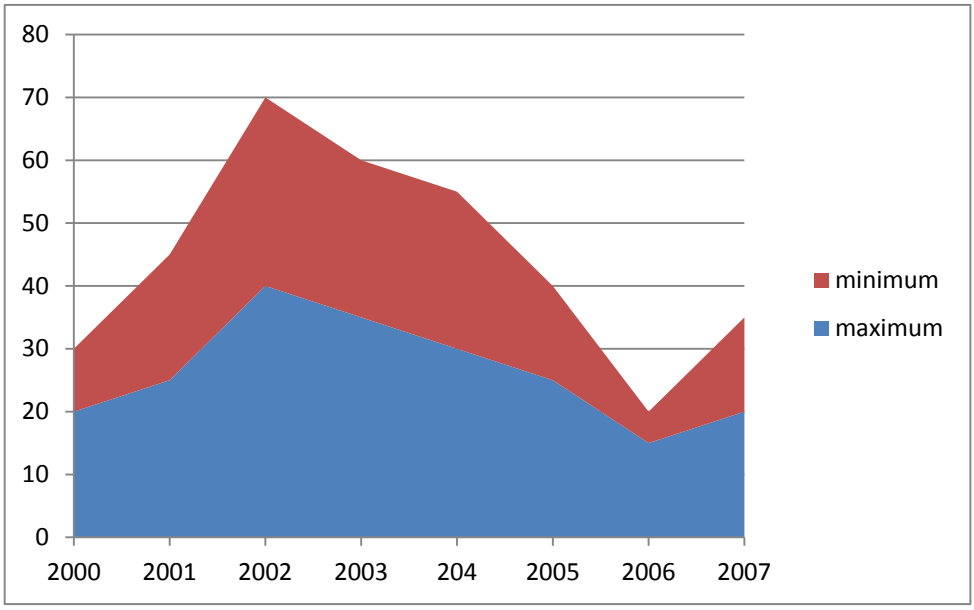
Solution: Graph showing income and expenses.



(c) Maximum-Minimum value graph: In some case a graph are used for showing the minimum and maximum data for comparison, in same period of time. This line represent the different type :

Example 3 : Show a following data in graph use of Capital in year.

Year	2000	2001	2002	2003	2004	2005	2006	2007
Maximum	20	25	40	35	30	25	15	20
Minimum	10	20	30	25	25	15	5	15



(d) Net balance graph:

A net balance graph also called the silhouette or difference graph. This graph given by “silhouette” so it’s called as silhouette chart.

A net balance graph is use for showing import and export, purchase and sales and Income and expenses etc.

It is constructed in two ways:

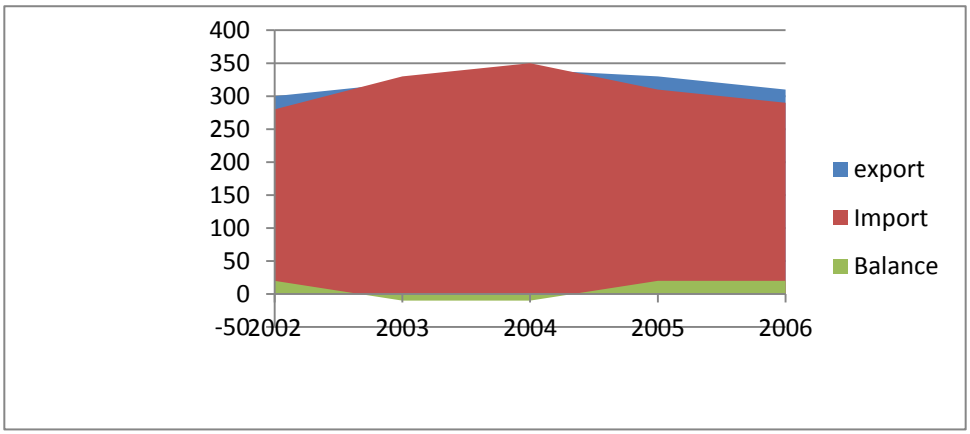
(i) Use the single graph paper for two histogram, and a difference of two histogram are shown by third histogram. The difference comes in positive or in negative

(ii) Plot two histogram in one graph paper and show the difference (positive or negative) with different colour.

Example : Represent the following data by net balance graph.

Year	Export	Import	Balance
2002	300	280	+20
2003	320	330	-10
2004	340	350	-10
2005	330	310	+20
2006	310	290	+20

Solution:



(e) Mixed graph:

When two measurement values are different in measure then used to be a graph. In mixed graph we used the two different scales. First scale is side on y axis and second on right side.

Example: Represent the following data relating to the arrival of foreign tourists in India and earning therefore with the help of a suitable graph:

Year	Tourists (lakh)	Earnings (crores)
1995	2	25
1996	3	33
1997	3	35
1998	4	40
1999	4	48
2000	5	70

(b) Range or Zone graph: A range or Zone graph is used for showing the highest and lowest position of an independent value. For

example, showing the temperature of person, highest and lowest value of a commodity.

Example:The following data show the maximum and minimum price quotations of the share of a company for five consecutive weeks. Present the data by on appropriate graph.

Week	1st	2nd	3rd	4th	5th
High Price	110	105	115	112	120
Low Price	100	103	110	108	115

11.5 GRAPH OF FREQUENCY DISTRIBUTION

For the graph frequency distribution the principal are same as for graphic and diagrammatic representation.

A frequency distribution is a contentment way of presentation a large mass of data in tabular form by grouping the data

There is type of frequency distribution, individual, discrete and continuous.

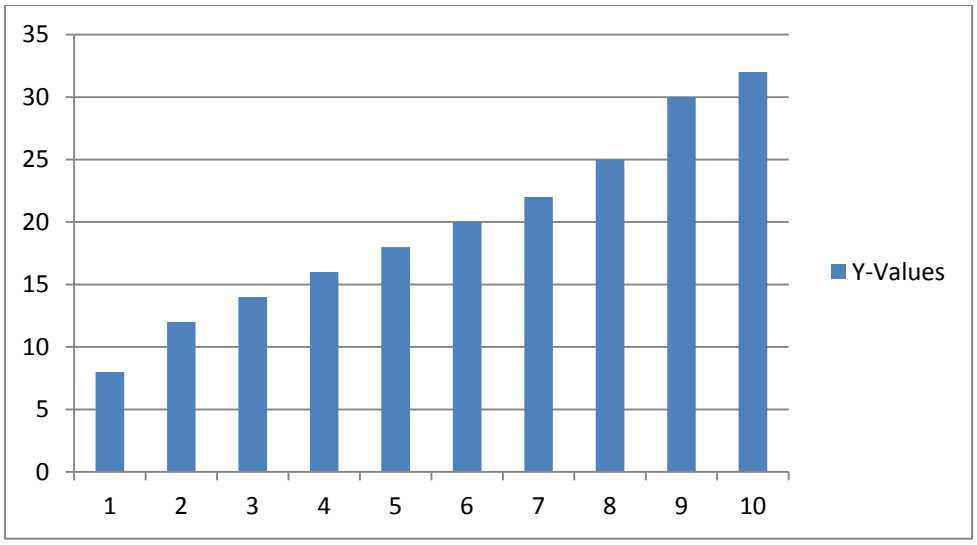
11.6 TYPES OF FREQUENCY DISTRIBUTION

Frequency distribution graphs are three types:-

1. Graph under series of individual observation: - In individual series first arranged the series in a sending or descending order and then it should be plotted in the graph paper.

Example: Following are marks obtained 10 students.

Serial No.	Marks.
1	8
2	12
3	14
4	16
5	18
6	20
7	22
8	25
9	30
10	32



2. **Graphs under discreet series:** In Discrete series values are fixed a definite of item. Here value is shown by line order.

Example: Following are the marks obtained by student in mathematic in class (iii).

Marks.	No. of student
1	5
2	8
3	12
4	15
5	25
6	10
7	6

3. Graphs under Continuous series:- In continuous series a graph is represented in different way :

   (i) Histogram : A histogram is used to represent the continuous series. In histogram use the equal continuous series.

Example : From the following data construct a histogram.

Marks	0-5	5-10	10-15	15-20	20-25	25-30	30-35
No. of student	10	20	15	12	25	30	18

(ii) Frequency Polygon: Frequency polygon is same as histogram, but here there is difference. In frequency polygon we connect here there is only are the midpoint with the help of a single line which we do not do in histogram.

Example: Prepare a frequency polygon from the following data:

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No. of student	5	8	12	20	15	10

(iii) Frequency curve: If the class intervals in a distribution are continuously reduced in size and if the number of items in the distribution is continuously increased the frequency polygon will re-simple a smooth curve called a smoothed frequency curve or simply a frequency curve.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of student	2	3	7	13	11	9	4

(iv) Ogive (or cumulative frequency curves): An ogive is a graphic presentation of a cumulative frequency distribution. Since a cumulative frequency distributed can be of “less then” or “more than” type, therefore there are two type of ogives “less then” ogive a “more than” ogive.

It is also called a cumulative frequency distribution polygon. This graph is useful to find median and quartiles in an easy manner.

Example: - Prepare an ogive curve from following data.

Marks	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	2	5	4	10	15	8
Cumulative	2	7	11	21	36	44

11.7 QUESTIONS

Question (1) : Prepare a frequency distribution from the following data taking the class width as 10 by exclusive method.

42	31	19	14	23	28	17	53	22	21
20	30	63	9	30	29	12	21	32	25
18	42	38	44	7	39	6	15	41	45
30	2	44	15	37	5	36	31	46	46
19	24	22	30	30	13	31	30	44	6

Question (2): Draw a histogram to represent the following frequency distribution.

Valve	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency	328	350	720	664	598	524	378	244

Question(3): Draw a histogram for the following distribution of marks of 150 students:

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of student	10	21	28	43	31	14	2	1

Question(4): Draw a histogram to represent the following frequency distribution:

Marks	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	12	30	35	65	45	25	18

Question(5):Represent the following by means of frequency distribution.

Daily wages (Rs.)	10-15	15-20	20-25	25-30	30-40	40-60	60-80
No. of Workers	1	19	27	15	12	12	8

Question(6): Draw a histogram to represent the following frequency distribution:

Mid-value (x)	5	15	25	35	45	55	65
Frequency (f)	9	19	27	15	12	10	8

Question(7) : Draw a frequency polygon for the following distribution of marks obtained by 200 students in an examination.

Marks	300-305	350-400	400-450	450-500	500-550	550-600	600-650	650-700	700-750
Frequency	5	14	23	50	52	25	22	7	2

* Question (8) : Prepare a histogram and frequency polygon from the following data:

Class	0-6	6-12	12-18	18-24	24-30	30-36
Frequency	4	8	15	20	12	6

* Question (9) : The following data represent the frequency distribution of the weekly wages (in Rs.) of 265 workers in a factory.

Weekly wages (in Rs.)	50-59	60-69	70-79	80-89	90-99	100-109	110-119
No. of workers	15	40	50	60	45	40	15

Present the above distribution graphically by drawing (1) Histogram (2) Frequency polygon.

Question(10): Given below is the distribution of marks obtained by 100 students in an examination:

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of student	4	10	20	35	15	6	10

Question(11): Draw the “less than” and “more than” ogive curves from of the above data.

Weekly wages (in Rs.)	0-20	20-40	40-60	60-80	80-100

No. of workers	10	20	40	20	10

Question(12): Draw the “less than” and “more than” ogive curves from the given below.

Weight	60-64	65-70	70-74	75-79	80-84	85-89	90-94	95-99	100-104
No. of student	10	15	24	30	22	18	13	8	5

11.8 SUMMARY

When data is presented in graphical manner it is known as graphical presentation, nature scale graph, mixed graph, frequency curve are some of important graphs.

UNIT-12 RATIO, PERCENTAE AND RATES

12.1 Objective

12.2 Introduction

12.3 Some importance Point in ration

12.4 In verse Ratio

12.5 Compound Ratio

12.6 Duplicate Ratio

12.7 Sub-Duplicate Ratio

12.8 Triplicate Ratio

12.9 Sub-Triplicate Ratios

12.10 Continued Ratio

12.11 Commensurable Ratio

12.12 Incommensurable Ratio

12.13 Percentage

12.14 Profit and Loss

12.15 Simple Interest

12.16 Compound Interest

12.17 Logarithm

12.18 Summary

12.1 OBJECTIVE

After studying this unit a student should know about Ratio, Types of Ratio, Percentage, Profit and Loss and logarithm, their application and calculation.

12.2 INTRODUCTION

A Ratio isa comparison of the sizes of two or more quantities of the some kind by division.

If a and b are two quantities of the same kind (in same unit), then the fraction a/b is called the ratio of a to b. it is written a: b. Thus the ratio of a

to b = a/b or a: b. The quantities a and b are called the term of ratio, a is called the first term or antecedent and b is called the second term or consequent.

12.3 SOME IMPORTANT POINT IN RATIO

1. Both term of a ratio can multiply or divided by the same number (non-zero). Usually ratio is repressed in lowest term or simplest form

   Example: 12:16 = 12/16

   $= \frac{3\times4}{4\times4}$

   =3/4

   =3:4

2. The order of the terms in a ratio is important.

   Example: 3:4 is not same as 4:3

3. Ratio exists only between quantities of same kind

   Example: (I) There is no ratio between no. of children in a class and the salary of a teacher.

   (ii) There is no ratio between the weight of one child and the age of another child.

4. Quantities to be compared (by division) must be in the same units.

   Example: (I) Ratio between 150gm and 2kg.

   = 150/2000

   = 3/40

   = 3:40

   (ii) Ratio between 25 minutes and 45 Seconds

   =1500/45

   = 100/3

   =100:3

5. To compare two ratios, convert them into equivalent like fractions.

   Example: - To find ratio is greater

   $2\frac{2}{3}: 3\frac{1}{3}$

   = 3.6:4.8

Solution: - $2\frac{2}{3}: 3\frac{1}{3}$

$= \frac{7}{3}: \frac{10}{3}$

=7:10

3.6:4.8 = 3.6/4.8

= 36/48

=3/4

L.C.M of 10 and 4 is 20.

So, $\frac{7}{2}$ = (7×2) / (10×2) = 14/20

And ¾ = (3×5) / (4×5) = 15/20

As 15>14, So 15/20>14/20, ¾> 7/10

Hence

3.6:4.8 is grater Ratio

Example: Rough weight 56.7kg. If he reduces his weight in the ratio 7:6 find his new weight.

Solution: Original weight of wrong = 56.7kg.

He reduces his weight in the ratio = 7:6

His new weight = (6 ×56.7)/7= 6×8.1 = 48.6kg.

Answer = 48.6kg.

Example: - Simplify the ratio 1/3: 1/8:1/6

Solution: L.C.M. of 3, 8 and 6 is 24

$=\frac{1}{3}:\frac{1}{8}:\frac{1}{6}$

$= \frac{1\times24}{3}: \frac{1\times24}{8}:\frac{1\times24}{6}$

= 8:3:4

Answer = 8:3:4

Example 3: The ratio of the no of boy to the no. of girls in a school in a school of 720 students is 3:5. If 18 new girls are admitted the school, find how many new boys may be admitted so that the ratio of the no. of boys to the no. of girls may change 2:3.

Solution : - The ratio of no. of boy to no. of girls=3:5

Sum of the ratio = 3+5= 8

So,

The no. of boy in the school (3×720) / 8

= 270

The no. of girls in the school = (5×720) / 8

= 450

Let the no. of boy admitted be x, then the no.of boy becomes (270+x)

After admitting of new girls the no. of girls become 450+18= 468

According to given description of the problem (270+x) / 468= 2/3

Or 3(270+x) 2×468

Or 8/0+3x = 936 or 3x = 126

X= 126/3

X= 42

Hence the no.of new boys admitted = 42

Answer = 42

12.4 INVERSE RATIO

One ratio is she in inverse of another if their product is 1. Thus a: b is the inverse of b: a.

12.5 COMPOUND RATIO

The ratio compound ratio a: b and c: d is ac:bd. For example (I) Compound ratio of 3:4 and 5:7 is 15:28

(ii) Compound ratio of 2:3, 5:7 and 4:9 is 2×5×4: 3×7×9= 40:189

12.6 DUPLICATE RATIO

The ratio compound of itself is called its duplicate ratio. Thus a^2:b^2 is the duplicate ratio of a: b.

For example: Duplication of 2:3 is 2^2:3^2 = 4:3

12.7 SUB-DUPLICATE

A sub-duplicate ratio is totally reversed of duplicate ratio. The sub-duplicate ratio of a: b is $\sqrt{a}$:$\sqrt{b}$ For example: - Sub duplicate ratio of 4:9 is$\sqrt{4}$: $\sqrt{9}$= 2:3.

12.8 TRIPLICATE RATIO

A triplicate ratio is cube of ratio, the triplicate ratio of a: b is a^3:b^3

For example: Tropical ratio of 2:3 is 8:27.

12.9 SUP-TRIPLICATE RATIO

A sub triplicate ratio is cube root of rote. The triplicate ratio of a: b is $\sqrt[3]{a}$: $\sqrt[3]{b}$.

For example: Sub- triplicate ratio of 8:27 is $\sqrt[3]{8}$: $\sqrt[3]{27}$ = 2:3

12.10 CONTINUED RATIO

The continued ratios are three similar quantities, means it is the relation between the magnitudes of there or more quantities of the same kind. A, B, C are written as a: b: c. for example: The continued ratio of Rs. 200, Rs. 400 and Rs. 600 is 200:400:600 = 1:2:3.

12.11 COMMENSURABLE RATIO

If the ratio of two similar quantities can be expressed as a ratio of two integers the quantities are said to be commensurable ratio.$\sqrt{a}$: $\sqrt{b}$= a: b

For example: $\sqrt{9}$: $\sqrt{4}$= 3:2 these are expressed as a ratio of two integers the quantities so it's a commensurable ratio.

12.12 INCOMMENSURABLE RATIO

If the ratio of two similar quantities cannot be expressed as a ratio of two integers the quantities are said to be incommensurable ratio.$\sqrt{a}$: $\sqrt{b}$.= a:b.

For example: The ratio of two integers and therefore $\sqrt{3}$ and $\sqrt{2}$ are incommensurable ratio.

Example: The monthly incomes of two persons are in the ratio 4:5 and their monthly expenditure is in the ratio 7:9 if such Rs. 50 per month fined monthly incomes.

Solution: Let the monthly incomes of two person be Rs. 4x and 5x, so that the ratio is Rs. 4x: 5x = 4:5. If such saves Rs. 50 per month, then the expenditures of two person are Rs. (4x-50) and 5x-50).

$\frac{4x-50}{5x-50} = \frac{7}{9}$

36x–450= 35x–350

36x–35x= –350+450

X= 100

The monthly income will the total wages of the workers of a factory be increased or decreased if there be a reduction in the number of workers in the ratio 15:11 and an increment in their wages in the ratio 22:25.

Solution : Let x be the original number of workers and Rs. Y the wages per persons.

Then total wages before change =by

After reeducation the number wages per workers = (11x) /15

After increment the average wages per workers = (25/y) /22

The total wages after change = (11/15) x + (25/22) y = 5xy/6

Thus, the total wages of workers get decreased from r × y to Rs. 5xy/6

Hence the required ratio in which the total wages decrease is by: 5xy/6= 6:5.

Answer : 6:5.

12.13 PERCENTAGE

The word percentage comes from per centum which means per hundred. Expressing quantities in percentage makes it easy to compare them, by the symbol %. Mark on a report Card, annual rate of interest in a bank and profit and loss figures are generally expressed as a percentage.

We shall study the following application of percentage:-

(i) Profit and loss (ii) Discount (iii) Sale, tax or value- added tax (VAT) (IV) Simple interest (v) compound interest.

12.14 PROFIT AND LOSS

The basic formula regarding profit and loss as follows:-

(i) If selling price (sp) > Cost price (cp), then

Profit = SP-CP

Profit % = $\frac{profit \times 100}{cp}$

(ii) If CP> SP then

Loss = CP-SP

Loss % = $\frac{loss \times 100}{cp}$

(iii) To find SP when CP and profit % or loss % are given:-

SP = $\left(\frac{100+profit\%}{100}\right)$ ×CP (in case of profit)

SP = $\left(\frac{100-loss\ \%}{100}\right)$ ×CP (in case of loss)

(iv) To find CP when SP and Profit % or loss % given.

CP = $\left(\frac{100\times Sp}{100+Profit\ \%}\right)$

CP = $\left(\frac{1000\times Sp}{100-Loss\%}\right)$

Overhead Expenses:

Expenses that are paid on repair, transportation, maintenance, cartage, salary of employees etc. are called overhead expenses and are added to the cast price of the goods.

The actual cost price includes the overhead expenses.

Some Important Point:-

(i) Profit or loss percent is always calculated on the cost price.

(ii) Overhead expenses are included or added to the cost price and profit or loss is calculated on this actual cost price.

(iii) Profit is also called gain.

Example : Tara bought a cooking range for Rs. 5000 and sold it for Rs. 6500. Find the profit percent.

Solution: CP= 5000

SP= 6500

SP>CP

Profit = SP-CP

= 6500-5000

= 1500

Profit % = $\frac{profit\times 100}{cp}$

= $\frac{1500}{5000}\times 100$

Profit % = 30%

Answer: Ishita made a profit of 30%.

Example: Rishi bought a Computer for Rs. 25,500 and sold it at a gain of 5%. Find his selling price.

Solution: CP= 25,500

Gain % = 5%

SP = $\left(\frac{100+gain\%}{100}\right)$ ×CP

SP = $\left(\frac{100+5}{100}\right)$ ×25,500

= $\frac{105}{100}$ ×25,500

SP= 26,775

Answer:- SP= 26,775

Example: On selling a mobile for Rs. 1216, Mohan loses 5%. Find the cost price.

Solution:- SP= 1216

Loss% = 5%

CP = $\left(\frac{1000\times Sp}{100-Loss\%}\right)$

= $\left(\frac{1000\times1216}{100-5}\right)$

= $\frac{1216}{95}$

CP= 1520

Answer: CP = 1520

Example: By selling a garment for Rs. 460, a shopkeeper incurs a loss of 8%. At what price should he sell the garment to get a profit of 5%?

Solution:-

Here the problem is divided into two parts. We are given the SP and the loss% from where we can calculate the CP it he CP it's known and the Profit % we can find the SP.

SP= 460

Loss% = 8%

CP = $\left(\frac{1000\times Sp}{100-Loss\%}\right)$

= $\left(\frac{100\times460}{92}\right)$

CP= 500

CP of the garment is Rs. 500

For a profit of 5%, the shopkeeper has to sell it at the following SP.

SP = $\left(\frac{100+profit\%}{100}\right)$ ×CP

SP = $\left(\frac{100+5}{100}\right)$ ×500

$= \frac{105}{100} \times 500$

SP= 525

The shopkeeper should sell the garment at a price of Rs. 525

Answer: 525

Example: Suhas sold two paintings for Rs. 99000 each on one he incurred a loss 10% and on the other the gained by 10. Find his actual gain or loss percent on the translation.

Solution:

Loss or gain is always on the CP , so in each case we need to find the CP.

Let the two paintings be painting A and Painting B

Painting A:- SP = 99000 Loss% = 10

CP = $\left(\frac{1000}{100-Loss\%}\right) \times$SP

$= \left(\frac{100\times 99000}{100-10}\right)$

$= \frac{9900000}{90}$

CP= 1,10,000

Painting B:-

SP: = 99000 Gain% = 10%

CP = $\left(\frac{100\times Sp}{100+gain\,\%}\right) \times$SP

$= \frac{100}{100+10} \times 99000$

$= \frac{100}{110} \times 99000$

CP= 90,000

Total CP = 1,10,000+90,000

= 2,00,000

Total SP= 99000+99000

= 198000

CP> SP

Loss = 200000-198000

= 2000

Loss% = $\frac{2000\times 100}{200000}$

Loss % = 1%

Suhas made a loss of 1% in the transaction.

Answer:Loss% 1%

Example:- Mr. Agrawal bought 6 dozen note books for Rs. 1080. On opening the packet he found that 18 notebook where damaged. At what price should he sell the remaining notebook to make a profit of 10%.

Solution:- 6 dozen notebooks are equal to 72 notebooks.

72 notebook were bought for Rs. 1080

1 notebook costs 1080/72= Rs. 15

18notebooks are damaged, so number of undamaged notebook is = 72-18

= 54notebooks

Profit is calculation on the CP

Mr. Agarwal has to earn a 10% profit

Profit = CP×Profit %

$1080\times \frac{10}{100}$= Rs. 108

SP= CP+ Profit = 1080+108

= 1188Rs.

Hence he has to sell the remaining 54 notebooks for Rs. 1188

Each notebook needs to be sold at a price of = 1188/54= 22Rs.

Answer: Remaining notebook are sold Rs. 1188

Example: If the cost price of a packet of 25 greeting cards is equal to the selling price of 20 greeting cards, find the gain or loss percent.

Solution: let the CP of each card be Rs. X

The CP of 20 cards = Rs. 20x

SP of 20 cards = Rs. 25x

CP of 20 cards = Rs.20x

SP of 20 cards = CP of 25 cards = Rs. 25x

SP>CP, hence there is a profit.

Profit = SP-CP

= 25x–20x

=5x

Profit % = $\frac{profit \times 100}{cp}$

= $\frac{5}{20} \times 100$

Profit % =25%

There is gain of 25%

Answer: 25%

Example: Sohil bought a plot of land for Rs. 96000. He sold 2/5 of it at a loss of 6%. At what profit percent should be shell the remaining part to get a profit of 10% on the whole.

Solution: Total CP for the plot of land. Rs. 96,000

2/5 of Rs. 96,000 = Rs. 38400

3/5of 96000 = 57,600

CP of 2/5 of land = 38400

Sohil increased a loss of 6% on the sale of 2/5 of the plot. SP of 2/5 of the plot:-

SP = $\left(\frac{100-loss\,\%}{100}\right) \times$CP

= $\left(\frac{100-6}{100}\right) \times 38400$

= $\frac{94 \times 38400}{100}$

= 36,095

10% of the profit on the whole means,

= $\frac{10}{100} \times 96{,}000$

= 9600

But he incurs a loss of Rs. 38,400 – Rs. 36,096

= 2304 on 2/5th of the plot.

So to get a profit of 10% on the whole transaction, he should get a profit of

Rs. 9600+Rs. 2304= Rs. 11,904

CP of the remaining 3/5th of the land is Rs. 57,600

He needs to make a profit of Rs. 11904

∴ Profit % = $\frac{11904}{57600} \times 100$

= $\frac{62}{3}$%

Profit = $20\frac{2}{3}\%$

Sohail should sell the remaining plot of land at a profit of $20\frac{2}{3}\%$

Answer: $20\frac{2}{3}\%$

Discount

An article that is to be sold has a price on it which is known as its market price (M.P). When stored offer a discount during sales as diwali sale, winter sales or monsoon sale. The articles are sold at a price that is less than the market price. This reduced price is a certain percentage of the market price and is known as the discount. To find the discount percent we use the following formula.

Discount%= $\frac{Discount}{MP} \times 100$

If on article has a MP of Rs. 500 and is sold at Rs. 400 then the discount is Rs. 100.

Thus, $\frac{100}{500} \times 100$

MP= $\frac{100 \times SP}{100 - Discount\%}$

Some times more than one discount is offered a particular. Brand of items. For example, an article has a 20% discount on it. If the article is bought in large number an additional 5% discount is offered. Such discount are called successive discount.

Discount is always calculated on the market price. Selling price = Market price – Discount

Example:- The market price of an air condition Rs. 25,000. The shopkeeper offers a discount of 10% on its purchase. Find its selling price.

Solution:- MP = 25,000

Discount = 10%

Discount = 10% MP

= $\frac{10}{100} \times 25000$

Discount= 2500

Now

SP = MP-Discount

= 25000-2500

= 22,500

SP = 22,500

Answer:- SP= 22,500

Example"- A the annual sate at a supermarket, a silk kurta market at Rs. 900 is sold for Rs. 765. Find discount percent

Solution:- MP = 900

SP= 765

Discount = MP-SP

= 900-765

= 135

∴ Discount%= $\frac{Discount}{MP} \times 100$

= $\frac{135}{900} \times 100$

= 15%

Answer = Discount = 15%

Example:- During a diwali sate a saree is sold for Rs. 3250 after allowing a discount of 35% find the market price.

Solution:

Method I (unitary method)

Let the MP be Rs. 100

Discount = 35%

SP- MP- discount

= 100-35

= 65

If SP is 65 then MP = $\frac{100}{65} \times 3250$

= 5000

Hence MP= 5000

Method II (using the formula)

MP= $\frac{100 \times SP}{100 - Discount\%}$

= $\frac{100 \times 3250}{100 - 35}$

= $\frac{100 \times 3250}{65}$

MP = 5000

Answer:- MP= 5000

Example:- The market price of a television set is Rs. 28600. After allowing a discount of 25% the dealer still makes a profit of 10%. Find the cost price of the television set.

Solution:- MP = Rs. 28,600

Rate of discount = 25%

Discount = 25% of Rs. 28600

$=\frac{25}{100}\times 28600$

= 7150

SP = MP- discount

= 28600-7150

= 21,450

Now CP= $\frac{100\times SP}{100+Profit\%}$

= $\frac{100\times 21450}{100+10}$

= $\frac{100\times 21450}{110}$

CP = 19500

Answer:- Cost price of the television Set is = 19500

Example:- During a book exhibition , a set of book were sold for Rs. 9000 after giving srccerrive discount of 20% and 10% respectively. Find the market price of the set of books.

Solution :- Let the MP be Rs. 100

First discount = 20%

SP= Rs. 100-20

= 80Rs.

Second discount = 10% of 80

= $\frac{10\times 80}{100}$

= 8Rs.

SP after second discount = 80-8= 72Rs.

When the SP is Rs. 72, the MP Rs. 100

When the SP is Rs. 9000, the MP is

= $\frac{100}{72}\times 9000$

MP = 12500

The market price of the set of book is 12500.

Answer:- MP of the set of book is 12500

SALES TAX OR VALUE- ADDED TAX (VAT)

Sales tax or value added tax is collected by the state government on the sale of items. This tax is charged by the shopkeeper who is then paid to the government. Values added tax is not a profit for the shopkeeper but is recorded in the bill book and then paid to the government.

There is various type of taxes that the government collects in the form of taxes. These include income tax wealth tax etc. that are collected as direct taxes. The direct tax is in the form of excise tax, sales tax , VAT etc.

Value added tax is calculated on the sale price of one item

The range of VAT is different for different kind of item. It is more on laxly items than on items of common use or daily use.

Example: The price of a refrigerator is marked at Rs. 15,400 if 15% VAT is charged find the total amount priti will have to pay to the shopkeeper to buy the refrigerator.

Solution : Since VAT is 15% priti will have to pay an additional 15% of Rs. 15400 as VAT. Therefore the additional amount priti will have to pay as VAT is

$= \frac{15}{100} \times 15400$

$= 15 \times 154$

$= 2310$

The total amount that priti has to pay is MP of refrigerator + VAT

= 15400+2310

= 17.710

Priti has to pay Rs. 17710 for the refrigerator.

Answer:- 17710

Example:- Rohini paid 12% VAT on an air conditioner bought for Rs. 28000, what was the amount selling price of the air conditioner?

Solution:- Let the SP= Rs. 100

12% VAT is paid on the SP. The amount paid will be Rs. 100+12 = Rs. 112

When Rs. 12 is paid the SP is Rs. 100

Rs. 2800 0 is paid the SP will be

$= \frac{100}{112} \times 28000$

= 25,000

The actual selling price of the air conditioner was Rs. 25,000.

12.15 SIMPLE INTEREST

We know that while calculating simple interest, it is calculated uniformly throughout the loan period on the original principal and is given by the following formula:

Simple Interest (SI) = $\frac{Principal \times rate \times time}{100}$

SI= $\frac{P \times R \times T}{100}$

Amount = Principal + SI

Interest = Amount – Principal

Example:- How much interest will be earned or Rs. 2000 at 6% simple interest for 2 year?

Solution:- SI = $\frac{P \times R \times T}{100}$

= $\frac{2000 \times 6 \times 2}{100}$

= 240Rs.

Answer: SI = 240Rs.

Example:- Sonia deposit Rs. 50,000 in a bank for two year with the interest rate of 5.5% P.a. How much interest would she earn? And find the final value of investment.

Solution: SI = $\frac{P \times R \times T}{100}$

= $\frac{50{,}000 \times 5.5 \times 2}{100}$

SI = 5500Rs.

Final value of investment:-

A= P(1+RT)

= 50,000(1+$\frac{5.5}{100} \times 2$)

= 50,000 $\left(1 + \frac{11}{10}\right)$

= $\left(\frac{50{,}000 \times 111}{100}\right)$

= 5,5500

Or

A= P+SI

= 50,000+5500

= 5,5500

Answer: SI = 5500, A = 55,500

Example: Find the rate of interest if the amount owned after 6 months is Rs. 1050, borrowed amount being Rs. 1000.

Solution: We know= A = P+ P×R×T

1050= 1000+1000× 6 ×6/12

50= 500r

R= 50/500

R= 1/10

= 10%

Answer = r= 10%

Example:- Rahul invest Rs. 70,000 in a bank at the rate of 6.5% or simple interest rate. He received Rs. 85925 after the end of term. Find out the period for which sum was invested by rahul.

Solution: A = P(1+RT)

85925= 7000(1+$\frac{6.5}{100} \times t$)

85925/7000= 6.5t

22.75 = 6.5t

T= 3.5 year

Answer:- Time = 3.5year

12.16 COMPOUND INTEREST

The method we have used for the calculation of compound interest in the problems discussed. So far is quite length and cumbersome. As the period for which interest is payable increases, the process of calculating the compound interest becomes lengthier and time consuming.

Example:- Suppose you deposit Rs. 50,000 in ICICI bank for 2 year at 7% P.a. Compounded annually interest will be calculated in the following way.

Interest for first year:

I= P×R×T

$= 50{,}000 \times \frac{7}{100} \times 1$

= 3500

Interest far second year :

50000+3500

= 53500

$= 53500 \times \frac{7}{100} \times 1$

= 3745

Total interest = Interest for first year+ Interest for second year

= 3500+3745

= 7245

Example: determine the compound amount and compound interest on Rs. 1000 at 6% compounded semi-annually for 6 year. Given that (l+i) = 1.42576 for I= 3% and n= 12

Solution:- I = 0.06/2 = 0.03

N= 6×2= 12

P= 1000

Compound Amount (A_{12}) = P $(l+i)^n$

$= 1000(1+0.03)^{12}$

= 1000×1.42576

= 1425.76

Compound Interest = 142.76 – 1000

= 425.76

Answer: Compound amount = 1425.76

Compound Interest = 425.76

Example: Sonia deposited Rs. 1,00,000 in a nationalized bank for three years, if the rates of interest are 7% P.a. calculated the interest that bank has to pay to Sonia after three year if interest is compound annually. Also calculated the amount at the end of third year.

Solution: Principal of First year Rs. = 1,00,000

Interest for first year = P×R×T

$= 1{,}00{,}000 \times \frac{7}{100} \times 1$

= 7000

Principal for the second year = Principal for first year + Interest of first year

= 1,00,000+7000

= 1,07,000

Interest for second year:- $1{,}07{,}000 \times \frac{7}{100} \times 1$

= 7490

Principal for third year= Principal of second year+ Interest for second year.

= 1,07,000+7490

= 114490

Interest for the third year = $114490 \times \frac{7}{100} \times 1$

= 8014.30

Compound Interest at the end of third year

= Rs. (7000+7490+8014.30)

=Rs. 22504.30

Amount at the end of third year

= Principal (initial deposit) + Compound interest

= Rs. (1,00,000+22504.30)

= 1,22,504.30

Answer: Amount end of third year= 122504.30

Conversion period

Conversion period	Description	No. of Conversion period
1day 1month 3 months 6 months 12 months	Compounded daily Compounded monthly Compounded Quartly Compounded Semi-Quartly Compounded Annually	365 12 4 2 1

Formula for compound interest:

Tacking the principal as 'P' the rate of interest per conversion period as 'I' (in decimal) , the number of conversion as 'n' the accrued amount after 'n' payment periods as an we have accrued amount at the end of first payment period.

$A_1 = P+Pi = (l+ I)$

At the end of second payment period

$A_2 = A_1 + A_1i = A_1 (l + I)$

$= P (l + I) (l+i)$

$= P (l +I)^2$

At the end of the third payment period

$A_3 = A_2 + A_2i$

$A_2 (l +I)$

$P (l+i)^2 (l+i)$

$P (l+i)^3$

$A_n = A_{n-1} + A_{n-1}i$

$= A_{n-1} (l+i)$

$= P (l+i)^{n-1} (l+i)$

$P(l+i)^n$

Thus the accrued amount A_n or a principal P after n conversion periods at 'I' rate of interest per conversion period is given by

$A_n = P (l +i)^n$

Where $I = \frac{Annual\ rate\ of\ interest}{Number\ of\ conversion\ perios\ per\ year}$

Interest = $A_n – P = P(l+i)^n - P$

$= P [(l+i)^n – 1]$

Example: Rs. 2000 is invested at annual rate of interest of 10% what is the amount after two year if compounding is done (a) Annually (b) Semi-annually (c) Quarterly (d) monthly.

Solution: (a) Compounding is done annually

Here principal P= Rs. 2000

Conversion Period n in = 2 years

Rate of interest per conversation period (1 year) = 0.10

$A_n = P (l +i)^n$

$A_2 = 2000 (1 +0.01)^2$

$= 2000 \times (1.1)^2$

$= 2000 \times 1.21$

$= 2420$

(b) For semiannual Compounding

$N= 2 \times 2 = 4$

$I = 0.1/2 = 0.05$

$A_4 = 2000 (1 +0.05)^4$

$= 2000 \times 1.2155$

$= 2431$

(c) For quarterly Compounding

$N= 4 \times 2 = 8$

$I= 0.2/4= 0.025$

$A_8 = 2000 (1 +0.025)^8$

$= 2000 \times 1.2184$

$= 2436.80$

(d) For monthly compounding

$N= 12 \times 2 = 24$, $i= 0.1/12 = 0.00833$

$A_{12} = 2000 (1 +0.00833)^{12}$

$= 2000 \times 1.22029$

$= 2440.58$

Answer: (a) Annually = 2420 (b) Semi annually = 2431 (c) Quarterly = 2436.80 (d) Monthly = 2440.58

12.17 LOGARITHM

The logarithm of a number to a given base is the index or the power to which the base must be raised to produce the number, to make it equal to the given number. If there are three quantities indicated by say a, x and n they are related as follows.

If $a^x = n$

Then x is said to be the logarithm of the number n to the base 'a' symbolically it can be expressed as follows.

Log $a^n = x$

The logarithm of n to the base 'a' is x, we give same illustrations below:-

$2^4 = 16$

$= \log_2 16 = 4$

Find the log Value of number

Step 1:- First subtract the one in total number and then write (.) and then see the log table.

Example 1:- Find the logarithm of 4597.

Log 4597 =m 3.6625 4-1=3

Log 459.7 = 2.6625

Log 45.94 = 1.6625

Log 4.594= 0.6625

Log .4594= -1.6625

Negative mantisa so must be converted into a positive mantissa before reference to a logarithm table. For example:-

-3.6872= -4+(4-3.6872)

$= \bar{4} + 0.3128$

$= \bar{4}.3128$

Here $\bar{4}$ is nagative while .3128 is positive

Add of Log

Example:- Add $\bar{4}$.74628 and 3.42367

= -4 + .74628

= 3+ .42367

= -1+1.16995

= 0.16995

Example: Find the value of log sib log2 is equal to .3010

Solution : log5 = log 10/2

= log10- log2

= .6990

Example: Find the number logarithm is 2.4678

From the antilog table for montisa. .467 The number = 2931

For mean difference 8 the number = 5

∴ For montisa .4678 the intergral part

Hence antilog 2.4678 = 293.6

Example: Find the number whose logarithm is – 2.4678

-2.4678 = -3+-2.4678

= -3+.5322

= $\bar{3}$ .5322

For mantissa .532 the number = 34404

For mean difference to the number = 2

∴ For mantissa .5322 the number = 3406

The characteristic is -3, therefore the number is less than one and there must be two zeros just after the decimal point.

Practice Question

Question 1: The ratio of the price two houses was 16:23. Two years later when the price of the first has increased by 10% and that of the second by Rs. 477, the ratio of the price becomes 11:20 find the original prices of the two houses.

Answer: Rs. 848 and Rs. 1219

Question 2: (I) in inverse ratio of 11:15

(ii) The ratio of the Quantities is 5:7 if the consequent of its inverse ratio is 5 the antecedent are.

Answer:- (I) 15:11 (ii) 20

Question 3: (I) find the compound ratio of 2:3, 9:4, 5:6 and 8:10

(ii) Find the duplicate ratio of 3:4

Answer: (I) 1:1 (ii) 9:16

Question 4: (I) find the sub- duplicate ratio of 25:36

(ii) Find the triplicate ratio of 2:3

Answer: (I) 5:6 (ii) 8:27

Question 5: (I) find the sub- triplicate ratio of 8:27.

(ii) Find the triplicate ratio of 2:3

Answer: (I) 2:3 (ii) 18:25

Question 6: Anand earns Rs. 80 in 7 hours and pramode Rs. 90 in 12 hour. Then find the ratio of their earning.

Answer: 32:21

Question 7 : Find the number if there ratio are 7:10 their difference is 105.

Answer: (245,350)

Question 8: P,Q and R there cities . The ratio of average temperature P and Q is 11:12 and that between P and R is 9:8. Then find the ratio between the average temperature of Q and R.

Answer: 27:22

Question 9: Find the number which when subtracted from each of the term of the ratio 19:31 reducing it to 1:4.

Answer : 15

Question 10: Daily earning of two persons are in the ratio 4:5 and their dailyexpenses are in the ratio 7:9. If each saves Rs. 50 perday then find their daily incomes in Rs.

Answer: (400,500)

Question 11: The ratio between the speeds of two trains is 7:8.if the second train runs 400 kms in 5 hours. Then find the speed of the first trains

Question 12: (a) Reena bought a book for Rs. 600 and sold it for Rs. 750. Find her profit or loss percent

(b) Ritu purchased a television trally far Rs. 3600 and later sold it to her friend Neetu for Rs. 3000. Find Reetu's gain or loss percent.

Answer: (a) 25% gains (b) $16\frac{2}{3}$ % loss

Question 13: (a) Sahari purchased a mobile phone for Rs. 4500 and sold it at a profit 5%. Find the amount of money she received on the deal.

(b) Karan bought a thermos flask for Rs. 970 and sold it at a loss of 10%. Find his selling price

Answer: (a) Rs. 4725 (b) Rs. 873

Question 14: (a) Lily sold a tea set far Rs. 4914 at a profit of 5% to her friend, chetna. At what price did lily purchase the tea set?

(b) Bhairav sold his bicycle for Rs. 4600 at a loss of 8%. Find his loss.

Answer: (a) Rs. 4680 (b) Rs. 5000

Question 15: (a) Manpreet bought an old computer for Rs. 18500 and spent Rs. 1500 on its repair. He then sold it to his friend Trarngit for Rs. 23000. Did manpreet incur a loss or profit and how much? Give your answer in percent.

Answer: (a) Rs. 3000, profit 5%

(b) Rahul bought a second – hand car for Rs. 1,90,000 and spent Rs. 3200 on its insurance and Rs. 10,500 on repair. He then sold it for Rs. 2,50,000. Find his gain or loss

Answer: (b) 22.73%

Question 16: (a) A television dealer sold a model of television for Rs. 24840 at a loss of 8% and another model of television for same price at a gain 8%. Find the total gain or loss percent

Answer: (a) 0.64% loss

(b) Madhav bought two pairs of jeans for Rs. 960 each he sold one of them at a loss of 4% and the other at a gain of 8%. Find madhav's gain or loss percent on the whole translation.

Answer: (a) 2% gain

Question 17: (a) Find the selling price of an article market at Rs. 750, if a discount of 20% is offered on it?

Answer: (a) Rs. 600

(b) Rini bought a pair of silver candle stands for Rs. 4650 in a discount sale of 7%. What was the market price of the silver candle stands?

Answer: (a) Rs. 5,000

Question 18: (a) the price of a raincoat was reduced from Rs. 960 to Rs. 720 during a sale. What was the discount percent offered?

Answer:Rs. 25%

(b) A Packet transistor is market at Rs. 2500. The shopkeeper allows a discount off 15% on it find the net selling price.

Answer: 2,125

Question 19: (a) During the festival season, the price of a washing machine is reduced form Rs. 14500 to Rs. 12,325. Find the percentage of discount offered.

Answer: 15%

(b) Shobhana paid Rs. 1260, for a tea set often a discount 10% was allowed on the marked price during a clearance sale. What was the market price of the tea set? How much dies shobhana due to the discount?

Answer: Rs. 1400, Rs. 140

Question 20: (a) A book seller marks his goods at a price that is 20% more than the cost price if the allows a discount of 10% what will be his gain or loss percent?

Answer: 8% gain

(b) During the monsoon sale at a loading saree shop the market price of a saree is Rs. 9575. The shopkeeper allows a discount of 12% on the market price and still makes a profit of 10%. Find the cost price of the saree.

Question 21: (a) At a poultry shop, all items are marked 30% above the cost price. Due to a bird flu epidemic, the shopkeeper had to sell the percent in the transaction.

Answer: 7,660

(b) At a poultry shop, all items are marked 30% above the cost price. Due to a bird flu epidemic the shopkeepers had to sell the goods at a discount of 30%. Find the gain or loss percent in the transaction.

Answer: 9% loss

Question 22: (a) The marked price of a dress was Rs. 650. All goals in the shop were being sold at a 20% discount. Anaya liked the shopkeeper offered a future discount of 10% to her. How much did Anaya pay for the dress/

Answer : Rs. 468

(b) Find the single discount which is equivalent to two successive discounts of 20% and 5%

Answer: 24%

Question 23 : (a) The marked price of a dining table set is Rs. 20,500. It the VAT charged is 10% how much does the customer pay?

Answer: Rs. 22,550

(b) Aparna bought a gold bracelet for Rs. 37,450 including 7% VAT, find the selling price of the bracelet.

Question 24: (a) Devyani liked a guitar that was priced at Rs. 12,000. A 10% VAT was to be charged on the bill, but shopkeeper agreed to give devyani a discount of 20%. Find the actual amount paid by devyani for the guitar.

Answer : Rs. 10,560

(b) Alka bought six T-shirts for Rs. 650 each and six long skirts for Rs. 800 each at a shop that was offering a 20% discount during its annual clearance sale. She had to pay a 5% VAT on the bill. How much did alka pay the shopkeeper? How much discount did she get on her purchase?

Answer: Rs. 7,308

Question 25: (a) The marked price of a book shelf is 25% more than its cast price. The shopkeeper gives a discount of 10% on it. If the marked price of the book shelf is Rs. 14,500. Find the cost price , selling price and the shopkeeper profit (as a percentage)

Answer: 11,600, Rs. 13,050

Question 26: There was a 20% discount on a laptop. A VAT of 18% was being charged on it. If dhiren bought the laptop for Rs. 236,600. What was its marked price.

Answer: 25,000

Question 27: Find the logarithm of these number:-

(i) 159456 (ii) 13922 (iii) 5946 (iv) 4597 (v) 6282 (vi) 18095 (vii) 4932

Answer: (I) 4.7742 (ii) 2.5935 (iii) 4.7943 (IV) 3.6625 (v) 15.798 (VI) 7.90882 (vii) 3.6930

12.18 SUMMARY

Comparison on two or more quantities is known as ratio, Different types of ratio are inverse ratio, compound, and duplicate, sub-duplicate, triplicate, sub- triplicate ratio.

Selling price minus cost price is known as profit while CP- SP = Loss

SI = $\frac{P \times R \times T}{100}$

Chi = A= P (1+r/n) not

A= amount of mummery accumulate in n years

P= principal amount

r= rate of interest,

t= total number of year

n= number of year

Uttar Pradesh Rajarshi Tandon
Open University

Bachelor of Business Administration

BBA-107

Business Statistics

BLOCK

4

MEASURES OF TENDENCY

UNIT-13 **211**

Measure of Central Tendency and Mean

UNIT-14 **233**

Median

UNIT-15 **263**

Mode

UNIT-16 **283**

Curriculum Design Committee

Prof. Omji Gupta **Director**

School of Management Studies, UPRTOU, Prayagraj

Dr Gyan Prakash Yadav

Assistant Professor

School of Management Studies, UPRTOU, Prayagraj

Dr Devesh Ranjan Tripathi

Assistant Professor

School of Management Studies, UPRTOU, Prayagraj

Course Preparation Committee

Dr T.M. Raini **Author**

Associate professor, Firoj Gandhi PGCollege, Raibarely

Dr Shiv Bushan Gupta

Associate professor, MMM PG college, Kalakankar, Pratapgarh

Dr Gaurav Sankalp

Academic consultant, School of Management Studies, UPRTOU, Prayagraj

Dr Ajay Tiwari

Assistant Professor

MONIRBA, Department of Commerce, University of Allahabad

Prof J.N. Mishra **Editor**

Professor, MONIRBA, Department of Commerce, University of Allahabad

Dr Gyan Prakash Yadav **Coordinator**

Assistant Professor

School of Management Studies, UPRTOU, Prayagraj

©UPRTOU, Prayagraj-2021

ISBN :

©All Rights are reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the **Uttar Pradesh Rajarshi Tondon Open University, Prayagraj.**

BLOCK INTRODUCTION

Measure of central tendency is the basic saturation and provide base in statics. This platform by which various statistical tools are modified and are applied. The present block IV measure of central tendency has four units.

Unit-13 : Discusses about concept of control tendency and mean

Unit-14 : Deals with median of various data, its computation and analysis

Unit-15 : Deals with Mode, Merits and limitations of modes with some examples.

Unit-16 : Deal with Gemrtic, Harmonic and moving average, its merits, limitations and application.

UNIT-13 MEASURE OF CENTRAL TENDENCY AND MEAN

13.1 Objective

13.2 Introduction

13.3 Essential of an ideal Average

13.4 Objective of Average

13.5 Different measure of Arithmetic mean

13.6 Calculation of AM – Discrete Series

13.7 Calculation of AM – Continuous Series

13.8 Weighted Arithmetic Mean

13.9 Property of Arithmetic Mean

13.10 Merits of Arithmetic Mean

13.11 Limitations of Arithmetic Mean

13.12 Questions

13.13 Summary

13.1 OBJECTIVE

After studying this unit students would know about.

- What is Average
- Objective and essentials of ideal average
- Calculation of Arithmetic Mean, its property, merit and limitation.

13.2 INTRODUCTION

In measure of central tendency we use a central set of data. In central tendency commonly, used methods are mean, median, mode, harmonic mean and geometric mean.

So we say that, it's a method of find the average of data, and for calculation we use a method called as "Mean".

According to "A.L, Bowely": "Average is statistical constants which enable us to comprehend in a single effort the significance of the whole."

According to "Ya-Lun-Chon". "An average is a typical value in a sense that it is sometimes employed to represent all individual value in a series or a variable."

13.3 ESSENTIAL OF AN IDEAL AVERAGE

(1) Absolute Value: An Average is always denoted absolute value and are not denoted by % and other unrelated number.

(2) In should be rigidly defined: An average should be rigidly defined. So that it has one and only one interpretation. In other words, the definition should not leave anything to the discretion of the investigator

(3) It should be easy to understand and simple to calculate :

An average should be easily understand and should not involve complex calculation. However it should not involve complex calculation. It should not be accomplished at the expense of other advantage.

(4) It should be based on all the observation: It is logically that an average which is supposed to stand for the whole group of values should be based on all the observations.

(5) It should be capable of future algebraic treatment: We would like to have on average that can be used for future mathematical treatment to enhance its utility.

(6) It should have sampling stability: By sampling stability we mean that if we take different random samples of the same size from a given population and compute the average of each sample we should expect to get approximately the same answer.

13.4 OBJECTIVE OF AVERAGE

1. To present a brief picture:- Average presents a help's in getting a clean picture
2. Facilitates Comparison: - With the help's of average it's easy to compare the two or more than two data, class and grouped.
3. To represent the entire group: - Entire group of sample is represented by averages.
4. Basis of statistical Analysis: - Average provides the value for statistical analysis.
5. Base for future Plans: - Average provides the base to make future plane and rule.
6. Presentation of Mathematical relationship: - Average is helpful to find the mathematical relationship in two classes and grouped.

13.5 DIFFERENT MEASURE OF ARITHMETIC MEAN

Calculation of Arithmetic Mean- Individual observations

In individual series a arithmetic mean denoted by $\bar{x}$

Arithmetic Mean $\overline{x} = \frac{\Sigma x}{n}$

Σx = Sum of different value.

N= No. of item

Example: - Find arithmetic mean of 10 student marks:-

Marks:- 12,8,17,13,15,9,18,11,6,1

Solution: - $\overline{x} = \frac{\Sigma x}{n}$

$= \frac{12+8+17+1315+9+18+11+6+1}{10}$

$= \frac{110}{10}$

=11

Answer = $\bar{x}$ = 11

Example 2:- Find arithmetic mean of data:-

4,6,12,2,10,20,18,15,7,10,11

Solution:-

$\overline{x} = \frac{\Sigma x}{n}$

$\bar{x} = \frac{4+6+12+2+10+20+18+15+7+10+11}{11}$

$\bar{x}$= 115/11

$\bar{x}$= 10.4545

Answer = $\bar{x}$ = 10.4545

Short cut method:

A short cut method is used when the value of x is very large. For calculation of short cut step are as follows:-

Step: Choose an assumed mean in the value of x, generally a assumed mean are taken mid value of 'x'

Step 2: Find value of'd' = x-A and take the algebraic sign + or –

Step 3: The arithmetic mean is given by

$$\bar{x} = A + \frac{\Sigma d}{n}$$

Example 3: Find Arithmetic mean by (me) direct method (ii) short cut method.

X: - 164,240,150,180,100,110,220,200,160

Solution:

S.no.	X	A= 100. D = x-A
1	160	-60
2	240	-140
3	150	-50
4	180	-80
5	100	0
6	110	-10
7	220	-120
8	200	-100
9	160	-60
	$\in x$= 1520	+620

(i) Direct method: $\bar{x} = \frac{\Sigma x}{n}$

= 1520/9

= 168.888

(ii) Short cut method: $\bar{x} = A + \bar{x} = \frac{\Sigma d}{n}$

=100+620/9

= 100+68.888

= 168.888

Answer = $\bar{x}$= (i) 168.888 (ii) 168.888

13.6 CALCULATION OF AM – DIRECT SERIES

In discrete series variable x is shown by the values x1, x2 ……..xn and frequencies f1,f2, ……..fn

Arithmetic mean calculation by direct and short cut method.

Formula:-

Direct Method: - $\bar{x} = \frac{f1x1+f2x2+fnxn}{f1+f2+\cdots\ldots\ldots\ldots fn}$

$= \frac{\in fx}{\in b}$ or $\frac{\in fx}{n}$

Short cut method: - $\bar{x}$ = A + $\frac{\in d}{n}$

Where:-

A= assumed mean

D= x-A

Example: - Calculate A.M. for following discrete frequency series.

Series: - 10, 20,30,40,50,60,70,80

F: - 2, 8,10,5,12,20,6,4

Series (x)	f.	fx
10	2	20
20	8	160
30	10	300
40	5	200
50	12	600
60	20	1200
70	6	420
80	4	320
	N= $\in f$= 67	$\in fx$= 3220

$\bar{x} = \frac{\in fx}{n}$

= 3220/67

$\bar{x}$= 487.059

Answer: - $\bar{x}$= 487.059

Example: - Find A. M. daily earning of 30 workers in a factory

Daily earning	50	100	150	200	250	300
Workers	5	2	8	10	3	2

Calculation by (I) Direct method (ii) Short cut method

Solution:

Daily earning (x)	Workers (f)	Fx	A= 200 D=x-A	Fd
50 100 150 200 250 300	5 2 8 10 3 2	250 200 1200 2000 750 600	-150 -100 -50 0 +50 +100	-750 -200 -400 0 +150 +200
	N= 30	$\in fx$= 5000		$\in fd$= -1000

Direct method: $\bar{x} = \frac{\in fx}{n}$

$\bar{x}$= 5000/30

$\bar{x}$=166.66

(ii) Short cut Method:-

$\bar{x}$ = A + $\frac{\in d}{n}$

$\bar{x}$= 200+ $\frac{-1000}{30}$

200-33.33

$\bar{x}$= 166.66

Answer = $\bar{x}$= (i) 166.66 (ii) 166.66

13.7 CALCULATION OF A.M.-CONTINUOUS SERIES

In continuous series it's applying by three methods:-

(i) Direct method:

$\bar{x} = \frac{f1x1+f2x2+fnxn}{f1+f2+\cdots\ldots\ldots fn}$

$= \frac{\in fx}{\in b}$ or $\frac{\in fx}{n}$

(ii) Short cut method:-

$\bar{x}$ = A + $\frac{\in d}{n}$

(iii) Step – deviation method:-

U= $\frac{x-a}{h}$

Where:

X= mid-value

h= Common size of class

$\bar{x}$ = A + $\frac{\in d}{n}$ ×h

Example: Find A.M. of data:

Mark	0-10	10-20	20-30	30-40	40-50	50-60
No. of student	5	7	8	14	10	6

Solution:

Marks	Mid-Value (x)	f.(no. of student)	Fx
0-10	5	5	25
10-20	15	7	105
20-30	25	8	200
30-40	35	14	490
40-50	45	10	450
50-60	55	6	330
	$\in f$=n= 50		$\in fx$= -1600

$\bar{x}$=$\frac{\in fx}{n}$

$\bar{x}$= 1600/50

$\bar{x}$= 32

Answer: $\bar{x}$= 32

Example: Find A.M. in (me) direct method (ii) Short cut method. (iii) Step – deviation method

Mark	0-10	10-20	20-30	30-40	40-50	50-60
No. of student	10	9	25	30	16	10

Solution:

Marks	Mid-Value (x)	f.(no. of student)	Fx	A= 200 D=x-A	H= 10 U= $\frac{x-a}{h}$	Fd	F4
0-10 10-20 20-30 30-40 40-50 50-60	5 15 25 35 45 55	10 9 25 30 16 10	50 135 625 1050 720 550	-30 -20 -10 0 10 20	-3 -2 -1 0 1 2	-300 -180 -250 0 160 100	-30 -18 -25 0 16 20
	n= 100		$\in fx$= -3130			$\in fd$= -370	$\in f4$= -37

Direct method: - $\bar{x} = \frac{\in fx}{n}$

$\bar{x}$= 3120/100

$\bar{x}$=31.20

(ii) Short cut Method:-

$\bar{x}$ = A + $\frac{\in d}{n}$

$\bar{x}$= 35+ $\frac{-370}{100}$

35-3.70

$\bar{x}$= 31.30

(iii) Step – deviation method:-

$\bar{x}$ = A + $\frac{\in d}{n}$ ×h

$\bar{x}$ = 35 + $\frac{-37}{100}$ ×10

= 35-3.70

= 31.30

Answer = $\bar{x}$= (i) 31.30 (ii) 31.30 (iii) 31.30

Example: If mean of following data are so and total frequency are 120 then find missing frequencies:

Class	0-20	20-40	40-60	60-80	80-100
f.	17	F1	32	F2	19

Solution :

Marks	Mid-Value (x)	f	U=x-50/20 (A=50, h=20)	F4
0-20 20-40 40-60 60-80 80-100	10 30 50 70 90	17 F1 32 F2 19	-2 -1 0 1 2	-34 -f1 0 f2 38
		N= 68+f1+f2		∈fu= 4-f1+f2

Given n= 120,

68+f1+f2= 120

F1+f2= 120-68

F1+f2= 52…….. (I)

Using the step – deviation method

$\bar{x}$ = A + $\frac{\in d}{n}$ ×h

$\bar{x}$ = 50 + $\frac{4-f1+f2}{120}$ ×20

50 = 50+$\frac{4-f1+f2}{6}$

50-50 = $\frac{4-f1+f2}{6}$

$0=\frac{4-f1+f2}{6}$

4–f1+f2= 0

F1–f2= 4 ……. (ii)

Adding (I) and (ii)

F1+f2= 52

F1-f2= 4

2f1 = 56

F1= 56/2

F1= 28

Put the value of f1 = 28 in 9i0 then

28+ f2= 52

F2 = 52-28

F2 = 24

Answer: f1= 28, f2 = 24

Example: Find missing frequency if A.M. = 33

Mark	0-10	10-20	20-30	30-40	40-50	50-60
No. of student	10	15	30	-	25	20

Solution:

Marks	Mid-Value (x)	f.(no. of student)	Fx	F4
0-10	5	10	-3	-30
10-20	15	15	-2	-30
20-30	25	30	-1	-30
30-40	35	X	0	0
40-50	45	25	1	25
50-60	55	20	2	40
	n= 100+x			$\in fu$= -25

Using step – devotion method: - $\bar{x}$ = A + $\frac{\in d}{n}$ ×h

For calculation the mean: - 33= 35 + $\frac{-25}{10+x}$ ×10

33-35= -250/100+x

-2 = -250/100+x

2x= 250-200

X= 50/2

x= 25

Answer = x= 25

13.8 WEIGHTED ARITHMETIC MEAN

A weighted arithmetic mean is very simple & easy to calculate and here only steps for calculation the weighted arithmetic mean:-

(i) Give an actual or expected weight of every value and item.

(ii) Multiply every value of x to weight (w) and find wx.

(iii) Add all value of wx, ∈wx.

(iv) Divide the value of ∈wx. by ∈W.

Formula: $\bar{x}_w$ = $\frac{\in wx}{\in w}$

Example: A candidate obtained the following percentage n B.Com examination: Statistics 60, Business org.:- 70, Account: - 80, Economic : 60, Banking: - 85 and Law: - 60. Find weighted arithmetic mean of marks if weight of these subject are 4,5,2,5,8,10

Solution: Weighted arithmetic Mean:

Subject	Mid-Value (x)	f	U=x-50/20 (A=50, h=20)
Statistics	60	4	240
Business org.	70	3	210
Account	80	2	160
Economic	60	5	300
Banking	85	8	680
Law	60	10	600
		∈w = 32	219

$\bar{x}_w = \frac{\epsilon wx}{\epsilon w}$

$= \frac{2190}{32}$

$\bar{x}_w$= 68.437

Answer- 68.437

Answer $\bar{x}_w$ = 68.437

Example: - Find the weighted arithmetic mean of the index number from the data given below:

Group	Index No. (x)	Wight
Food	100	5
Clothing	150	8
Fuel and light	120	4
House rent	200	6
Miscellaneous	125	3

Solution: Weighted arithmetic mean:

Group	Index No. (x)	Wight	wx
Food	100	5	500
Clothing	150	8	1200
Fuel and light	120	4	480
House rent	200	6	1200
Miscellaneous	125	3	375
		26	3755

$\bar{x}_w = \frac{\epsilon wx}{\epsilon w}$

$= \frac{3755}{26}$

$\bar{x}_w$= 144.423

$Answer = \bar{x}_w = 144.423$

13.9 PROPERTY OF ARITHMETIC MEAN

Some important and useful property of arithmetic mean.

Property 1: The algebraic sum of the deviations of the given set of oblations taken from the arithmetic mean is zero. That it $\in(x-\bar{x}) = 0$ or for a frequency distribution.

$\in f(X-\bar{x}) = 0$

Example: Verify property 1 for the following data:-

X= 2, 3,4,5,6

Solution :

X	$x-\bar{x}$
2 3 4 5 6	-2 -1 0 1 2
$\in x = 20$	$\in(x-\bar{x}) = 0$

$\bar{x} = \frac{\in x}{n} = \frac{20}{5} = 4$

Property 2: The sum of the squares of deviations of the given observation in minimum when taken from the arithmetic mean.

That is the sum $\in (x-A)^2$ which represent the sum of the squares of deviation of a given set of observation from an arbitrary value A. is minimum when $A = \bar{x}$.

Example: Verify problem 2 for following data

X= 2, 4, 6,8,10

Solution:

X	$(x-\bar{x})^2$	$(x-4)^2$	$(x-8)^2$
2	16	4	36
4	4	0	16
6	0	4	4
8	4	16	0
10	16	36	4
$\in X=30$	$\in (x-\bar{x})^2=40$	$\in (x-4)^2=60$	$\in (x-8)^2=60$

$\bar{x} = \frac{\in x}{n}$

$= \frac{30}{5}$

$= 6$

Here sum of the squared deviations taken from the arithmetic mean is equal to 40. If deviation is taken from any value 'A' different from the arithmetic mean, the sum of the squared deviations would be greater than 40. This is shown by taking two different values of A, 4 and 8.

Property 3: If each observation of a data is increased or decreased by a constant k. then the arithmetic mean of the new data also gets increased or decreased by k. that is , if y= x+k, then $\bar{y} = \bar{x} + k$.

Example: The arithmetic mean of a set of 5 observations 5,10,15,20 and 25 is

$=\frac{5+10+15+20+25}{5}$

$= 75/5$

$=15$

If each item is increased by 3, then the item is 8,13,18,23 and 28 and their sum

=n 8+13+18+23+28 = 90

Mean = 90/5= 18

New mean also gets increased by 3.

Property 4:

If each observation of a data is multiplied by a constant k, then the arithmetic mean of the data so obtained also gets multiplied by it.

Example: The arithmetic mean of a set of 5 observations 2, 4, 6,8,10 is

$=\frac{2+4+6+8+10}{5}=\frac{30}{5}$

=6

If each item is multiplied by 2 then the new item are 4, 8,12,16,20 and their arithmetic meaning is:

$\bar{x}=\frac{4+8+12+16+20}{5}$

= 60/2

=12

A new mean also get multiplied by 2.

Property 5: $\in x = n\bar{x}\left(\bar{x}=\frac{\in x}{n}\right)$

The lost property is quite useful to find corrected mean whenever one, two or more of the observations were wrongly copied down.

For example:- Suppose we have computed the mean $\bar{x}$ of n observattion and later on it is found that two observation x_1 and x_2 were wrongly copied down as x_1^1 and x_2^1. It si important to correct the mean by replacing the wrong observation by the correct ones.

By the help of using properly 5, we can first obtain the uncorrected sum of the observations which is given by $n\bar{x}$, and subtract the wrong observation x_1^1 and x_2^1, and add the corresponding correct observations x_1 and x_2 to get the corrected sum.

$n\bar{x}= (x_1^1 + x^1_2) – (x_1 + x_2)$+corrected Σx

Dividing this by n, we get the corrected mean.

In general if x observation are misread as x^1_1 , x^1_2,……..x^1_n, while correct observation are (x_1 + x_2……..x_n then corrected sum of observations is given by

$n\bar{x}= (x^1_1 + x^1_2$+……. $x^1_n) + (x_1 + x_2$……..$x_n)$

Example: The mean marks of 100 students were found to be 40. Later on it was discovered that a score 53 was misread as 83. Find the correct mean corroding to the correct score.

Solution:

N=100 and $\bar{x}$= 40

$\bar{x}=\frac{\in x}{n}$

$\therefore \in x = n\bar{x}$

= 100×40

= 4000

But this is not the correct $\in x$

Correct $\in x$ = 4000-wrong score + correct score

= 4000-83+53

Correct $\bar{x} = \frac{Corrected \in x}{n}$

$= \frac{3970}{100}$

Correct $\bar{x}$= 39.7

Answer: Correct $\bar{x}$= 39.7

13.10 MERIT'S OF ARITHMETIC MEAN

(i) Arithmetic mean is rigidly defined.

(ii) It's very easy to calculate and simple to understand.

(iii) It's based on all observations.

(iv) It is suitable for further mathematical treatment.

(v) Of all the averages, arithmetic mean is affected least by fluctuations of sampling.

(vi) An arithmetic mean always fixed, for use of different formula in an observation.

13.11 LIMITATIONS OF ARITHMETIC MEAN

(i) Mean are not presented in the series. For example a mean of the 4, 8, 6 and 2 are 5 but are not visible in series.

(ii) In some case an arithmetic mean provides the confusing results. For example mean of series 3, 5, 10, 13 and 7.75, so it's not representing the actual mean.

(iii) An arithmetic mean affected the extreme series. For example:- If income A,B,C, and D are 10,15,20 and 15then if E are enter in A,B,C, D, and his income are 300 then it's mean are 72, but income of A,B,C and D are less than 72.

Properties of Arithmetic mean:

(i) It is very simple and easy for calculation and interpretation.

(ii) Arithmetic mean is based on advanced statistical tools.

(iii) It is very useful for comparative study of two or more series.

(iv) Mean are determined in unarranged series.

(v) Mean are easily calculated if only the aggregate of scores and their number are given.

Uses of weighted arithmetic mean:

(i) It's useful for finding grouped mean of different class Example: - income, expenditure, wages and salary etc.

(ii) Its use to find the mean of ratio, rate and percentage.

13.12 QUESTIONS

Question 1: State two important objects of measures of central value.

Question 2: what do you mean by arithmetic mean? Discuss its merits and demerits also state its important properties.

Question 3: State the algebraic properties of arithmetic mea with examples.

Question 4: The following table gives the monthly income of 10 families in a city.

Family No.	1	2	3	4	5	6	7	8	9	10
Income (in Rs.)	7800	7600	7500	8400	9900	11000	0	8100	10500	9500

Calculate the arithmetic mean of incomes.
(Answer: 8650)
Question 5: The following table gives the monthly income of 10 families in a city.

Roll No.	1	2	3	4	5	6	7	8	9	10
Mark obtain	43	48	65	60	48	59	78	31	57	37

Calculate the arithmetic mean.

(Answer: 52.6)

Question 6: Calculate the arithmetic mean for the following discrete frequency distribution.

X	10	15	20	25	30
y	2	5	4	9	5

(Answer: 22)

Question 7: Calculate the arithmetic mean for the following data:

Height in inches	58	60	61	62	63	64	65	66	68	70
No. of Person	4	6	5	10	20	22	24	6	2	1

(Answer:63.47 inches)

Question 8: Calculate the arithmetic mean for the following data:

Weight in pound	100	110	1120	130
No. of Person	15	20	25	30

(Answer: 120 lks)

Question 9:- The daily earnings (in rupees) of employee working n a daily basis in a firm are:

Daily earning	60	80	100	120	160	180	200
No. of employee	5	8	12	22	10	7	6

(Answer: Rs. 126.286)

Question 10: Computer the arithmetic mean from the following frequency distribution.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of students	6	5	8	15	7	6	3

(Answer: 33.4)

Question 11: Compute the arithmetic mean from the following frequency distribution.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of students	8	12	10	8	3	2	7

(Answer: 29)

Question 12: Compute the arithmetic mean from the following frequency distribution.

Marks	20-29	30-39	40-49	50-59	60-69	70-79
No. of students	5	11	18	22	16	8

(Answer: 51.6)

Question 13: Find the arithmetic mean of the weekly income from the following frequency distribution.

Weekly income (in Rs.)	10-15	15-20	20-25	25-30	30-35	35-40
No. of workers	200	700	900	800	600	400

(Answer: 31)

Question 14: Calculate the mean for the following data:

Marks (more than)	0	10	20	30	40	50	60	70
No. of students	100	90	75	50	25	15	5	0

(Answer: 31)

Question 15: Calculate the arithmetic mean for the following data:

Temp. ºC	No. of days	Temp. ºC	No. of days
–40 to –30	10	0 to10	65
–30 to –20	28	10 to20	180
–20 to –10	30	20 to30	10
–10 to 0	42		

(Answer: 4.288°C)

Question 16: Calculate the mean from the following data:

Mark (less than)	80	70	60	50	40	30	20	10
No. of students	100	90	80	60	32	20	13	5

(Answer: 46.2)

Question 17: For the data given below, find the missing frequency if the arithmetic mean is 28.

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No. of students	12	18	27	-	17	6

(Answer: 20)

Question 18: The mean of the following frequency distribution is 33. But the frequency f1, and f2 in class 10-20 and 30-40 are missing. Find the missing frequencies:

Class	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	5	F1	25	F2	20	10

(Answer: f1= 10, f2= 30)

Question 19: The mean weight of 100 students (boys and girls) in a class is 50 kg. The mean weight of boy students is 52kg. And that of girl students is 42 kg. Find the number of boys and girls in that class.

(Answer: Boys=80, Girls = 20)

Question 20: Find the missing information from the following data.

	Group I	Group II	Group III	Combined
Number	100	120	80	300
Mean	55	?	25	45

(Answer: 50)

Question 21: A candidate obtained the following percentage of marks in different subjects in the final examination: English=60, Statics= 75, Cost Accounting =63, Economics=59, Income Tax=55, Find the candidates weighted arithmetic mean if weights 1, 2,1,3,3 representively are allotted to the subjects.

(Answer: 61.5)

Question 22: Calculate the simple and weighted arithmetic mean from the following data:

Designation	Monthly Salary in (Rs.)	Strength of the cadre
Class I officers	15000	10
Class II officers	8000	20
Subordinate staff	5000	70
Clerical staff	2500	100
Lower staff	1000	150

(Answer: - $\bar{x}$ = Rs. 6300, $\bar{x}w$= Rs. 3028.57)

Question 23: Calculate the weighted arithmetic mean from the following data.

Group	Index No. (x)	Wight
Food	370	71
Clothing	423	3
Fuel and light	469	9
House rent	110	7
Miscellaneous	279	10

(Answer: 353.20)

13.13 SUMMARY

Averages are very important in statics and are always denoted by absolute unit not in percentage. There are short cut method and direct method to find averages. There are five properties of averages. Advantages of averages are easy to calculate, based on all observation, can be further mathematical exertion. It is a fixed value.

UNIT-14 MEDIAN

14.1 Objective
14.2 What is median?
14.3 Merits of Median
14.4 Limitation of Median
14.5 Individual Series
14.6 Discrete series
14.7 Median in class interval
14.8 Median in unequal class interval
14.9 Zero Frequency in Continuous Series
14.10 Median in Inclusive Series
14.11 Descending class Interval
14.12 Determination of unknown frequency
14.13 Median in continues series
14.14 Quartiles
14.14.1 Discrete Series
14.14.2 Continues Series
14.15 Dectiles
14.15.1 Discrete Series
14.115.2 Continues Series
14.16 Percentile
14.16.1 Discrete Series
14.16.2 Continuous Series
14.17 Graphical determination of median and other partition values.
14.18 Practice Questions
14.19 Summary

14.1 OBJECTIVE

After studying this unit a student should understand about

- Meaning of median its merits and limitation

- Median in individual, discrete and continues series
- Graphical determination of median and

14.2 WHAT IS MEDIAN?

"Median is a made or central value of a series after arranging the series in ascending or descending order."

A series is of three types (I) Individual series (ii0 Discrete (iii) Continuous series

When a series is individual then firstly we arranged the series in ascending or descending order, and then checked that number of series are odd or even. If number of item is odd then made or central value is median, but if number of item is even then odd the two mid value and divide by '2' then the divided value is median.

Median is simple calculation and very easy to formula. It's also helpful to find out the made value of any series without any difficulty.

14.3 MERITS OF MEDIAN

Median had following merits:-

(i) It's easy to calculate and simple to understand.

(ii) It's very less time taking.

(iii) Calculation result of median is fixed and clear.

(iv) Median is essaying to calculate of incomplete figure if number of item and human of median are given.

(v) Median are presented by point diagram

14.4 LIMITATIONS OF MEDIAN

1. No. use of median in geometric problems
2. Find the median firstly arrange the number in ascending and descending order and this take time.
3. If number of item in any series is small then median ceases to be typical.
4. In continues series the calculation of median are difficult.

In different series we use formula for calculation of median.

Median is calculated in three different series. (I) Individual (ii) Discrete (iii) Continuous

14.5 INDIVIDUAL SERIES

Where only one observation is given then it's called Individual series.

Formula: When observation is odd

M= $\left(\frac{n+1}{n}\right)^{Th}$ item

M= median

N= no. of item.

When observation is even

Mino. $\left(\frac{n+1}{n}\right)^{th}$ or $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n+1}{n}\right)^{th}$

Median Value of $\left(\frac{n}{2}\right)^{th}$ + Vol. of $\left(\frac{n+1}{2}\right)^{2}$

Example 1: Calculate median from the following data:-

Roll No.	1	2	3	4	5	6	7	8	9	10	11
Marks.	4	10	3	8	12	20	5	9	16	14	18

Solution: Marks arranged in ascending order

3,4,5,8,9,10,12,14,16,18,20

Median = $\left(\frac{n+1}{n}\right)^{th}$ item

= $\left(\frac{11+1}{2}\right)^{th}$ item

12/2 th item.

Median = 6th item

= 6th item = 10

Answer = 10 marks.

Example 2: Determine median for following data:-

20, 25, 23, 25, 27, 40, 23, 15, 25

Solution: Arranged in ascending order

15, 20, 23, 23, 25, 25, 25, 27, 40

N=9

Median = $\left(\frac{n+1}{n}\right)^{th}$ item

= $\left(\frac{9+1}{2}\right)^{th}$ item

= $10/2^{th}$ item

= 5th item

A 5th item are 25

Answer: 25

Example3: Mark's obtained by 8 students in a test in economics is given below:

10, 8, 24, 20, 5, 12, 18, 14

Calculate median marks

Solution: Arranged in ascending order

5,8,10,12,14,18,20,20

N=8

Median = $\left(\frac{n+1}{n}\right)^{th}$ item

= $\left(\frac{8+1}{2}\right)^{th}$ item

= $\frac{9}{2}^{th}$ item

=4.5^{th} item

Average of 4th and 5th item = $\frac{12+14}{2}$

=$\frac{26}{2}$

=13

Median = 13, Answer:- 13

Example 4: Calculate the median

5, 28, 6, 20, 30, 12, 4, 10, 14

Solution: Arranged in ascending order

2, 4, 5, 6, 8, 10, 12, 14, 20, 30

N=10

Median = $\left(\frac{n+1}{n}\right)^{th}$ item

= $\left(\frac{10+1}{2}\right)^{th}$ item

= $\frac{11}{2}^{th}$ item

=5.5^{th} item

Average of 4th and 5th item = $\frac{8+10}{2}$

$=\frac{18}{2}$

=9

Median = 9,

Answer: 9

14.6 DISCRITE SERIES

In discrete series we use two series first series are totally depended to second series. Second series are represented by the frequency of first series.

For calculation of median first odd a frequency one by one and find cumulative frequency (cf) and then use the formula.

Formula:- M= $\left(\frac{n+1}{n}\right)^{th}$ item

Example 1:- Mark's obtained by 35 students in a test in banking us given below:-

Marks:	2	6	4	8	10	5
No. of student:	8	4	3	8	2	10

Calculate median marks.

Solution:

Marks (m)	No. of student (f)	Cumulative frequency (cf)
2	8	8
4	3	11
5	10	21
6	4	25
8	8	33
10	2	35
	35	

M= $\left(\frac{n+1}{n}\right)^{th}$ item

= $\left(\frac{35+1}{2}\right)^{th}$ item

= $\frac{36^{th}}{2}$ item

=18th item

A 18th item (student) are comes in cumulative frequency 11 to 21, and here seared 5 marks each

Median= 5

Example 2: Calculate the median:-

X : 11, 14, 12, 18, 13, 15, 19, 17, 16, 20

F : 2, 14, 10, 15, 8, 6, 16, 20, 26, 18

Solution : Median

X	F	cf
11	2	2
14	10	12
12	8	20
18	14	34
13	6	40
15	26	66
19	20	86
17	15	101
16	16	117
20	18	135
	135	

M= $\left(\frac{n+1}{n}\right)^{th}$ item

= $\left(\frac{135+1}{2}\right)^{th}$ item

= $\frac{136_{th}}{2}$ item

=68th item

A 68th item are comes in cf 86 and its Value (x) are 17. M =17

Answer = Median = 17

Example 3:- Calculate of Median

X : 2	5	3	6	8	10
F : 4	10	8	0	0	6

Solution :

X	F	cf
2	9	4
5	8	17
3	10	27
6	0	27
8	0	27
10	6	33
	33	

M= $\left(\frac{n+1}{n}\right)^{th}$ item

= $\left(\frac{33+1}{2}\right)^{th}$ item

= $\frac{34^{th}}{2}$ item

=17th item

A 17th item are come in cf second thermo and its value are 3Answer = Median =3

Solution 4 : Find out median from the following data:-

Marks in test :	2	3	4	5	6	7	8	9	10
No. of student:	9	6	2	2	2	4	3	3	3

Solution : Median

X	f	cf
2	9	9
3	6	15
4	2	17
5	2	19
6	2	21
7	4	25
8	3	28
9	3	31
10	3	34
	34	

M= $\left(\frac{n+1}{n}\right)^{th}$ item

= $\left(\frac{34+1}{2}\right)^{th}$ item

= $\frac{35^{th}}{2}$ item

=17.5^{th} item

17.5^{th} item are come after 17th item, so here tack the both value.

Median = $\frac{4+5}{2}$

= 9/2

= 4.5

Answer = Median = 4.5

14.7 MEDIAN IN CLASS INTERVALS

Classes intervals are divide many parts:-

Median in unequal class intervals.

1. Zero in continues series.
2. Median in inclusive series
3. Descending class intervals
4. Determination of unknown frequency
5. Median in continuous series

14.8 MEDIAN IN UNEQUAL CLASS INTERVALS

Example: Find out the median from the following frequency:

Class	0-3	3-6	6-10	10-15	15-25	25-30
F	10	15	30	15	10	6

Solution :

Class	f	cf
0-3	10	10
3-6	15	25
6-10	30	55
10-15	15	70
15-25	10	80
25-30	6	86

Formula: M.No. $\left(\frac{n}{2}\right)$, Median = $L_1 + \frac{i}{f}$ (m-c)

Where:

L_1= Lower limit

i=Class interval of median class

f= Frequency median class

m= Median No. of cf

C = Preceding value of median No.

M.No. $\left(\frac{n}{2}\right)^{th}$ item

$\frac{86^{th}}{2}$ item

= 43th item

A 43th item is come in 55 cf and his class intervals are 6-10, 50.

Median = $L_1 + \frac{i}{f}$ (m-c)

= $6 + \frac{4}{30}$ (43-25)

= $6 + \frac{4 \times 18}{30}$

= 6+2.4

= 8.4

Answer = M= 8.4

14.9 ZERO FREQUENCY IN CONTINUOUS SERIES

Example : Find the median of following data :

Class	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
f	3	4	6	12	0	14	6	5

Solution :

Class	f	cf
0-5	3	3
5-10	4	7
10-15	6	13
15-20	12	25
20-25	0	25
25-30	14	39
30-35	6	45
35-40	5	50

M. No. $\left(\frac{n}{2}\right)^{th}$ item

= $\left(\frac{50}{2}\right)^{th}$ item

= 25th item

A 25th item are come two time so here end the 0 frequency and divide it's class interval in two equal part (2.5) next and previous class.

So, new classes are:

Class	f	cf
0-5	3	3
5-10	4	7
10-15	6	13
15-22.5	12	25
22.5-30	14	39
30-35	6	45
35-40	5	50

M. No. = 25 are came in cf 25 and it's median class are 15-22.5

Median = $L_1 + \frac{i}{f}$ (m-c)

= 15+ $\frac{7}{12}$ (25-13)

= 6+ $\frac{7.5}{12} \times 12$

= 15+7.5

= 22.5

Answer = M=22.5

14.10 MEDIAN IN INCLUSIVE SERIES

Example: Find the median following data :

Mark	10-14	15-19	20-24	25-29	30-34
No. of Student	5	8	15	10	4

Solution :

Class interval	f	cf
9.5-14.5	5	5
14.5-19.5	8	13
19.5-24.5	15	28
24.5-29.5	10	38
29.5-34.5	4	12

M. No. $\left(\frac{n}{2}\right)^{th}$ item

= $\left(\frac{42}{2}\right)^{th}$ item

= 21th item

Median = L_1+ $\frac{i}{f}$ (m-c)

= 19.5+ $\frac{7}{12}$ (21-13)

= 6+ $\frac{5}{15} \times 8$

= 19.5+2.66

= 22.166

Answer = M=22.166

14.11 DESCENDING CLASS INTERVALS

Example :

Number	50-40	40-30	30-20	20-10	10-0
No. of student	10	8	6	4	2

Find the median.

Solution :

Class	f	cf
50-40	10	10
40-30	8	8
30-20	6	24
20-10	4	28
10-0	2	30

M. No. $\left(\frac{n}{2}\right)^{th}$ item

= $\left(\frac{30}{2}\right)^{th}$ item

= 15th item

M.No. 15 are present in of 24,50

Median = $L_2 + \frac{i}{f}$ (m-c)

= 20- $\frac{10}{6}$ (15-8)

= 20- $\frac{10}{6} \times 7$

= 20-11.66

= 8.334

Answer = M=8.334

14.12 DETERMINATION OF UNKNOWN FREQUENCY

Example: Find the unknown frequency, if sum of frequency = 100 median = 50

Example: 0-20 20-40 40-60 60-80 80-100

No. of family: 14 ? 27 ? 15

Solution :

Expenses	f	cf
0-20	14	14
20-40	F1	14+f1
40-60	27	41+f1
60-80	F2	41+f1+f2
80-100	15	56+f1+f2

N= 56+f1+f2

(i) 56+f1+f2 = 100

F1+f2 = 100-56

F1+f2= 44……..(I)

(ii) M.N. = . $\left(\frac{n}{2}\right)$

=100/2

=50th item

According to question median are given 50,80, median class = 40-60,

Median = $L_2+ \frac{i}{f}$ (m-c)

= 40+$\frac{20}{17}${50(14+f1)}

50-40= $\frac{20}{17}$ {36-f1}

10= $\frac{20}{17}$ {36-f1}

$\frac{10\times27}{20}$= 36-f1

$\frac{270}{20}$= 36-f1

13.5= 36-f1

F1 = 36-13.5

F1= 22.5

Because frequency are always whole number, So

22.5= 23

F1+f2 = 44

23+f2= 44

F2 = 44-23

F2 = 21

Answer: Unknown frequency is

20-40=23

60-80= 21

14.13 MEDIAN IN CONTINUOUS SERIES.

Example : Find out the Median :

Marks	5-10	10-15	15-20	20-25	25-30	30-35	35-40
No. of student	6	8	4	10	16	20	2

Solution :

Class	f	cf
5-10	6	6
10-15	8	14
15-20	4	18
20-25	10	28
25-30	16	44
30-35	20	64
35-40	2	66

M. No. $\left(\frac{n}{2}\right)^{th}$ item

= $\left(\frac{66}{2}\right)^{th}$ item

= 33th item

Median = $L_2 + \frac{i}{f}$ (m-c)

= $25+\frac{5}{16}$ (33-28)

= 25+1.5625

= 26.56

Median = 26.56

Answer = M=26.56

14.14 QUARTILES

In Quartiles data are divided in four equal parts for w1 and we gain four equal parts of data.

In quartiles we use two series first discrete and second continuous series.

14.14.(I) Discrete Series:- In discrete series following formula are used for quartile:-

First Quartile (Q_1) = $\left(\frac{n+1}{4}\right)^{th}$ item

Second Quartile (Q_3) = $\left(\frac{n+1}{4}\right) \times 3^{th}$ item

Example 1: 8,10,2,6,12,18,20,4,9,7,3,22,27,29,24,30,26 find the Q1, Q2, Q3 and Q6

Solution: Firstly arrange in ascending order

2,3,4,6,8,8,9,10,12,18,20,22,24,26,27,28,29,30

$Q_1 = \left(\frac{n+1}{4}\right)$th item $= \frac{11+1}{4}$ =12/4 3th item $Q_1 = 4$	$Q_3 = \left(\frac{n+1}{4}\right) \times 3^{th}$ item $= \frac{11+1}{4} \times 3$ = 36/4 $=9^{th}$ item $Q_3 = 12$
$Q_4 = \left(\frac{n+1}{4}\right) \times 4^{th}$ item $= \frac{11+1}{4} \times 4$ =48/4 $=12^{th}$ item $Q_4 = 22$	$Q_6 = \left(\frac{n+1}{4}\right) \times 6^{th}$ item $= \frac{11+1}{4} \times 6$ = 72/4 $=18^{th}$ item $Q_6 = 30$

Answer: Q_1=4, Q_3= 12, Q_4= 22, Q_6 = 30

(ii) Continuous series: formulas for continuous Series are

$Q_1 = (n/4)^{th}$ item and for interpolation

$Q_1 = L_1 + \frac{l2-l1}{f1}(q_1\text{-}c)$

Q_3 = (n/4) × 3th item

= L_1+ $\frac{l2-l1}{f1}$ (q_3-c)

Example: Find the Q_1, $Q_3$and Q_4

Class	0-5	5-10	10-15	15-20	20-25	25-30
	4	6	1	3	2	3

Solution:

Class	f	cf
0-5	4	4
5-10	6	10
10-15	10	20
15-20	3	23
20-25	2	25
25-30	3	28

Q_1 = $\left(\frac{n+1}{4}\right)$th item

= $\frac{28}{4}$

=7th item

7th are come in cf 10, so middle class are 5-10

Q_1 = L_1+ $\frac{l2-l1}{f1}$ (q_1-c)

= 5+ $\frac{10-5}{6}$ (7-4)

= 5+$\frac{5}{6}$×3

= 5+2.5

Q_1 = 7.5

Q_3 = $\left(\frac{n+1}{4}\right)$ ×3th item

= $\frac{28}{4}$ ×3

= 7×3rd item

=21th item

$Q_3 = 21$

According to cf. a middle class of 21th item are 15-20, 50

$Q_3 = L_1 + \frac{l2-l1}{f1}(q_3\text{-}c)$

$= 15 + \frac{20-15}{3}(21\text{-}20)$

$= 15 + \frac{5}{3} \times 1$

$= 15 + 1.666$

$Q_3 = 16.666$

$Q_4 = \left(\frac{n+1}{4}\right) \times 4^{th}$ item

$= \frac{28}{4} \times 4$

$= 7 \times 4^{th}$ item

$= 28^{th}$ item

$Q_4 = 28$

According to cf. a middle class of 28th item are 25-30, 50

$Q_4 = L_1 + \frac{l2-l1}{f1}(q_4\text{-}c)$

$= 25 + \frac{30-25}{3}(28\text{-}25)$

$= 25 + \frac{5}{3} \times 3$

$Q_4 = 30$

Answer: $Q_1 = 7.5$, $Q_2 = 16.666$, $Q_3 = 30$

DECILES

In deciles the No. of data are divides in 10 equal parts. In deciles here also two series (I) discrete (ii) Continuous.

14.15 (I) DISCRETE SERIES

$D_1 = \left(\frac{n}{10}\right)^{th}$

$D_3 = \left(\frac{n}{10}\right) \times 3^{th}$ item

Example : Find D_1 and D_3.

8,10,4,7,15,2,14,5,3

$D_1 = \left(\frac{n}{10}\right)^{th}$

$D_1 = \left(\frac{9+1}{10}\right)^{th}$

$= 10/10=1^{th}$

$D_1 = 2$

$D_3 = \left(\frac{n}{10}\right) \times 3^{th}$ item

$= \left(\frac{9+1}{10}\right) \times 3$

$= \frac{10\times 3th}{10}$

$= 3^{th}$

$D_3= 4$

Answer: $D_1 = 2$ and $D_3= 4$

Example: Find D_1 and D_6

82,56,65,90,50,120,75,75,80,82,90,120,130

$D_1 = \left(\frac{n}{10}\right)^{th}$

$= \left(\frac{10+1}{10}\right)^{th}$

$= 11/10^{th}$

$= 1.1^{th}$

$= 1^{th}$ value + .10 × difference d/w 2th and 1th value.

= 50+.10 ×(56-56)

= 50+.10×6

= 50+.60

$D_1= 50.60$

$D_6 = \left(\frac{n}{10}\right) \times 6^{th}$

$D_6 = \left(\frac{10+1}{10}\right) \times 6$

$= 6.6^{th}$ value

$= 6^{th}$ value + .60 × difference d/w 7th and 6th value.

= 80+.60 ×(82-80)

= 50+.10×2

= 80+1.20

D_6= 81.20

Answer: D_1= 50.60 and D_6= 81.20

14.15 (II) CONTINUOUS SERIES

Example: Find D_1 and D_6:

Class: 0-5 5-10 10-15 15-20

6 8 4 2

Solution :

Class	f	cf
0-5	6	6
5-10	8	14
10-15	4	18
15-20	2	20

$D_1 = \left(\frac{n}{10}\right)^{th}$

$= \left(\frac{20}{10}\right)^{th} = 2^{th}$

$D_1 = L_1 + \frac{l2-l1}{f1}(d_1\text{-}c)$

$= 0+\frac{5-0}{6}(6\text{-}0)$

$= 0+\frac{5}{6}\times 6$

$= 0+\frac{30}{6} = 0+5$

d_1= 5

$D_6 = \left(\frac{n}{10}\right)^{th}$

$= \left(\frac{20}{10}\right) \times 6^{th}$

$\frac{120^{th}}{10} = 12^{th}$

$D_6 = L_1 + \frac{l2-l1}{f1}(d_6\text{-}c)$

$= 5+\frac{5}{8}(12\text{-}6)$

$= 5+ \frac{5}{8} \times 6$

$= 5+3.75$

$D_6= 8.75$

Answer: $d_1= 5$, $D_6= 8.75$

14.16 PERCENTILE

Percentile are divided 100 equal part in percentile no. of item's are divided by 100. A percentile defined in series (I) discrete (ii) Continuous.

14.16 (I) DISCRETE SERIES

$P_1 = \left(\frac{n+1}{100}\right)^{th}$ item

$P_3 = \left(\frac{n+1}{100}\right) \times 3^{th}$ item

Example:- Find P_{82}

82,56,90,100,40,120,75,115,110,70

Solution:- Arrange in ascending order:-

40,56,70,75,82,90,100,110,115,120

$P_{82} = \left(\frac{n+1}{100}\right) \times 82^{th}$ item

$= \left(\frac{10+1}{100}\right) \times 82^{th}$ item

$= \frac{11\times82}{100}$

9.02^{th} value

$=9.02^{th}$ value + .02 × diff bow 10th and 9th value

$= 115+0.2(120-115)$

$= 115+0.2\times0.1$

$P_{82}= 115.1$

Answer = $P_{82}= 115.1$

14.16 (II) CONTINUES SERIES

Example: Find the P_{500}

Class	0-10	10-20	20-30	30-40	40-50
No. of Student	480	8500	10,000	55,000	65,000

Solution:

Class	f	cf
0-10	480	480
10-20	8500	8980
20-30	10,000	18980
30-40	55,000	93,980
40-50	65,000	80,480

$P_{500} = \left(\frac{n}{100}\right)^{th}$ item

$= \left(\frac{80480}{100}\right)^{th}$

$= 804.8^{th}$

$= L_1 + \frac{l2-l1}{f1} (P_{500}\text{-}c)$

$= 10 + \frac{20-10}{8500} (805\text{-}480)$

$= 10 + \frac{20-10}{8500} \times 325$

$= 10 + 0.382$

$P_{500} = 10.382$

Answer = $P_{500} = 10.382$

14.17 Graphic determination of median and other Partition Values

The value of median and other partition value like Quartile deciles and percentile are also be located graphically with the help of a curve.

Example: From the following data find the median.

Class	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60
Fre.	6	9	13	23	19	15	9	6

Solution :

Class	f	cf
20-25	6	6
25-30	9	15
30-35	13	28
35-40	23	51
40-45	19	70
45-50	15	85
50-55	9	94
55-60	6	100

N= 100

Median = n/2th item

= 100/2

=50th item

= $L_1 + \frac{i}{f}$ (M-C)

= $35+\frac{5}{23} \times 22$

= 35+4.782

=39.782

Example 2: Find Q_1,Q_3, P_{40} and d_6 by graphical.

Class	10-15	15-20	20-25	25-30	30-35	35-40
Fre.	5	10	15	20	10	5

Solution:

Class	f	cf
10-15	5	5
15-20	10	15
20-25	15	30
25-30	20	50
30-35	10	60
35-40	5	65

$Q_1 = \frac{n}{4}^{th}$, 65/4 = 16.25^{th}

Q_1 lies in the class 20-25

$Q_1 = L_1 + \frac{l2-l1}{f1}(Q_1\text{-}c)$

$= 20 + \frac{25-20}{15} \times (16.25\text{-}15)$

$= 20 + \frac{5}{15} \times 1.25$

= 20+.416

= 20.416

= 20.41

$Q_1 = 20$

$Q_3 = \left(\frac{n}{4}\right) \times 3^{th}$, $\frac{65\times3}{4}$= 48.75

Q_3 = lies in the class 25-30

$Q_3 = L_1 + \frac{l2-l1}{f1}(Q_3\text{-}c)$

$= 20 + \frac{30-25}{15} \times (49\text{-}30)$

= 25+4.75

= 29.79

$Q_3 = 30$

$P_{40} = \left(\frac{n}{100}\right) \times 40^{th} = \left(\frac{65}{100}\right) = 26^{th}$

P_{40} lies in the class 20-25

$P_{40} = L_1 + \frac{l2-l1}{f1}(P_{40}\text{-}c)$

$= 20 + \frac{25-20}{15} \times (30\text{-}15)$

$= 20 + \frac{5}{15} \times 15$

=20+5

P_{40} =25

$D_6 = \left(\frac{n}{10}\right) \times 6^{th}$, $\frac{65\times6}{10}$= 39th

D_6 lies in the class 25-30

$D_6 = L_1 + \frac{l2-l1}{f1}(D_6\text{-}c)$

$= 25 + \frac{30-25}{20} \times (39\text{-}30)$

$= 25+ \frac{5}{20} \times 9$

$= 25+2.25$

$D_6 = 27.25$

14.18 PRACTICE QUESTION

Question1: The marks obtained by 9 students in a test are 15,20,15,45,18,7,10,38,12 find the median marks.

(Answer = 18)

Question 2: Find the value of median form the following data:

88,72,33,29,70,86,54,91,61,57

(Answer: 65.5)

Question 3: Find the median farm the following data:-

(i) 8,3,11,7,12,6,9

(ii) 94,33,86,68,62,80,48,70

(Answer: (I) 8 (ii) 69)

Question 4: Find the median from the following data:-

(i) X: 1,5,7,11,8

F: 15,20,15,18,12

(ii) X: 85,90,95,100,105,110,115,120,125

F: 15,18,25,19,15,7,13,16,22

(Answer:I) 8 (ii) 100)

Question 5: Calculate the median from the following data:-

Daily earning (Rs.) : 100,150,80,200,250,180

No. of employees: 24,26,16,20,6,30

(Answer: Rs. 150)

Question6: The heights in cm. of 30 persons are given below:-

133,125,137,129,130,130,131,125,137,147,128,127,147,141,148,149,145, 148,139,125,145,134,129,127,147,132,128,130,131.

Prepare the frequency distribution table and find the median

(Answer:133)

Question 7: The mark obtains by 100 students in a certain examination are given below:

Class	0-10	10-20	20-30	30-40	40-50	50-60
No. of Student	12	16	18	25	20	9

(Answer: 31.6)

Question 8: The marks obtained by 200 students in a certain examination are given below:

Marks	0-25	25-50	50-75	75-100
No. of student	30	50	80	40

(Answer: 56.25)

Question 9:- Find out the median weight from the following data of the weights of 60 persons:

Weights (in kg)	30-34	35-39	40-44	45-49	50-54	55-59	60-64
No. of person	3	5	12	18	14	6	2

(Answer: 47.28)

Question 10: Find out the median from the following data:

Wages per week (inch)	50-59	6069	70-79	80-89	90-99	100-109	110-119
No. of Workers	15	40	50	60	45	40	15

(Answer: 84.08)

Question 11:Find out the value median for the following data:

Weight (in lbs)	60-64	65-74	75-79	80-84	85-94	95-99	100-104
No. of person	10	39	35	20	26	9	6

(Answer: 77.86)

Question 12: Following is the distribution of marks in law obtained by 50 students:

Marks (more than)	0	10	20	30	40	50
No. of students	50	46	40	20	10	3

(Answer: 27.5)

Question 13: Calculate median for the following distribution:

Production per day (in tons)	21-22	23-24	25-26	27-28	29-30
No. of days	7	13	22	10	8

(Answer: 25.41)

Question 14: Following is the distribution of marks in statistics obtained by 50 students:

Marks (less than)	10	20	30	40	50	60
No. of students	4	10	30	40	47	50

(Answer: 27.5)

Question 15: Find the missing from the following distribution given that the median is 24.

Class Interval	0-10	10-20	20-30	30-40	40-50
Frequency	5	25	---------	18	7

(Answer : 25)

Question 16:- An incomplete distribution is given below:

Class Interval	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	10	----------	30	----------	25	20

Total : 125

Find out missing frequencies if median Value is 33 (Answer:- 15,25)

Question 17: An incomplete distribution is given below:

Class Interval	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	10	----------	30	----------	25	20

Total : 100

Find out missing frequencies if median Value is 30 (Answer:- 15,10)

Question 18: Calculate median from the following data:

Marks	0-10	10-30	30-60	60-80	80-90
No. of students	5	15	30	8	2

(Answer : 5575)

Question 19: Find the median from the following data:

Income (Rs.) below	0-2000	2000-4000	4000-6000	6000-8000	8000-10000	Abv 10000
	25	40	80	75	20	16

(Answer: 5575)

Question 20: Find the median from the following data.

Mid-Value	105	115	125	135	145	155	165	175
Frequency	4	6	20	32	33	17	8	2

(Answer: 139.68)

Question 21: Draw a “less than” and “more than” q gives curves for the following data and find the median value:

Weights (in kg)	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of person	8	12	10	8	3	2	7

(Answer : 25)

Question 22: Draw a “less than” and “more than” q give curves for the following data and find the median value also verify your result by actual calculation :

Weights (in kg)	0-10	10-20	20-30	30-40	40-50	50-60
No. of person	4	6	20	10	7	3

(Answer : 27.5)

Question 23: The marks obtained by 15 students in a test are given below, find the values Q_1,Q_2,D_5, D_7 ,P_{65} and P_{90} for the data:-

Roll No. : 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15

Marks obtain: 13,17,7,11,18,3,12,12,6,2,19,12,10,11,14

(Answer: Q_1=15, Q_2= 45, Q_3= 14, D_5=12, D_7 = 13.2, P_{65}= 12.4, P_{90}= 18.4)

Question 24: From the following data compute the values of Q_1,Q_3,D_4, P_{60} and P_{82} .

X	1	15	25	35	45	55	65
y	8	12	10	8	3	2	7

(Answer: Q_1=15, Q_3= 45, D_4=25, P_{60}= 35, P_{82}= 55)

Question 25: Find the 45th and 65th percentiles for following data on marks obtained by 40 students:

Marks	0-5	5-10	10-15	15-20	20-25	25-30
No. of students	5	7	10	8	6	4

(Answer: P_{45}=13, P_{65}= 17.5)

Question 26: Find the values of lower and upper quartiles 7th docile and 90th percentile from the following data:

Marks	Below 10	10-20	20-40	40-60	60-80	Above 80
No. of students	8	10	22	25	10	5

(Answer: Q_1=21.82, Q_3= 56, D_7=52.8, P_{90}=74)

14.19 SUMMARY

Median is a mid or central value. A number series is of three types, individual, discrete and combination series, Median is simple to calculate and less time taking.

Disadvantage of median is not used in geometric problems, in continuous series median is difficult to calculate.

Formula of median = M= $\left(\frac{n+1}{2}\right)^{\text{Th}}$ item

UNIT-15 MODE

15.1 Objective

15.2 Introduction

15.3 Limitation of mode.

15.4 Merits of mode.

15.5 Mode in individual series

15.6 Mode in discrete series

15.7 Mode in continues series

15.8 Graphical determination of mode

15.9 Empirical representation between mean, median and mode.

15.10 Summary

15.1 OBJECTIVE

After studying this unit a student should understand about

- Meaning of mode its merits and limitation
- Mode in individual, discrete and continues series
- Graphical determination of mode and
- Empirical representation between mean, median and mode.

15.2 MEANING

A mode is the French word "la-mode" which means a most popular phenomenon. In other word mode is maximum number occurring in the given series.

For example: x is a given series and in this number are. 2,5,2,4,6,5,8,10,2 etc. then in given x '2' is the mode. Because '2' is written 4 time and 5, 2 time and 4, 6, 8, and 10 comes only one time. So '2' is the mode because '2' comes maximum time in comparison other.

In mode series are divided in three parts. Individual, discrete and continues.

15.3 LIMITATIONS OF MODE

Mode has following limitation:-

(i) In small and individual series it's very difficult to find the mode.

(ii) In some case mode value are overlook more than one times in a series.

(iii) For continues series mode is a time taking activity

(iv) Mode is not suitable for future mathematical treatment.

(v) Mode is not based al the observation.

(vi) Its values always determined.

15.4 MERITS OF MODE

Mode possesses the following merits.

(i) Mode is easy and simple for calculation and to understand.

(ii) Mode are easily plated in graph by the help's of histogram

(iii) Value of mode is fixed.

(iv) In some core it can be located merely by inspection.

(v) It can be conveniently determined for distribution with open end class.

15.5 MODE IN INDIVIDUAL SERIES

In individual series it's very simple to find mode for a given sum as in individual series an item which comes maximum time is called mode.

So, it's very simply and less time taken.

Example 1: find out the mode for following data:-

5, 10, 15, 9, 8, 7, 3, 7, 7, 5, 2, 7, 10, 7, 20

Solution: Firstly arranged this series in ascending order:-

2, 3, 5, 5, 7, 7, 7, 7, 7, 8, 9, 10, 10, 15, 20

7 is the maximum item in given series so mode (z) – 7.

Answer. Z= 7

Example 2:

Given number of origination of 20 student find the mode:-

8, 4, 5, 5, 8, 1, 8, 7, 6, 8, 10, 10, 9, 8, 8, 7, 10, 8, 2, 8

Solution: Arranged the number in ascending order

2, 4, 5, 5, 6, 7, 7, 8, 8, 8, 8, 8, 8, 8, 8, 9, 10, 10, 10, 10

Z= 8

Because 8 is the maximum observation (8) time, So Z=8

Answer: z=8

15.6 MODE IN DISCRETE SERIES

In discrete series here item's and its maximum observation or frequencies are given.

So with the help's of frequency we see that which frequency is largest in item. The largest one is mode.

Example: - Calculate the mode from the following data:

Salary	400	200	150	75	300	250	350
Frequency	6	8	2	1	4	7	10

Solution: Z= 350

Because from the series here 10 is the largest frequency, 50, 350, is a mode of following salary.

Answer : Z = 350

Example 2: Find out mode of the following series:

Marks	1	2	3	4	5	6	7	8	9	10
No. of student	3	4	2	9	7	4	10	6	5	4

Solution: Z= 7

Because frequency of marks 7 area largest, so 7 is the mode of given marks.

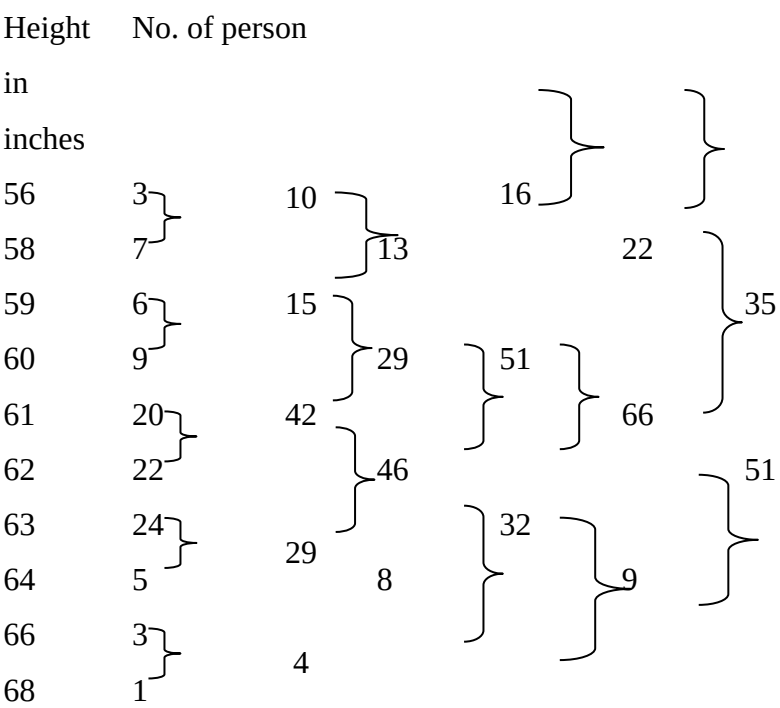
Note: in discrete and continues series mode can find by the help's of table (grouping table) and analysis table.

Example 3: Find out the following series:-

Heights in inches: 56,58,59,60,61,62,63,64,66,68

No. of person: 3,7,6,9,20,22,24,5,3,1

Solution: Grouping table



Col.No	56	58	60	61	62	63	64	66	68
1						1			
2				1	1				
3					1	1			
4			1	1	1				
5				1	1	1			
6					1	1	1		
Total			1	3	5	4	1		

The value 62 has comes the maximum number of times (5), 50 mode is 62.

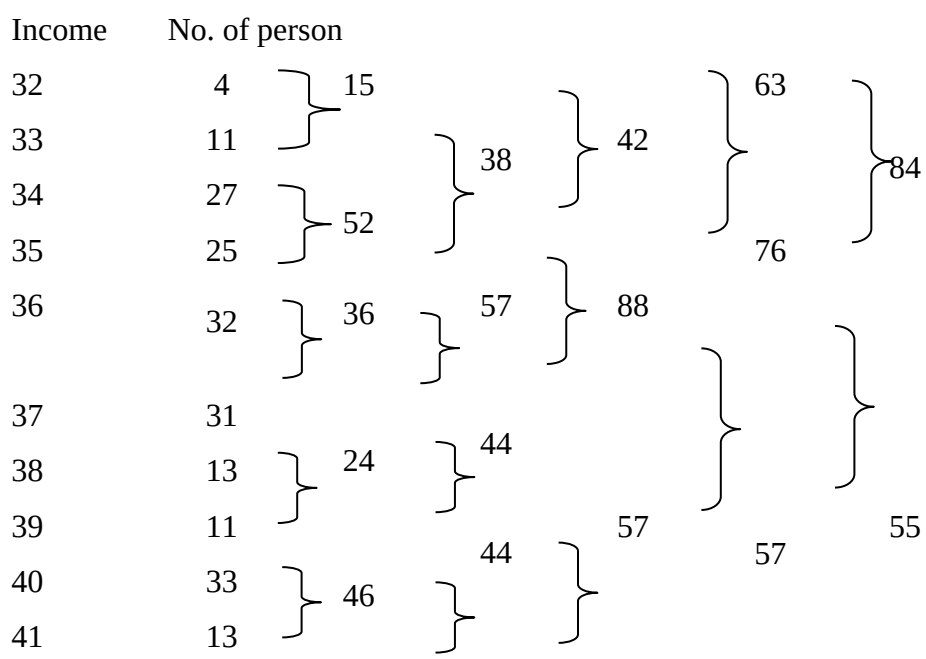
Answer = Z= 62

Example 2 : find out the method size of the income.

Income : 32, 33, 34, 35, 36, 37, 38, 39, 40, 41

No. of person : 4, 11, 27, 25, 32, 31, 13, 11, 33, 13

Solution: Grouping table :



Antilog table:

Col.no	32	33	34	35	36	37	38	39	40	41
1									1	
2					1	1				
3				1	1					
4				1	1	1				
5					1	1	1			
6				1	1					
Total	-`	-	1	3	5	3	1	-	1	-

15.7 MODE IN CONTINUES SERIES

In continues series first check that the different of class are equal if difference are not equal their firstly equal the class interval.

In continuous series a mode is find with the help of following formula:-

Formula:-

$Z = l_1 + \frac{f1-f0}{2f1-f0-f2} \times i$

Here:

Z= value of the mode.

L1= Lower limit of the modal group

f_1 = Frequency of modal class.

f_o = Frequency of pre-modal class

f_2 = Frequency of the next – model class.

i = Class Interval.

Example 1: Find mode of the data :

Class	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	4	8	2	10	6	5

Solution:

class	f	
0-10	4	
10-20	8	
20-30	2	
30-40	10	model class
40-50	6	
50-60	5	

The class 30-40 is model class because no. of observation is maximum (20):

Mode: $Z = l_1 + \frac{f1-f0}{2f1-f0-f2} \times i$

$= 30 + \frac{10-8}{2\times10-8-6} \times 10$

$= 30 + \frac{2}{20-14} \times 10$

$= 30 + \frac{20}{6}$

$= 30 + 3.333$

Z= 33.333

Answer: 33.333

Example: Find the mode of student

Class :	0-5	5-10	10-15	15-20	20-25	25-30	30-35
No. of student:	10	15	8	20	12	5	4

Solution:

Class	f.
0-5 5-10 10-15 15-20 20-25 25-30 30-35	10 15 8 20 12 5 4

A class 15-20 is a model class.

$Z = l_1 + \frac{f1-f0}{2f1-f0-f2} \times i$

$= 15 + \frac{20-8}{2\times20-8-12} \times 5$

$= 15 + \frac{12}{40-20} \times 5$

$= 15 + \frac{12\times5}{20}$

= 15+3

Z= 18

Answer = Mode = 18

Example 3: The distribution of weights of a group of 60 students in a class.

Weight	30-34	35-39	40-44	45-49	50-54	55-59	60-64
No. of student	3	5	12	18	14	6	2

Find the mode of distribution.

Solution: Given class is not a continuous class. So first convert the class into continuous series.

Weight	Class	Frequency
30-34	25.5-34.5	3
35-39	34.5-39.5	5
40-44	39.5-44.5	12
45-49	44.5-49.5	18
50-54	49.5-54.5	14
55-59	54.5-59.5	6
60-64	59.5-64.5	2

A model class is 45.5-49.5, 50

$Z = l_1 + \frac{f1-f0}{2f1-f0-f2} \times i$

$= 45.5 + \frac{18-12}{36-12-14} \times 5$

$= 45.5 + \frac{6}{10} \times 5$

$= 44.5 + 3$

$Z = 47.5$

Answer = 47.5

Example4: Find the mode of student.

Class	10-14	15-19	20-24	25-29	30-34
No. of student	5	8	15	10	4

Solution :

Class	No. of class	Frequency
10-14	9.5-14.5	5
15-19	14.5-19.5	8
20-24	19.5-24.5	15
25-29	24.5-29.5	10
30-34	29.5-34.5	4

A model class is 19.5-24.5

$Z = l_1 + \frac{f1-f0}{2f1-f0-f2} \times i$

$= 19.5 + \frac{15-8}{2\times15-8-10} \times 5$

$= 19.5 + \frac{7}{30-18} \times 5$

$= 19.5 + \frac{7\times5}{12} \times 5$

= 19.5+2.916

Z= 22.416

Answer= Z= 22.416

Example 5 : calculate mode for following data :

Marks	0-10	10-20	20-40	40-50	50-70
No. of student	2	7	18	15	8

Solution: The class interval they should first by mode equal by adjusting the frequency.

Marks	No. of student
0-10	2
10-20	7
20-30	9
30-40	9
40-50	15
50-60	4
60-70	4

The modal class is 40-50

$Z = l_1 + \frac{f1-f0}{2f1-f0-f2} \times i$

$= 40 + \frac{15-9}{2\times15-9-4} \times 10$

= $19.5+\frac{6\times10}{17}$

= 40+3.529

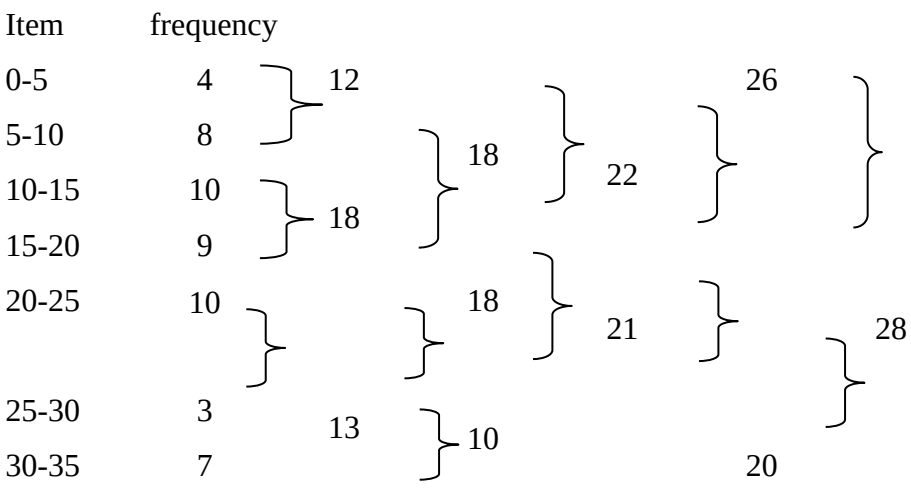
Z= 43.529

Answer= Z= 43.529

Example 6: Find out mode in the following series:

Item	0-5	5-10	10-15	15-20	20-25	25-30	30-35
Freq.	4	8	10	8	10	3	7

Solution: Grouping table



Analysis table:

Columns	0-5	5-10	10-15	15-20	20-25	25-30	30-35
1			1		1		
2			1	1			
3	1		1	1	1		
4	1	1	1				
5		1	1	1			
6			1	1	1		
total	1	2		4	3	0	0

A modal class is 10-15

$Z = l_1 + \frac{f1-f0}{2f1-f0-f2} \times i$

$= 10 + \frac{10-8}{2\times10-8-8} \times 5$

$= 10 + \frac{2\times5}{20-16}$

$= 10 + \frac{10}{4}$

= 10+2.25

Z= 12.25

Answer= Z= 12.25

Example 7: From the following distribution find the missing frequencies if median and mode are given on Rs. 33.5 and Rs.34.00

Wages (Rs.)	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	4	16	?	?	?	6	4

Solution: Calculate for missing frequencies.

Wages	F	L
0-10	4	4
10-20	16	12
20-30	F1	20+f1
30-40	F2	20+f1+f2
40-50	F3	20+f1+f2+f3
50-60	6	26+f1+f2+f3
60-70	4	30+f1+f2+f3

N= ∈f=30+f1+f2+f3

Given-

N= ∈f=230

30+f1+f2+f3= 230

F1+f2+f3=230-30

F1+f2+f3= 200…… (i)

Since median is given 33.5, so a class is 30-40,

$33.5= 30+ \frac{115-(20+f1)}{f2} \times 10$

$33.5-30 = \frac{95-f1}{f2} \times 10$

3.5f2=950-10f1

10f1 + 3.5f2 = 950…….(ii)

Further mode is given to be 34, so a class is

$34=30 + \frac{f1-f0}{2f1-f0-f2} \times 10$

$34-30=\frac{f1-f0}{2f1-f0-f2} \times 10$

4(2f1-f0-f2) = f1-f0 ×10

4(2f2-f1-f3) = 10f2+10f1

8f2 – 10f2-4f1+10f1-4f3=0

8f2-2f2-4f3 = 0…….(iii)

Multiplying equation. (i) by 4 and adding to equation (3) we get

4f1+4f2+4f3 = 800

6f1-2f2- 4f3 = 0

10f1+2f2 = 800……..(4)

Subtract equation (4) and equation(2)

10f1+2f2=800

6f1+2f2-4f3 = 950

\- + -

1.5f2=150

1.5f2=150

$F2= \frac{150}{1.5}$

F2 = 100

Put the value of f2 in equation (4)

10f1+2×100-800

10f1 = 800-200

$F1= \frac{600}{10}$

F1= 60

Put the value off1 and f2 in equation (1)

F1+f2+f3= 200

60+100+f3=200

160+f3=200

F3=200-160
F3= 40
Answer : Missing frequency are f1= 60
F2= 100, f3= = 40
Example 8: The distribution of age of patients turns out in a hospital

Age	No. of patients
More then 10	148
More then 20	124
More then 30	109
More then 40	71
More then 50	30
More then 60	16
More then 70	01
Up to 80	

Find the model age
Solution:

Age	No. of patients (f)
10-20	148-124=24
20-30	124-109= 15
30-40	109-71= 38
40-50	71-30= 41
50-60	30-16= 14
60-70	16-1 = 15
70-80	01

A model co class is 40-50

$Z = l_1 + \frac{f1-f0}{2f1-f0-f2} \times i$

$= 40 + \frac{41-38}{2\times41-38-14} \times 10$

$= 40+\frac{3\times10}{82-52}$

$= 40+\frac{30}{30}$

= 40+1

Z= 41

Answer= Z= 41

15.8 GRAPHICAL DETERMINATION OF MODE

A value of mode can also be located graphical from the histogram of frequency distribution.

Example 9:- Prepare histogram from the following distribution and find out mode.

Wages (Rs.)	0-10	10-20	20-40	40-50	50-60	60-70	70-90	90-100
Frequency.	4	6	14	16	14	8	16	5

Solution :

Marks	F
0-10	4
10-20	6
20-30	7
30-40	7
40-50	16
50-60	14
60-70	8
70-80	8
80-90	8
90-100	5

15.9 EMPIRICAL REPRESENTATION BETWEEN MEAN, MEDIAN AND MODE

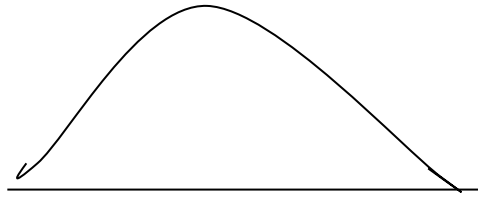
The mean tends to lie on the some side of the mode as the larger side and the median lies in between them.

Empirical relationship given by Prof. karl pearson.

Mean – Mode = 3(mean-Median)

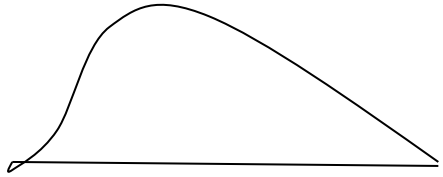
Mode = 3 Median – 2 median

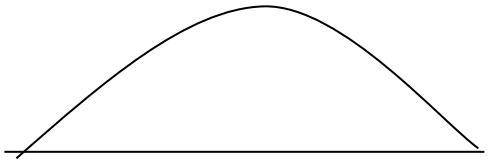
Median = mode + 2/3 (mean – mode)



Mode = Mode = Median

If we know any two values out of the three , we can compute the third from these.





Mode Median Mean Mean Median Mode

Example : 10: Here given : Mena 35.4 , and mode = 32.1 calculate the median

Solution :

Mean – Mode = 3 (Mean- Median)

35.4-32.1 =n 3(35.4 - median)

3.3 = 106.2- 3median

3median = 106.2-3.3

Median = $\frac{102.9}{3}$

Median = 34.3

Answer = Median = 34.3

*Question – 1: Find the mode of the number:

i. 5,3,27,.5,9,3,8,5

ii. 4,3,2,5,3,4,5,3,7,3,2,6

iii. 7,4,10,15,7,3,5,2,9,12

Answer : (i) 5 (ii) 3 (iii) 7

*Question – (2) : Calculate the mode form the following data of the marks obtain by 10 students:-

S. No.:	1	2	3	4	5	6	7	8	9	10
Mark obtain:	10	27	24	12	27	27	20	18	15	30

Answer : 27

*Question –(3): Find the value of mode from the following data:-

X:	1	2	3	4	5	6	7
F:	1	4	12	9	2	1	1

Answer : 3

*Question (4) : Calculate mode form the following data:

Monthly wages (Rs.)	200-250	250-300	300-350	350-400	400-450	450-500	500-550	550-600
No. of Students	4	6	12	19	33	17	8	2

Answer : 428.38

* Question (5) : Calculate mode from the following data:

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No. of student	10	9	25	30	16	10

Answer: 44

* question (6) : Given below is the distribution of marks obtains by 140 students in an examination:

Marks	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99

No. of student	7	15	18	25	30	20	16	7	2

Answer: 52.83

* Question (7) : Calculate mode from the followingdistribution:

Production	21-22	23-24	25-26	27-28	29-30
No. of days	7	13	22	10	8

Answer: 23.36

* Question (8) : Calculate mode from the following data:

Marks	0-10	10-20	20-40	40-50	50-60
Frequency	5	15	40	32	28

Answer: 44

*Question (9): Find out the model weight from the following data of the weights of 128 person:

Weights: 100-110, 110-120, 120-130, 130-140, 140-150, 150-160,160-170, 170-180

No. of person: 4,6,20,32,33,17,8,2

Answer : 140.05

*Question (10): Calculate mean, median and mode for the following data:

Class Interval: 93-97, 98-102, 103-107, 108-112, 113-117, 118-122, 123-127, 128-132

Frequency: 3 5 12 17 14 6 3 1

Answer:- $\bar{x}$= 110.6 Mode = 110.59, Mo= 110.625

*Question (12): The frequency distribution of marks obtained of marks obtained by 60 students of a class in a college 12 given below.

Marks	30-34	35-39	40-44	45-49	50-54	55-64
No. of student	3	5	12	18	14	8

Answer: 47.5

*Question (13): Draw a histogram for the following distribution and find the model value. Also verity your result by algebraic method.

Size	140-150	150-160	160-170	170-180	180-190	190-200	200-210
Frequency	40	60	100	180	40	30	20

Answer : 173 (app)

*Question 14: Determine the value of mode of the following distribution graphically and verity your result by actual calculations.

Class Interval	0-100	100-200	200-300	300-400	400-500
Frequency	14	23	27	21	15

Answer: 240

*Question 15:

(i) Give the empirical relation between mean , median and mode for a moderately skewed distribution-

Answer :- Mode = 3 Median – 2 mean

(ii) In a moderately asymmetrical distribution the value of mean is 75 and value of mode is 60, find the value of median.

Answer :- 70

(iii) Given mean = 70.2 and mode = 70.5 Find median using empirical relationship among them .

Answer :- 70.3

(iv) In a moderately skewed distribution the value of means is 10 and value of median is 9 find the value of mode.

Answer :- 7

(v) In a moderately skewed distribution, the value of mode is 12 and that of median is 140. Find the value of arithmetic mean.

Answer :- 150

Question 16:- Find the f missing value if given Mode = 17, Total number = 79

Class	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Freq.	4	8	10	?	5	6	14	12

Question 17:- Find the f missing value if given Mode = 44.285, Total number = 86, median= 43.5

class	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Freq.	4	8	10	?	?	?	18

Answer : f1 = 14, f2 = 20, f3 = 12

15.10 SUMMARY

A number which occurs maximum times its known as mode three types of number series is used in mode, continues, discrete and individual.

Mode = $l_1 + \frac{f1-f0}{2f1-f0-f2} \times i$

Z= value of the mode

L1= Lower limit of the modal group

f_1 = Frequency of modal class.

f_o = Frequency of pre-modal class

f_2 = Frequency of the next – model class.

i = Class Interval.

UNIT-16

16.1 Objective

16.2 Introduction

16.3 Geometric mean is discrete series

16.4 Geometric mean in continues series.

16.5 Weights Geometric mean.

16.6 Harmonic mean

16.7 Calculation of harmonic meaning discrete series

16.8 Harmonic mean in continuous series

16.9 Weights harmonic mean

16.10 Relationship between AM, GM and HM

16.11 Moving Average

16.12 Question

16.13 Summary

16.1 OBJECTIVE

After studying this unit student should known about geometric mean, weight, geometric mean, harmonic mean, weights harmonic mean, Relationship between AM, GM and HM, Moving Average.

16.2 INTRODUCTION

A geometric mean is totally different from simple mean. Geometric mean is the nth root of the product of n values of a series. If two number a and b are given then Geometric mean are $\sqrt{a \times b}$ so on if given number are three a,b and c then $\sqrt[3]{a \times b \times c}$

Formula: $\sqrt[n]{x_1.x_2.x_3.x_4.....x_n}$

Find the geometric mean of 7,8,2,5,10,9

Then G.M. = Formula: $\sqrt[6]{7 \times 8 \times 2 \times 5 \times 10 \times 9}$

Formula: $\sqrt[6]{50400}$

Use of geometric mean. It is very limited because to find out the roots so a mathematical term is very difficult and to calculate. We take the help of log. Table

Formula of geometric mean in different series.

(i) Individual series:- G= Anti log $\left(\frac{\sum log}{n}\right)$

(ii) Discrete and continues series:- G= Anti log $\left(\frac{\sum log \times b}{n}\right)$

Use of Geometric mean:-

A Geometric mean has various uses:-

(i) It's useful for find out appropriate average of data

(ii) A Geometric mean is suitable for give a weight age to smaller item and vice-versa.

(iii) Geometric mean are use which where a one data are objective and second smaller

(iv) Geometric mean is used primarily to average data for which the ratio of consecutive terms remains approximately constant. This occurs for example, with such data its rates of change, ratio percent increase in sale, population size.

Limitations of geometric mean:

In geometric mean has some limitation:

(i) A calculation of Geometric mean is very trouble and complicated.

(ii) Its calculation is done by only experts in mathematical terms.

(iii) When zero and negative then geometric mean are are imaginary.

Example : Find Geometric mean of 2, 4, 8, 12, 16, and 24

Solution : Calculation of G.M.

X	Log
2	0.3010
4	0.6021
8	0.9031
12	1.0792
16	1.2041
24	1.3802
N= 6	∈Logx= 5.4697

G= Anti log $\left(\frac{\in log}{n}\right)$

= Anti log $\left(\frac{5.4697}{6}\right)$

Al (0.9116)

= 8.159

Example 2: Find Geometric mean: - 156, 542, 2158, 68,332

Solution: Calculation of G.M.

X	Logx
156 542 2158 68 332	2.1931 2.7340 3.3340 1.8325 2.5211
N= 5	∈Logx= 12.6147

G.M. = Anti log $\left(\frac{\in log}{n}\right)$

= Anti log $\left(\frac{12.6147}{5}\right)$

Al (12.6147)

= 12.6147

Answer= G.M. = (12.6147)

16.3 GEOMETRIC MEAN IS DISCRETE SERIES

Example 3: Find Geometric mean from the following data:-

Diameter: 130,135,140,145,143,148,149,150

No. of screws: 3, 4,6,63,5,2,2

Solution: Calculation of G.M.

Diameter	No. of student(f)	Log x	f logx
130	3	2.1139	6.3417
135	4	2.1303	8.5212
140	6	2.1461	12.8766
145	6	2.1614	12.9684
143	3	2.1553	6.4649
148	5	2.1703	10.8515
149	2	2.1732	4.3464
150	2	2.1761	4.3522
	N= ∈f = 31		∈Logx= 66.7239

G.M.= Anti log $\left(\frac{\in logx}{n}\right)$

= Anti log $\left(\frac{66.7239}{31}\right)$

Al (2.1524)

= 142.0

Answer= G.M.= (142.0)

Example 4: Calculate Geometric mean:-

15 14 10 16 13 12 11

2 3 2 1 3 5 4

Solution:

X	f	Log	f logx
15	2	1.1761	2.3522
14	3	1.1461	3.4383
10	2	1.000	2.0000
16	1	1.2041	1.2041
13	3	1.1139	3.3417
12	5	1.0792	5.3960
11	4	1.0414	4.1656
	N= ∈f =20		∈Logx =21.8979

G.M. = Anti log $\left(\frac{\in logx}{n}\right)$

= Anti log $\left(\frac{21.8979}{20}\right)$

Al [1.0949]

= 12.45

Answer= G.M. = 12.45

16.4 GEOMETRIC MEAN IN CONTINUOUS SERIES

Example 5:- Find Geometric mean

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No. of student	10	9	25	35	16	10

Marks	Mid-Value	f	Logx	f log x
0-10	5	10	0.6990	6.99
10-20	15	9	1.1761	10.5849
20-30	25	25	1.3979	34.9475
30-40	35	35	1.5441	46.323
40-50	45	16	1.6532	26.4512
50-60	55	10	1.7404	17.404
	N= ∈f =100			∈Logx =142.7006

G.M. = Anti log $\left(\frac{\in flogx}{n}\right)$

= Anti log $\left(\frac{142.7006}{100}\right)$

Al [1.4270]

= 26.73

Answer= G.M. = 26.73

Example 6: Calculate Geometric mean

Marks	0-10	10-20	20-30	30-40	40-50
No. of student	10	9	25	35	16

Solution: Calculation of G.M.

Marks	Mid-Value	f	Logx	f log x
0-10	5	5	0.6990	3.4950
10-20	15	7	1.1761	8.2327
20-30	25	15	1.3979	20.9685
30-40	35	28	1.5441	38.6025
40-50	25	8	1.6532	13.3256
	N= 60			∈Logx =84.5243

G.M. = Anti log $\left(\frac{\in flogx}{n}\right)$

= Anti log $\left(\frac{84.5253}{60}\right)$

Al [1.4087383]

= 25.63

Answer=

G.M. = 25.63

16.5 WEIGHTS GEOMETRIC MEAN.

Formula: W.G.M. $\left[\frac{\in WLOGX}{\in W}\right]$

Example: Calculate weighted geometric mean.

Group	Index number	Weight
Food	350	5
Fuel and lighting	250	1
Clothing	240	1
House rent	160	1
miscellaneous	250	2

SOLUTION : CALCULATION OF WEIGHTED GEOMETRIC MEAN

Group	Index number	Weight	Weight (w)	W log n
Food	350	5	2.5441	12.7205
Fuel and lighting	250	1	2.3010	2.3010
Clothing	240	1	2.3802	2.3802
House rent	160	1	2.2041	2.2041
miscellaneous	250	2	2.3979	4.7959
		∈fw=10		∈Logx =142.7006

Formula: W.G.M. $\left[\frac{\in WLOGX}{\in W}\right]$

AL= $\left[\frac{24.4016}{10}\right]$

= AL (2.4401)

Answer: GM_W= 275.5

Example 8: The weighted geometric mean of 5 number 10,15,25,20 and 15. If the weights of the first four number are 2, 3, 5, and? Responsibility, find the weight of the fifty member.

Solution: Let the weight of fifty member be w:-

Calculation of weighted geometric mean.

X	w	Log x	w logx
10	2	1.000	2.000
15	3	1.1761	3.5283
25	5	1.3979	6.9895
12	2	1.0729	2.1584
20	W	1.3010	1.3010w
	∈w=12+w		∈ $WLOGX$=21.8979

Applying the formula: - log G.Mw=. $\left[\frac{\in WLOGX}{\in W}\right]$

Log (17.15) = $\frac{14.6762+1.3010w}{12+w}$

1.2343 = $\frac{14.6762+1.3010w}{12+w}$

1.2343 (12+w) = 14.6762 + 1.3010w

1.3010-1.2343=14.6762-14.8116

0...667w = 0.1354

W= $\frac{0.1354}{0.0667}$

W= 2(app.)

Answer = w= 2

16.6 HORMONIC MEAN

A harmonic mean is a reciprocal of the overage of the reciprocal of the given set of observation. If a number is zero or negative the resulting harmonic mean sets in circumstances.

When one factor is constant and the average rate is to be find for other factors.

The various formulas for harmonic mean:-

Calculation of harmonic mean individual series

H.M. = $\frac{n}{\frac{1}{x1}+\frac{1}{x2}+\cdots\ldots\ldots\ldots\ldots\frac{1}{xn}}$ = $\frac{n}{\in\frac{1}{x}}$

Example 1:- Find the harmonic mean of 5 numbers

4, 5,6,10 and 12

Solution: H.M. = $\frac{n}{\in\frac{1}{x}}$

$= \frac{5}{\frac{1}{4}+\frac{1}{5}+\frac{1}{6}+\frac{1}{10}+\frac{1}{12}}$

$=\frac{5}{\frac{15+12+10+6+5}{60}}$

$=\frac{5\times60}{48}$

$=\frac{300}{48}$

$=\frac{25}{4}$

6.25

Answer= H.M. = 6.25

Example 2: Find H.M. of number

80, 10, 20,40,50,30

Solution: H.M. = $\frac{n}{\in\frac{1}{x}}$

$= \frac{6}{\frac{1}{80}+\frac{1}{10}+\frac{1}{20}+\frac{1}{40}+\frac{1}{50}+\frac{1}{30}}$

$=\frac{6}{\frac{3+24+12+6+4.8+8}{240}}$

$=\frac{6\times240}{57.8}$

$=\frac{1440}{57.8}$

24.913

Answer= H.M. =24.913

16.7 CALCULATION OF HARMONIC MEANING DISCRETE SERIES

H.M. = $\frac{f1+f2+\cdots\ldots\ldots\ldots\ldots fn}{\frac{f1}{n1}+\frac{f1}{n2}+\cdots\ldots\ldots\ldots\ldots\frac{fn}{xn}}$

$=\frac{\in f}{\in f/x}$

$=\frac{\in f}{\in f/n}$

Example 3: Find Harmonic mean

X:	10	15	20	25	30
F:	2	5	4	9	5

Solution: Calculation of H.M.

X	f	f/x
10	2	0.2
15	5	0.34
10	4	0.2
25	9	0.36
30	5	0.167
	∈f=25	∈f/n=1.267

H.M. $=\frac{\in f}{\in f/x}$

= 25/1.267

= 19.73

Answer= H.M. = 19.73

Example 4: Find H.M. of data

X : 5 25 10 15 35 40 30

F : 5 6 2 8 10 3 2

X	F	f/x
5	5	1.0
25	6	4.166
10	2	5.0
15	8	1.875
35	10	3.5
40	3	13.333
30	2	15.0
	∈f=36	∈f/n=43.874

H.M. $=\frac{\in f}{\in f/x}$

= 36/43.874

= 0.82053

Answer= H.M. = 0.82053

16.8 HARMONIC MEAN IN CONTINUOUS SERIES

H.M. = $\frac{f1+f2+\cdots\ldots\ldots\ldots\ldots fn}{\frac{f1}{x1}+\frac{f1}{x2}+\cdots\ldots\ldots\ldots\ldots\frac{fn}{xn}}$

$=\frac{\in f}{\in f/x}$

$=\frac{n}{\in f/x}$

Example 5: Calculation of H.M.

Class interval: 10-20, 20-30, 30-40, 40-50, 50-60

Frequency: 15, 20, 30, 10, 5

Solution : Calculation of H.M.

Class	Mid-value (x)	f.	f/x
10-20 20-30 30-40 40-50 50-60	15 25 35 45 55	15 20 30 10 5	1 0.8 0.857 0.223 0.091
	N= 80		$\in f/x$=2.971

H.M. =$\frac{n}{\in f/x}$

= 80/2.971

= 26.93 (app.)

Answer= H.M. = 26.93 (app.)

Example 6: Find H.M. of data

Class : 0-10, 10-20, 20-30, 30-40, 40-50, 50-60

Frequency: 5, 15,10,25,8,5

Solution: Calculation of H.M.

Class	Mid-value (x)	f.	f/x
0-10	5	5	1
10-20	15	15	1
20-30	25	10	2.5
30-40	35	25	0.1714
40-50	45	8	0.177
50-60	55	5	0.9090
	∈N= 68		$\in f/x$=5.4819

H.M. =$\frac{n}{\in f/x}$

= 68/ 5.4819

Answer= H.M. = 12.404

Limitation of Harmonic mean:

1. A Harmonic mean is time taken comparison to simple average mean.
2. It's difficult to calculate and not easy to understand.
3. Harmonic mean is not popular among common man.
4. When a value is zero or negative in series then harmonic mean cannot be calculated.

16.9 WEIGHTS HARMONIC MEAN

Weighted Harmonic mean can be find by the help of following formula:-

H.Mw. = $\frac{\in w}{\in \frac{1}{x} \times w}$

Example: Kishor travels 900km. by train at an average speed of 60kms per hour. 3000km. by ship at an average speed of 25km.per hour; 400kms by a airplane at 350 kms per hour and finally 15kms by bus at 25kms per hour. Calculate his average speed for the entire journey.

Solution:

Speed x	Distance(w)	w/x
60	9000	15
25	3000	120
350	400	1.143
25	15	0.6
	∈w= 4315	∈/w = 136.743

Average speed = $\frac{\in w}{\in w/x}$

= 4315/ 136.743

= 31.556km/h

Answer: 31.556Kmh

Example 2: An individual purchases three qualities of pencils. The relevant data are given blow.

Quality	Price Per Pencil (Rs.)	Money Spent(Rs.)
A	1.00	50
B	1.50	30
C	2.00	20

Calculate the average price per pencil

Solution: - Average price = H.Mw = $\frac{\in w}{\in w/x}$

= $\frac{50+30+20}{\frac{50}{1.0}+\frac{30}{1.50}+\frac{20}{2.00}}$

= 100/ 50+20+10

= 100/80

1.25

Answer: - 1.25

16.10 RELATIONSHIP AMONG THE AM, GM AND HM

The value of AM, GM and HM are connected by the following relation.-

AM≥GM ≥HM

Note: - For two number, we have $(G.M.)^2 = A.M. \times H.M.$

Example 1: The A.M. and G.M. of two number are 15, 12 responsibility. Find the H.M. of the number.

Solution: Let two number be a and b. then we are given

A.M. = $\frac{a+b}{2}$ = 15

= a+b = 30

G.M. = $\sqrt{ab}$

G.M. = $\sqrt{ab}$ = 12

Ab= 144

H.M. = $\frac{2}{\frac{1}{a}+\frac{1}{b}}$

= $\frac{2}{\frac{b+a}{ab}}$

= $\frac{2ab}{a+b}$

= $\frac{2\times144}{30}$

= 9.6

Or

For two number, we know that

$(G.M.)^2$= A.M. × H.M.

$(12)^2$= 15× H.M.

144=15 × H.M.

H.M. = 144/15

H.M. 9.6

Answer = 9.6

16.11 MOVING AVERAGE

A moving average is use for the measurement of long term trends. This average is used for minimising the short term period fluctuations it's of arithmetic average.

For a given of observation $Y_1, Y_2, Y_3 \ldots\ldots\ldots\ldots\ldots Y_n$ we define moving totals of orders K as the sequence of sums.

$Y_1+Y_2+Y_3 \ldots\ldots\ldots\ldots\ldots Y_k, , Y_2+Y_3+Y_4 \ldots\ldots\ldots Y_{k+1},$

$Y_3 + Y_4+Y_5+\ldots\ldots\ldots Y_{k+2}$, so on

And moving average of order (k) as the sequence of arithmetic means.

$$\frac{Y_1+Y_2+Y_3 \ldots\ldots\ldots\ldots\ldots Y_k, ,}{k} \qquad \frac{Y_2+Y_3+Y_4 \ldots\ldots\ldots Y_{k+1},}{k}$$

$$\frac{Y_3 + Y_4+Y_5+\ldots\ldots\ldots Y_{k+2}, \ldots\ldots\ldots\ldots\ldots\ldots}{k}$$

in moving average we get two type of series (i) odd (ii) even

In case of odd number of series it's easy for calculation because calculation of median is easy. But in case of even number of series the calculation are longer.

For example: Find out the four yearly moving average here calculation for moving average there steps (i) first odd the year's four grouped (ii) add two yearly total of four years, total (iii) divided by eight add of two yearly totals.

Example : Give 3 yearly moving average of the following series.

Give	Production	Year	Production
1990	25	2001	4
1991	35	2002	18
1992	20	2003	20
1993	45	2004	22
1994	30	2005	40
1995	15		
1996	10		
1997	22		
1998	42		
1999	37		
2000	16		

Solution:

Year	Production	3year total	3yearly moving average
1990	25	-	-
1991	35	80	26.66
1992	20	100	33.33
1993	45	95	31.66
1994	30	90	30.00
1995	15	55	18.33
1996	10	47	15.66
1997	22	74	24.66
1998	42	101	33.66
1999	37	95	31.66
2000	16	57	19.00
2001	4	38	12.66
2002	18	42	14.00
2003	20	60	20.00
2004	22	82	27.33
2005	40	-	-

Example: Calculate five yearly moving average :

Solution:

Year	Area	5year total	5yearly moving average
1997	420	-	-
1998	340	1040	208
1999	280	1140	228
2000	520	1470	294
2001	670	1480	296
2002	290	1170	234
2003	210	680	136
2004	380	890	178
2005	300	720	144
2006	140	620	124
2007	180	580	116
2008	260	760	152

2009	320	830	166
2010	250	720	144
2011	150	-	-

Example : find four yearly moving average :

Year	Area
1995	80
1996	210
1997	150
1998	50
1999	100
2000	240
2001	300
2002	320
2003	250
2004	90
2005	140
2006	160

Solution :

Year	Sales	4yearly total	2yearly total of 4 yearly total	4yearly moving average
1995	80	-	-	-
1996	210	-	-	-
1997	150	490	1,000	250.0
1998	50	510	1150	287.0
1999	100	540	1230	307.5
2000	240	690	1650	412.5
2001	300	960	2070	517.5
2002	320	1110	2070	517.5
2003	250	960	1760	440.0

2004	90	800	1440	360.0
2005	140	640	-	-
2006	160	-	-	-

Example : find six yearly moving average.

Year	Income
1990	500
1991	450
1992	320
1993	550
1994	400
1995	220
1996	100
1997	150
1998	200
1999	250
2000	300
2001	350

Solution:

Year	Income	6yearly total	2yearly total of 6 yearly total	6yearly moving average
1990	500	-	-	-
1991	450	-	-	-
1992	320	2440	-	-
1993	550	2040	4480	746.66
1994	400	1740	3780	630.00
1995	220	1620	3360	560.00
1996	100	1320	2940	490.00
1997	150	1220	2540	423.33

1998	200	1350	2570	428.33
1999	250	-	-	-
2000	300	-	-	-
2001	350	-	-	-

16.12 QUESTION

Question 1: Point out the merits of geometric mean.

Question 2: Define weighted geometric mean.

Question 3: Find the geometric mean of :-

(i) 3,6,24 and 48 (ii) 24,72,108 and 144

Answer (i) 12 (ii) 72

Question 4: Calculate geometric mean of the following:

(i) 1,7,18,65,91 and 103

(ii) 1,7,29,92,115 and 375

Ans. (i) 20.61 (ii) 30.50

Question 5 : Calculation the geometric mean from the following data.

Daily earning (Rs.): 100 150 80 200 250 180

No. of employees: 24 26 16 20 6 30

Answer : 143.4

Question 7: Calculate the geometric mean from the following data:

Weight (in pounds)	100	110	120	130	140
No. of person	15	20	25	30	10

Answer: 119.3

Question 5: Find the geometric mean for the following distribution:

Class interval	10-20	20-30	30-40	40-50	50-60
Frequency	15	20	30	10	5

Answer: 49.64

Question 6: Calculate the geometric mean for the following distribution:

Marks	20-29	30-39	40-49	50-59	60-69	70-79
No. of students	5	11	18	22	16	8

Answer : 49.64

Question 7: Find the geometric mean for the following distribution:

Mid value	105	115	125	135	145	155	165	175
Frequency	4	6	20	30	33	17	8	2

Answer: 138.8

Question 8: Three groups of observation contain 8,7 and 8 observation. Their geometric are 8.52, 10.12 and 7.75 respectively. Find the geometric mean of the 20 observations in the single group formula by pooling the three groups.

Answer: 8.837

Question 9: Calculate the weighted geometric mean from the following data:

Group	Index number	weights
Food	225	52
Fuel and lighting	155	10
Clothing	175	8
House rent	250	14
miscellaneous	150	16

Answer: 202

Question 10: The weighted geometric mean of four numbers 8,25,17 and 30 is 15.3. If the weight of the first three numbers is 5.3 and 4 respectively, find the weight of the fourth number.

Answer: 2

Question 11: If arithmetic mean and geometric mean of two numbers are 15 and 12 respectively, find numbers.

Answer: 24,6

Question 1: Compute 3-yearly moving average value for the following data:

Year	1997	1998	1999	2000	2001	2002	2003
Value	2	4	6	8	13	12	14

Question 2: Calculate the 3 yearly moving averages of the production figures given below and determine the trend value. Find the short- term fluctuations.

Questions 3: Find the trend of the following time series by the method of moving averages assume a four yearly cycle.

Question 4: form the following data, calculate the trend values using 4-yearly moving averages.

Question 5: obtain the trend values for the quarterly data by the moving averages method.

Question 6: Calculate 5–yearly moving averages for the following data.

Year	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Value (Rs.)	123	140	110	98	104	133	95	105	150	135

Question 7: for the following data verify that give yearly weighted moving average trend values with weights 1,2,2,2,1 respectively is equivalent to 4 yearly centered moving average trend values:

Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
Sales(Rs.-in lakh)	5	3	7	6	4	8	9	1	8	9	9

Question: from the following data, calculate the trend values using 3-yearly moving averages.

Year	2000	2001	2002	2003	2004	2005
Value	20	24	22	32	27	31

Question: Find the trend of the following time series by the method of moving average (assume a four – yearly cycle).

Year	Value	Year	Value
1970	53	1977	87
1971	79	1978	89
1972	76	1979	104
1973	66	1980	97
1974	69	1981	92
1975	94	1982	11
1976	105	1983	105

Question: Calculate the trend values for the following time series using 4 – yearly moving averages method.

Year	Value	Year	Value
1977	12	1984	100
1978	25	1985	82
1979	39	1986	65
1980	54	1987	49
1981	70	1988	34
1982	87	1989	20
1983	105	1990	7

Question : Calculate 4–yearly moving averages for the following data.

Year	Value	Year	Value
1976	24	1982	78
1977	28	1983	94
1978	34	1984	112
1979	42	1985	132
1980	52	1986	154
1981	64	1987	178

Question : Estimate the trend values using the data given below taking 4-yearly moving average and eliminate the trend.

Year	Value	Year	Value
1991	24.1	1998	45.3
1992	25.1	1999	39.3
1993	27.3	2000	41.3
1994	28.3	2001	42.2
1995	28.1	2002	46.4
1996	29.1	2003	46.6
1997	30.1	2004	49.2

Question: Calculate the trend values for the following time series using 5 – yearly moving averages method.

Year	Value	Year	Value
1996	242	2002	251
1997	250	2003	257
1998	252	2004	260
1999	249	2005	265
2000	253	2006	262
2001	255		

16.13 SUMMARY

Geometric mean is the nth root of the product of n values of a series.

Formula of G.M.

In Individual series = G= Anti log $\left(\frac{\in logx}{n}\right)$

Discrete and continuous series = G= Anti log$\left(\frac{\in flogx}{n}\right)$

Formula for harmonic mean

H.M. = $\frac{n}{\frac{1}{x1}+\frac{1}{x2}+\cdots\ldots\ldots\ldots\ldots\frac{1}{xn}}$ = $\frac{n}{\in\frac{1}{x}}$

Uttar Pradesh Rajarshi Tandon
Open University

Bachelor of Business Administration

BBA-107

Business Statistics

BLOCK

5

UNIT-17 **309**

Measures of Dispersion-I

UNIT-18 **323**

Measures of Dispersion-II

UNIT-19 **337**

Measures of Skewness-I

UNIT-20 **341**

Measures of Skewness-II

Curriculum Design Committee

Prof. Omji Gupta **Director**

School of Management Studies, UPRTOU, Prayagraj

Dr Gyan Prakash Yadav

Assistant Professor

School of Management Studies, UPRTOU, Prayagraj

Dr Devesh Ranjan Tripathi

Assistant Professor

School of Management Studies, UPRTOU, Prayagraj

Course Preparation Committee

Dr T.M. Raini **Author**

Associate professor, Firoj Gandhi PGCollege, Raibarely

Dr Shiv Bushan Gupta

Associate professor, MMM PG college, Kalakankar, Pratapgarh

Dr Gaurav Sankalp

Academic consultant, School of Management Studies, UPRTOU, Prayagraj

Dr Ajay Tiwari

Assistant Professor

MONIRBA, Department of Commerce, University of Allahabad

Prof J.N. Mishra **Editor**

Professor, MONIRBA, Department of Commerce, University of Allahabad

Dr Gyan Prakash Yadav **Coordinator**

Assistant Professor

School of Management Studies, UPRTOU, Prayagraj

©UPRTOU, Prayagraj-2020

ISBN :

©All Rights are reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the **Uttar Pradesh Rajarshi Tondon Open University, Prayagraj.**

UNIT-17 MEASURES OF DISPERSION-I

17.1 INTRODUCTION

We have learnt that a single and central value of a given data set may be found out. This central value represents the characteristics of the entire data set. It is known as measure of central tendency. The three main measures of this central tendency are viz. mean, median, & mode. However, these measures of central tendencies (average) are not able to reveal the full details of the data set. Data sets may have the same averages but they may differ from each other. Suppose there are three students A, B and C in a class. Three class tests have been conducted by the teacher in the class. In the class tests their scores were as follows:

Table: 17.1 Example

	Student A	Student B	Student C
Class Test -1	20	18	0
Class Test -2	20	20	20
Class Test -3	20	22	40
Mean score	20	20	20

In this example, we see that the mean score of each student is 20, although the scattering of scores are different in each case. In statistics, the name given to this scattering of data is dispersion or deviation. In other words, deviation is the study of how far each element is from some measure of central tendency (average). A small deviation among values in the data set hints that data are clustered closely around the mean. The mean is therefore considered good representative of the data set and it is reliable. In contrast to this, a large deviation among values in the data set implies that mean is not a true representative of data set and it is not reliable. Meaning thereby, we are interested in studying the deviation or variability among data to get the true picture of the data set.

There are several ways to measure the dispersion or variability of the data. Although the most common and most important is the standard deviation, which provides an average distance for each element from the mean, yet several others are also important, and are hence discussed here. In this unit we will first try to understand the concept of dispersion. In the later part of the unit, various methods to study dispersion like:

range, inter quartile range and quartile deviations, and mean deviations are discussed.

17.2 DEFINITIONS

Some important definitions of dispersion or deviation are as follows:

(i) According to A L Bowley, ”Dispersion is the measure of the variations of the items.”

(ii) 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 data.”

The arithmetic mean, median, mode, geometric mean, harmonic mean etc. are known as *averages of first order*. In the computation of measures of dispersion, the average value is derived from the averages of first order, thus, the measures of dispersion are known as *averages of second order*.

17.3 SIGNIFICANCE OF MEASURING DISPERSION

(1) Measures of dispersion help us to know whether an average is true and reliable representative of the data set. If the difference between central value and the value of the other items of the data set is large, the average or central value is not a good representative of the data set.

(2) Measures of dispersion put forward the nature and causes of variation. Such analysis is useful in controlling the variation. *Spurr* and *Bonimi* have very correctly said,”in matters of health, variations in body temperature, pulse beats and blood pressure are basic guides to diagnosis. Prescribed treatment is designed to control their variation. In industrial production, efficient operation requires control of quality variation, the causes of which are sought through inspection and quality control programmes.”

(3) Measures of dispersion aids in making comparison the variability of two data sets by calculating relative measures of dispersion. The consistency and uniformity in the values of variable is decided by it. Lower the degree of variability the higher is the degree of consistency or uniformity in the data set, whereas, higher the degree of variability the lesser is the consistency or uniformity in the data set.

(4) The use of other statistical techniques such as correlation and regression analysis, quality control, hypothesis testing, forecasting etc. are being facilitated by the measures of dispersion.

17.4 PROPERTIES OF A GOOD MEASURE OF DISPERSION

The properties of a good measure of dispersion are mentioned below. On the basis of these properties the merits and demerits of individual measures of dispersion are identified.

(i) It should be rigidly defined.

(ii) It should be based on all the elements (values) in the data set.

(iii) It should be accurate, easy to calculate and simple to follow.

(iv) It should not be affected by the extreme values of the data set. Similarly fluctuations of sampling should not affect it.

(v) It should be capable of further mathematical and algebraic treatment.

17.5 ABSOLUTE AND RELATIVE MEASURES OF DISPERSION

The several measures of dispersion may be put into two categories:

(i) Absolute measures,

(ii) Relative measures

Absolute measures are described by a number or value to show the amount of variation among values in data set. It is expressed in the same measuring unit as the unit of the original data set such as meter, kilograms, feet, rupees, pounds, tonnes. Absolute measures are very useful in making comparison of two or more data sets in the same measuring unit.

Relative measures are expressed as a percentage or ratio. It is termed as coefficient of variation. Since relative dispersion is a ratio, it has no unit. In the computation of relative measures, the average value used as base should be same from which the absolute deviations were computed.

17.6 RANGE

Range is the simplest method of measuring dispersion. Range is defined as the difference between the value of the smallest observation and the value of the largest observation present in the data set. It is given by:

Range = (L – S) ---(eq^n 17.1)

Where, L = Largest value

S = Smallest value

For a grouped frequency distribution or a continuous frequency distribution, range is defined as the difference between the upper limit of the highest class and the lower limit of the smallest class.

Range = Upper limit of the highest class - Lower limit of the lowest class ---(eq^{n} 17.2)

17.6.1 COEFFICIENT OF RANGE

The range calculated above is not very useful for comparison if the data sets are in different units. For example, we cannot compare the range of the weights of employees with range of their heights, as the range of weights would be expressed in kilograms and that of heights would be expressed in centimeters or inches.. Hence, for making comparison purposeful, a relative measure of range is needed, which is known as coefficient of range. It is expressed as follows:

$$\text{Coefficient of Range} = \frac{\text{largest value} - \text{smallest value}}{\text{largest value} + \text{smallest value}}$$

$$= \frac{L - S}{L + S} \qquad \text{---(eq}^{n}\text{ 17.3)}$$

Table 17.2 presents the information about 200 employees in an organization in various age groups. You are required to calculate range and the coefficient of range.

Table 17.2: Ages of 200 Employees

Class (Yrs)	**Number of employees (f)**
36-40	14
41-45	32
46-50	78
51-55	56
56-60	20
Total	**200**

Range = (Upper limit of the highest class) – (Lower limit of the lowest class)

= 60-36 = 24.

And, coefficient of Range $= \frac{L-S}{L+S} = 24/96 = 0.25$

17.6.2 MERITS

(i) Range is simple to understand and easy to calculate.

(ii) Range is the quickest way to get a measure of dispersion, although its accuracy is always doubtful.

17.6.3 DEMERITS

(i) Range is not based on all the observations in the data. It is only based on highest and lowest values of the data set and overlooks the nature of dispersion among other values of the given data set.

(ii) Range cannot be calculated from frequency distributions with open-end classes.

(iii) Range is highly affected by extreme values of the data set and hence fluctuates from sample to sample of a population.

(iv) Range is unable to explain about the character of the distribution within two extreme observations.

(v) Range is not a reliable measure of dispersion of the values within a data set.

17.6.4 USES OF RANGE

Despite many limitations and short-comings, the range, still as a measure of dispersion is very useful. Following points explains some of its uses:

(i) In industry for the quality control of products without complete inspection, the range is used. Range plays an important role in construction of charts used for quality control.

(ii) Range is also useful in analyzing the fluctuations in share markets.

17.7 INTER QUARTILE RANGE AND QUARTILE DEVIATION

The shortcomings of range are removed up to some extent in the computation of inter-quartile range. As we know, range as a measure of dispersion is based on two extreme observations and it fails to explain the scatter within the range. So when these extreme observations are discarded, the limited range would become more accurate. Again it represents the entire data too. This range calculated based on the middle 50 percent of the observations is called inter-quartile range. This inter-

quartile range is calculated from observations obtained after discarding one quartile of the observations at the lower end and another quartile of the observations at the upper end of the data set. Thus, inter-quartile range is the difference between the third quartile and the first quartile. The quartiles (Q_1, Q_2, and Q3) are the highest values in each of the four parts of the distribution.

$\therefore$ Inter-quartile range = (Q_3-Q_1) ---(eq^n 17.4)

Figure 17.1 displays the concept of inter-quartile range. The area under the curve is divided into four equal parts (25% each).

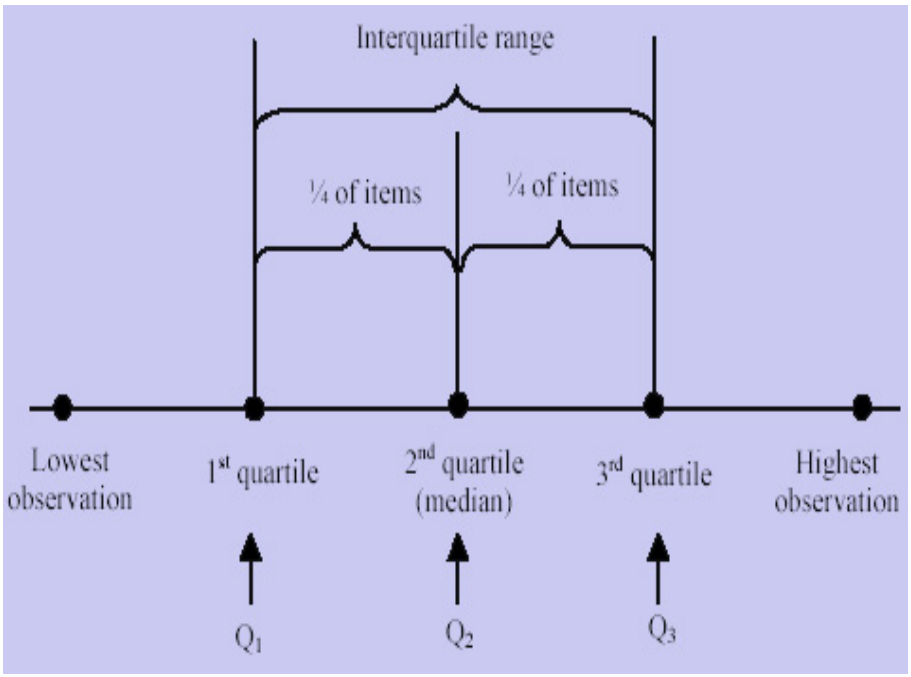


Figure 17.1 Inter-quartile Range

When inter-quartile range is divided by two, we get Quartile deviation. It is defined as one half of the inter-quartile range.

Therefore, Quartile deviation (Q.D.) = $\frac{Q_3 - Q_1}{2}$ ---(eq^n 17.5)

Quartile deviation computes the average value by which the two quartiles differ from the median. In a symmetrical distribution or data set, the quartiles Q_3 and Q_1 are equidistant from the median i.e.

(Median - Q_1) = (Q_3 – Median) ---(eq^n 17.6)

This difference can be taken as a measure of variation.

Hence, a small quartile deviation denotes less variation in the central 50% of the observations, while, a high quartile deviation indicates large variations.

17.7.1 COEFFICIENT OF QUARTILE DEVIATION

The relative measure of quartile deviation is called coefficient of quartile deviation. It is utilized to compare the degree of variation in different data sets. It is given by as below:

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \qquad \text{---(eq}^n\text{ 17.7)}$$

17.7.2 COMPUTATION OF QUARTILE DEVIATION

Computation of Q.D is very easy and quick. We need to calculate the values of the upper (Q_3) and lower (Q_1) quartiles only.

In case of Ungrouped Data

$$\text{Lower quartile } (Q_1) = \frac{(N+1)}{4}\text{th observation} \qquad \text{---(eq}^n\text{ 17.8)}$$

$$\text{Upper quartile } (Q_3) = \frac{3(N+1)}{4}\text{th observation} \qquad \text{---(eq}^n\text{ 17.9)}$$

Where, N = total number of observations.

In case of Grouped Data

$$Q_1 = L_1 + \frac{(\frac{1}{4}N - C)}{f} \times h \quad \text{---(eq}^n\text{ 17.10)}$$

$$Q_3 = L_3 + \frac{(\frac{3}{4}N - C)}{f} \times h \quad \text{---(eq}^n\text{ 17.11)}$$

Where,

L_1 = the lower boundary of the first quartile class (Q_1)

L_3 = the lower boundary of the third quartile class (Q_3)

N = Total cumulative frequency

f = Frequency of the quartile class

h = width of the Class interval

C = Cumulative frequency of the class just before the quartile class

The following example explains the computation of Quartile Deviation for the data presented in Table 17.2.

Table 17.2 Weights of Students

Weight (Kg)	55-60	60-65	65-70	70-75	75-80
Number of students	10	18	14	16	12

Table 17.3 : Cumulative Frequency of Weights

Weights (Class)	**Number of students (frequency) f**	**Cumulative frequency**
55-60	10	10
60-65	18	28
65-70	14	42
70-75	16	58
75-80	12	70

Q_1 = ¼ N = ¼ (70) = 17.5^{th} item

This item will fall in class (60-65)

∴ L_1 = 60, C= 10, f = 18, h = 5

With the help of eq^n 17.10

$$Q_1 = 60 + \frac{17.5 - 10}{18} \times 5$$

$$= 60 + \frac{7.5}{18} \times 5$$

= 60 + 2.083 = 62.083

Similarly, Q_3 = ¾× 70 = 52.5^{th} item

It comes in class interval of (70-75). With the help of eq^n 17.11;

$$Q_3 = 70 + \frac{52.5 - 42}{16} \times 5$$

= 70 + 3.28

= 73.28

$\therefore$ Quartile Deviation = $\frac{Q_3 - Q_1}{2} = \frac{73.28 - 62.0823}{2} = \frac{11.197}{2} = 5.5985$

Now, Coefficient of Q.D. = $\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{73.28 - 62.0823}{73.28 + 62.0823} = \frac{11.197}{135.3623} = 0.0827$

17.7.3 MERITS

1. Q.D can be used as a measure of variation to open-ended distributions.
2. As Q.D. is not affected by the presence of extreme values, hence, it is a better measure of variation for highly skewed distribution and distribution with extreme values.

17.7.4 DEMERITS

1. As the Q.D is computed using only 50% of the total observations it cannot be regarded as a good measure of variation.
2. The value of Q.D. is affected by sampling fluctuations.
3. Q.D. is not a true measure of variation as it does not measure the spread of observation from the average.

17.8 MEAN DEVIATION

The two measures of dispersion, range and quartile deviation, discussed earlier in this unit do not tell how values in a data set are scattered about a central value, therefore, it is always better to have such a measure of dispersion which takes into account all the items of a series and is computed in relation to a central value. This method of computing dispersion is known as the ***method of averaging deviation***. In this method, the deviations of items from a measure of central tendency are averaged to get the dispersion of the data set. Mean deviation is such a measure of dispersion. It is also known as the ***first moment of dispersion***. Mean deviation is obtained by calculating the absolute (i.e. ignoring the positive and negative signs) deviations of each observation from mean.

17.8.1 MEAN DEVIATION FOR UNGROUPED DATA

To compute mean deviation for ungrouped data, firstly, the mean of the sample is calculated. Then the absolute value of the difference between each item in the data set and the mean is calculated. i.e., first we subtract the mean from every value in the data set and ignore the positive or negative signs, (consider everything to be positive). Secondly, all

those differences are added and finally this sum is divided by the number of items in the sample.

Thus, Absolute Mean Deviation (AMD) = $\frac{\sum |x - \bar{x}|}{n}$ ---(eq^{n} 17.12)

Where,

x = value of observation

$\bar{x}$ = mean

n = number of observations in the data set

Now, we shall learn the calculation of the mean deviation with the help of following example. Consider the marks obtained by 10 students in a Class test given in Table 17.4

Table: 17.4 Marks obtained by 10 students in a Class test

S.No.(N)	Observation in days (x)
1	10
2	15
3	18
4	20
5	20
6	22
7	23
8	25
9	27
10	30

Table 17.5 : Calculation of Mean Deviation of Marks of 10 Students in a Class Test

S. No.(N)	Observation (x) Marks obtained	Deviation from mean ($x - \bar{x}$)	Absolute deviation ($\lvert x - \bar{x} \rvert$)
1	10	-11	11
2	15	-6	6
3	18	-3	3
4	20	-1	1
5	20	-1	1
6	22	1	1
7	23	2	2
8	25	4	4
9	27	6	6
10	30	9	9
N=10	$\sum x = 210$		$\sum \lvert x - \bar{x} \rvert = 44$

Mean ($\bar{x}$) = $\frac{\sum x}{N}$ = 210/10 = 21 marks

∴ Absolute Mean deviation = $\frac{\sum \lvert x - \bar{x} \rvert}{N}$ = 44/10 = 4.4 marks

17.8.2 MEAN DEVIATION FOR GROUPED DATA

Mean deviation (M.D.) for a grouped or continuous frequency distribution can be calculated about average (mean) using following equation.

M.D. (About the mean $\bar{x}$) = $\frac{1}{N}\sum f\,|\,x - \bar{x}\,|$

$$= \frac{1}{N}\sum f\,|\,d\,| \qquad \text{---(eq}^{n}\text{ 17.13)}$$

Where,

x = mid value of the class interval

f = the corresponding frequency

N= total cumulative frequency

$|\,x - \bar{x}\,|$ = the absolute value of the deviations

$d = (x - \bar{x})$ = difference of the values of x from the mean $\bar{x}$.

Consider the data given in Table 17.6 and compute the Mean Deviation.

Table 17.6

Class Interval	0-4	4-8	8-12	12-16
Frequency	4	2	1	3

Table 17.7 : Computation of Mean Deviation

Class Interval	**Frequency (f)**	**Mid-value of class interval (X)**	**d=X-A**	$f \times d$	$\lvert X - \bar{X} \rvert$	$f\lvert X - \bar{X} \rvert$
0-4	4	2	-6	-24	5.2	20.8
4-8	2	6	-2	-4	1.2	2.4
8-12	1	10	2	2	2.8	2.8
12-16	3	14	6	18	6.8	20.4
	N=10			$\sum f \times d = -8$		= 46.4

Let the value of assumed mean (A) = 8

Actual Mean = Assumed Mean + $\sum f \times d$ /N = 8 + (-8/10) = 7.2

Absolute Mean Deviation = $\frac{1}{N} \sum f \mid d \mid$

Therefore, Absolute Mean Deviation = $\frac{46.4}{10}$ = 4.64

17.8.3 MERITS

1. The calculation of Absolute mean deviation is not very difficult. It is simple and easy to understand.
2. Absolute mean deviation is a more comprehensive measure of dispersion as it is dependent on all observations of a distribution.
3. It is a true measure of dispersion, as it is obtained by taking the average of the deviations from the mean of every observation of the data set.
4. It is not affected very much by the value of the extreme values in the data set.

17.8.4 DEMERITS

1. Absolute Mean deviation is not very reliable as it is based on the arithmetic mean of the absolute values (ignoring the positive and negative signs).
2. Any further algebraic treatment cannot be done with the help of Absolute Mean Deviation.
3. Computation of Absolute Mean deviation for distributions with open-end classes is not possible.

17.9 Questions

(1) What do you understand by dispersion? Explain.

(2) What is range? What are its advantages and disadvantages as a measure of dispersion?

(3) How is mean deviation computed? Discuss its merits and limitations.

(4) The following table gives the height of students. Find the quartile deviation.

Height (inches)	50 - 53	53 – 56	56 – 59	59 – 62	62 – 65	65 – 68
No. of students	02	07	24	27	13	03

(5) Calculate the mean deviation from the following data:

Class Interval	0 – 9	10 – 19	20 – 29	30 – 39	40 – 49	50 – 59
Frequency	15	36	53	42	17	02

UNIT-18 MEASURES OF DISPERSION– II

18.1 INTRODUCTION

In the previous unit (Unit – 17; Measures of Dispersion – I), we learnt that the dispersion hints how much the observations in a given data set are spread out. The simplest of these measures is the range. The range is the distance between the largest and the smallest value in the data set. Of all measures of dispersion, the range is most sensitive to extreme values. Similar to the range, but not influenced by extreme values, is the inter-quartile range. It is the difference between the third and first quartiles. The inter-quartile range is the range of the middle half of the data. The mean deviation shows how much the average value in the data set differs from the mean. The mean deviation is just that - the average distance of the measurements from the mean.

To overcome the drawbacks of the over mentioned measures of dispersion, few other measures of dispersion are discussed in this unit. The standard deviation is one of the rationale and popular measure of dispersion. It is usually calculated by first finding the variance and then extracting the square root of the variance. Although more difficult to compute than the other measures of dispersion, the standard deviation and the variance are the most useful and accurate measures of dispersion.

18.2 VARIANCE

To study the concept of standard deviation, first we have to understand about the variance. It is similar to mean deviation expect for that, it is calculated using the sum of the squared distances between the mean and each observation divided by the total number of elements in the distribution. While calculating variance the differences (deviations) are squared to make them positive.

The formula for calculating variance (σ^2) is as follows.

$$\sigma^2 = \frac{\Sigma f_i \left(x_i - \bar{x}\right)^2}{N} \qquad (eq^n\ 18.1)$$

Where,

σ^2 = Variance

X_i = the value of observation;

$\bar{X}$= Mean;

N = Total cumulative frequency;

f_i= frequency of a class.

18.3 STANDARD DEVIATION

The concept of standard deviation was first introduced by Karl Pearson. It fulfills most of the properties laid down for an ideal measure of dispersion. Standard deviation is the square root of the arithmetic average of the squares of the distances measured from the mean. In other words, it is the square root of the variance. The standard deviation is expressed in same units as those used in the data set. It is conventionally represented by the Greek letter σ (sigma).

σ = Standard Deviation = $\sqrt{\frac{\sum(x-\bar{x})2}{N}}$ (eq^n 18.2)

18.4 CALCULATION OF STANDARD DEVIATION

18.4.1 1 SERIES OF INDIVIDUAL OBSERVATIONS

(1) Direct Method No. 1

Under this method, the following steps are involved:

(i) Find the arithmetic average of the series;

(ii) Find the deviation of each observation from the arithmetic average of the series and represent it by (d) i.e. calculate (X-$\overline{X}$)for each X;

(iii) Square these deviations and sum up them to get $\sum d^2$;

(iv) Divide $\sum d^2$ by the number of observations in the series to get $\frac{\sum d^2}{N}$;

(v) Standard deviation = $\sqrt{\frac{\sum d^2}{N}}$. (eq^n 18.3)

(2) Direct Method No. 2

In this method, we follow these steps:

(i) Calculate the sum of the squares of the items in the series, i.e. find $\sum X^2$;

(ii) Get the sum of all the values to find ($\sum X$) and square it to find $(\sum X)^2$;

(iii) Divide $(\sum X)^2$ by N i.e. find $\frac{\sum X^2}{N}$;

(iv) S. D. = $\sqrt{\frac{\sum X^2}{N} - (\frac{\sum X}{N})^2}$ (eq^n 18.4)

Short cut Method

$$\text{S. D.} = \sqrt{\frac{\sum dx^2}{N} - \left(\frac{\sum dx}{N}\right)^2} \qquad (\text{eq}^{n}\ 18.5)$$

Where, dx = (X – A).

Example 18.1 Calculate the standard deviation of the weights of the 10 persons given below:

Weights (in kg): 60, 60, 61, 62, 63, 63, 63, 64, 64, 70.

Solution: Calculation of standard deviation of weights:

(With the help of Direct Method 1)

Weights in kg (X)	Deviation from mean 63 (d)	Deviation squared (d^2)
60	-3	9
60	-3	9
61	-2	4
62	-1	1
63	0	0
63	0	0
63	0	0
64	+1	1
64	+1	1
70	+7	49
$\sum X = 630$		$\sum d^2 = 74$

Arithmetic Average = $\frac{\sum X}{N} = \frac{630}{10} = 63$ kg.

Standard Deviation = $\sqrt{\frac{\sum d^2}{N}} = \sqrt{\frac{74}{10}} = \sqrt{7.4} = 2.27$ kg.

(With the help of Direct Method 2)

Weights in kg (X)	(X^2)
60	3600
60	3600
61	3721
62	3844
63	3969
63	3969
63	3969
64	4096
64	4096
70	4900
$\sum X = 630$	$\sum X^2 = 39{,}764$

Standard Deviation = $\sqrt{\frac{\sum X^2}{N} - (\frac{\sum X}{N})^2}$

$= \sqrt{\frac{39764}{10} - (\frac{630}{10})^2}$

$= \sqrt{3976.4 - (63)^2}$

$= \sqrt{3976.4 - 3969} = \sqrt{7.4} = 2.72$ kg.

18.4.2 CALCULATION OF STANDARD DEVIATION OF DISCRETE SERIES

(A) Direct Method

(i) Calculate the arithmetic mean;

(ii) Get the deviations of (d) various values, from the mean value;

(i) Square these deviations i.e. get (d^2);

(ii) Multiply d^2 with the respective frequencies (f);

(iii) Sum up all these values i.e. ($\sum fd^2$);

(iv) Divide ($\sum fd^2$) by the total no. of items in the series (N);

(v) Find out the square root of the above to get standard deviation i.e. find $\sqrt{\frac{\sum fd^2}{N}}$. ---------- (eq^n 18.6)

(B) Short cut Method

(i) Assume a mean (A) and get deviations (dx) from this assumed mean and get dx^2;

(ii) Multiply dx^2 with the respective frequencies f to find (fdx^2);

(iii) Calculate ($\sum fdx^2$);

(iv) Compute $\frac{(\sum fdx2)}{N}$;

(v) From $\frac{(\sum fdx2)}{N}$ subtract the square of the difference between actual and Assumed Mean i.e. $(\bar{X} - A)^2$;

(vi) Now to get the value of Standard Deviation, take the square root of the above value i.e. the formula to compute S.D. will be:

S.D. = $\sqrt{\frac{(\sum fdx2)}{N} - (\bar{X} - A)^2}$ (eq^n 18.7)

Oralternatively, S.D. = $\sqrt{\frac{(\sum fdx2)}{N} - (\frac{\sum fdx}{N})^2}$ (eq^n 18.8)

As, $(\bar{X} - A) = \frac{\sum fdx}{N}$

Example 18.2: Find the Standard Deviation of the following discrete series:

X	Frequency	X	Frequency
18	2	23	17
19	7	24	13
20	10	25	8
21	14	26	5
22	18	27	4

Solution:

X	Frequency (f)	(f.X)	Deviation from the Average (22) (d)	d^2	fd^2
18	2	36	-4	16	32
19	7	133	-3	9	63
20	10	200	-2	4	40
21	14	294	-1	1	14
22	18	396	0	0	0
23	17	391	+1	1	17
24	13	312	+2	4	52
25	8	200	+3	9	72
26	5	130	+4	16	80
27	4	108	+5	25	100
	N = 100	$\sum fX = 2200$			$\sum fd^2 = 470$

Arithmetic Average = $\frac{\sum fX}{N} = \frac{2200}{100} = 22.$

Standard Deviation = $\sqrt{\frac{\sum fd^2}{N}} = \sqrt{\frac{470}{10}} = \sqrt{47} = 6.856.$

18.4.3 CALCULATION OF STANDARD DEVIATION OF CONTINUOUS SERIES

In a continuous series, the data are given in class intervals. We take middle points of the respective class intervals. Further, the technique of computing standard deviation is exactly same as in discrete series. We calculate S.D. in the following manner:

(i) Find the middle value of each class interval and represent it by *x*;

(ii) Calculate the deviations of each middle value from some value and show these deviations as *dx*;

(iii) S.D. = $\sqrt{\frac{\sum(f.dx2)}{N} - \{\frac{\sum(f.dx)}{N}\}^2}$ -----(eq^n 18.9)

(iv) If the deviations are divided by some common factor, say *i*define*d'x*

d'x = dx/i then

S.D. = $\sqrt{\frac{\sum(f.dx2)}{N} - \{\frac{\sum(f.dx)}{N}\}^2 \times i}$ -----(eq^n 18.10)

Example 18.3: Compute the S.D. from the data given in the following table representing the age distribution in a social-cause organization.

Age	No. of Persons
20 – 30	3
30 – 40	61
40 – 50	132
50 – 60	153
60 – 70	130
70 – 80	51
80 – 90	2
Total	532

Solution: Calculation of S.D. of the age distribution of 532 members of the social-cause organization:

Age Group (X)	Mid Value = *x*	Frequency (*f*)	Deviations from assumed mean(55) *dx*	*f.dx*	dx^2	$f.dx^2$
20 – 30	25	3	- 30	-90	900	2700
30 – 40	35	61	-20	-1220	400	24400
40 – 50	45	132	-10	-1320	100	13200
50 – 60	55	153	0	0	0	0

60 – 70	65	130	+10	+1300	100	13000
70 – 80	75	51	+20	+1020	400	20400
80 – 90	85	2	+30	+60	900	1800
		N = 532		$\sum f.dx=250$		$\sum fdx^2 =75500$

$$\text{S.D.} = \sqrt{\frac{\sum(f.dx2)}{N} - \{\frac{\sum(\text{f.}dx)}{N}\}^2}$$

$$= \sqrt{\frac{75500}{532} - (\frac{-250}{532})^2}$$

$$= \sqrt{141.92 - (-0.47)2} = \sqrt{141.699} = 11.9 \text{ years.}$$

18.5 PROPERTIES OF STANDARD DEVIATION

1. As it is possible to compute combined mean of two or more data sets, it is also possible to calculate combined standard deviation of two or more data sets. Combined standard deviation denoted by σ_{12} is computed as follows:

$$= \sigma_{12} = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1d_1^2 + n_2d_2^2}{n_1 + n_2}} \text{ -----(eq}^n \text{ 18.11)}$$

Where,

σ_1 = standard deviation of first data set

σ_2 = standard deviation of second data set

$d_1 = \overline{x1} - x$

and, $d_2 = \overline{x2} - x$

$$\bar{x} = \frac{n1\overline{x1} + n2\overline{x2}}{n1 + n2}$$

2. **The value of standard deviation does not change, if in a series each of the observation is increased or decreased by a constant quantity. In statistical language, we say that standard deviation is independent of change of origin.**

For example, for the observations 3, 10 and 12

$\bar{x}$ = 8.33, σ = 3.859

If we increase the value of each observation by 5.5 we get the observations 8.5, 15.5 and 17.5.

Now $\bar{x}$ = 13.833

σ = 3.859

Hence, although $\bar{x}$ has increased by 5.5, σ remains the same.

3. **If each observation in a data set is multiplied or divided by a constant quantity, standard deviation will also beaffected in the same way. In statistical language we say that standard deviation is dependent of change of scale.**

Consider the above example again, where observations were 3, 10 and 12.

σ = 3.859 as shown in the above calculation.

Suppose we multiply each observation by 6, the observations become 18, 60 and 72.

$\bar{x}$ = 50

$$\sigma = \sqrt{\frac{(18-50)^2+(60-50)^2+(72-50)^2}{3}}$$

= 23.152.

It is shown that the standard deviation has also been multiplied by 6 i.e. $3.859 \times 6 = 23.152$.

Similarly if the observations 3, 10 and 12 with σ = 3.859, are divided by a constant value, say 2. The observations now become 1.5, 5 and 6.

$\bar{x}$ = 4.16.

$$\sigma = \sqrt{\frac{(1.5-4.16)^2+(5-4.16)^2+(6-4.16)^2}{3}}$$

=1.92

Thus the standard deviation is also divided by 2 i.e. (3.859 ÷ 2) =1.92.

Hence, we conclude that the standard deviation is independent of any change of origin, but is dependent on the change of scale.

4. Standard deviation is the minimum root-mean- square deviation. In other words, the sum of the squares of the deviations of

observations of any data set from a value other than the arithmetic mean would always be greater.

18.6 MERITS AND DEMERITS OF STANDARD DEVIATION

Merits :

(i) Standard deviation is rigidly defined and its value is always definite.

(ii) SD is less influenced by the fluctuations of sampling than any other measure of dispersion.

(iii) SD is based on all the observations of data set.

(iv) It is possible to use SD for further algebraic treatment.

(v) SD overcomes the difficulty of algebraic signs of mean deviation successfully.

Demerits :

(i) SD is not very easy to calculate. It is cumbersome in computation, in comparison to other measures of dispersion.

(ii) SD considers more weight to extreme observations and less weight to those observations which are near to mean. The deviations 2 and 8 are in the ratio 1:4, however, their squares 4 and 64 would be in the ratio of 1:16.

Although the Standard Deviation is not free from certain limitations, yet it is the best measure of dispersion.

18.7 COEFFICIENT OF VARIATION

The Standard Deviation is an absolute measure of dispersion. It is expressed in terms of the unit of the given data set or series. For the purpose of comparison, we need a relative measure which is free from unit.

The coefficient of variation is a measure of relative dispersion and is given by:

$$\text{Coefficient of Variation} = \frac{\text{Standard deviation}}{\text{Mean}} \text{ -----(eq}^{n}\text{ 18.12)}$$

It is generally expressed as percentage

$$\text{i.e., Coefficient of variation (\%)} = \frac{\text{Standard deviation}}{\text{Mean}} \times 100 \text{ -----(eq}^{n}\text{ 18.13)}$$

Hence the coefficient of variation measures the spread of a set of data as a proportion of its mean. When we want to compare the variability, homogeneity, stability, uniformity and consistency of two or more data sets, it is used. The data set for which the coefficient of variation is greater is said to be more variable i.e. less consistent or less homogeneous. On the other hand, if the coefficient of variation is less it is said to be less variable i.e., more consistent or more homogeneous data set.

Example 18.4: Companies A and B are manufacturing CFLs. The length of life of their CFLs is given in the following table. You are required to decide, the CFL of which Company is more consistent from the perspective of length of life?

Length of Life (Hours)	Company A	Company B
5000-7000	50	40
7000-9000	110	300
9000-11000	260	120
11000-13000	100	80
13000-15000	80	60
	600	600

Solution:

Computation of Mean and Standard Deviation

Length of Life (Hours)	Mid Value X	Step Deviation dx	**Company A**			**Company B**		
			f	fdx	fdx^2	f	fdx	fdx^2
5000-7000	6000	-2	50	-100	200	40	-80	160
7000-9000	8000	-1	110	-110	110	300	-300	300
9000-11000	10000	0	260	0	0	120	0	0
11000-13000	12000	+1	100	+100	100	80	+80	80
13000-15000	14000	+2	80	+160	320	60	+120	240
			600	+50	730	600	-180	780

Company A:

$$\textbf{Mean} = 10000 + \left(\frac{50}{600} \times 2000\right) = 10160.67 \text{ hours}$$

$$\textbf{S.D.} = \sqrt{\frac{730}{600} - \left(\frac{50}{600}\right)^2} \times 2000 = 2200 \text{ hours.}$$

$$\textbf{Coefficient of Variation or C.V.} = \frac{\mathbf{2200}}{\mathbf{10160.67}} \times 100 = \mathbf{21.64}$$

Company B:

$$\textbf{Mean} = 10000 + \left(\frac{-18}{600} \times 2000\right) = 9400 \text{ hours}$$

$$\textbf{S.D.} = \sqrt{\frac{780}{600} - \left(\frac{-18}{600}\right)^2} \times 2000 = 2200 \text{ hours.}$$

$$\textbf{Coefficient of Variation or C.V.} = \frac{\mathbf{2200}}{\mathbf{9400}} \times 100 = \mathbf{23.41}$$

On the basis of the values of coefficient of variation of two companies, we conclude that the CFLs of the Company A are more consistent in comparison to Company B.

E**xample 18.5:** Compute the Variance, Standard Deviation and Coefficient of Variation given the profitability of 500 companies given in the following table:

Profitability of 50 Companies

Profit in crores (x_i)	**Number of companies (f_i)**
10	150
15	100
20	150
25	60
30	40

$$\bar{x} = \frac{(10\times150)+(15\times100)+(20\times150)+(25\times60)+(30\times40)}{150+100+150+60+40} = \mathbf{17.4}$$

Computation of Variance and Coefficient of Variation

xi	f_i	$f_i x_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$f_i (x_i - \bar{x})^2$
10	150	1500	-7.4.	54.76	8210.40
15	100	1500	-2.4	5.76	570.60
20	150	3000	2.6	6.76	1010.40
25	60	1500	7.6	57.76	3460.56
30	40	1200	12.6	158.76	6350.04
Total	**500**	**8700**			**19620.00**

Mean = 8700 / 500 = 17.4

Variance = 19620 / 500 = 39.24

So variance of profitability among 50 companies is 39.24

Now, Standard Deviation of profits of 50 companies is:

SD = $\sqrt{39.24}$ = 6.26

Coefficient of variation=$\frac{6.26}{17.4}$ ×100 = 36%

18.8 LORENZ CURVE

Yet another measure of dispersion is Lorenz curve method. It is a graphical method to study dispersion. This concept was given by Dr. Lorenz who first studied the dispersion of distribution of wealth by graphical method. This technique involves the following steps:

(i) The size of the observations as well as frequencies are cumulated;

(ii) Take the final cumulated figures as 100 percentages are computed for all cumulative figures;

(iii) Calibrate X-axis starting from 100 – 0 and Y-axis starting from 0 - 100 to represent frequencies;

(iv) Draw a line from 0 of the X-axis to the 100 of Y-axis. This line is known as line of equal distribution;

(v) Now, plot the different points of x and y and draw the curve.

Conclusion: The drawn curve will be away from the line of equal distribution, if the distribution is not equal proportionately. The farther the curve from the line of equal distribution, the greater is the variability in the given data set. Similarly, the closer the curve to the line of equal distribution, he lesser is the variability in the series or data set.

Although to study dispersion through Lorenz curve method is very easy and simple to understand, yet it has a great limitation. Lorenz curve is merely a pictorial representation of the dispersion of the series; it does not provide any numerical value of the measure of dispersion. Therefore, it is suggested to be used along with any numerical measure of dispersion.

18.9 QUESTIONS

(1) What is Standard Deviation? Explain its important properties.

(2) Define Coefficient of Variation. What purpose does it serve? Distinguish between 'Variance' and 'Coefficient of Variation'.

(3) Find out the standard deviation of the following series:

Size of holdings (hectares)	No. of Cultivators
0 – 20	100
20 – 40	230
40 – 60	360
60 – 80	240
80 – 100	170
100 – 120	300
120 – 140	50

(4) Goals scored by two teams A and B in a football championship were as follows:

No. of Goals scored in a Match		0	1	2	3	4
No. of Matches playedby teams	Team A	27	9	8	5	4
	Team B	17	9	6	5	3

Draw inference about the consistency of the performance of Team A and Team B.

UNIT-19 MEASURES OF SKEWNESS-I

19.1 INTRODUCTION

In the previous units of this block, we discussed measures of dispersion (deviation) to explain the spread of values for a given set of data around a central value. However, the measure of dispersion does not talk about the degree to which these individual values in the data set disperse from symmetry on both sides of the central value. A frequency distribution of the data set that is not symmetrical or normal is called skewed (asymmetrical). If the given frequency distribution is skewed, the tail of the curve is lengthened by the extreme values of data set. These extreme values rest aside either towards the right tail or towards the left tail of the distribution. When these extreme values move towards the right or upper tail the distribution is known as positively skewed. Similarly, when these values move towards the left or lower tail, the distribution is termed as negatively skewed.

Skewness is the opposite of symmetry. If skewness is present in a distribution, it means that distribution is not symmetrical or skewed. The averages tell us that us about the clustering of items (data) around the central value whereas measures of skewness tells us whether the dispersal of items (data) from an average is symmetrical.

Now, we can define skewness in the following way:

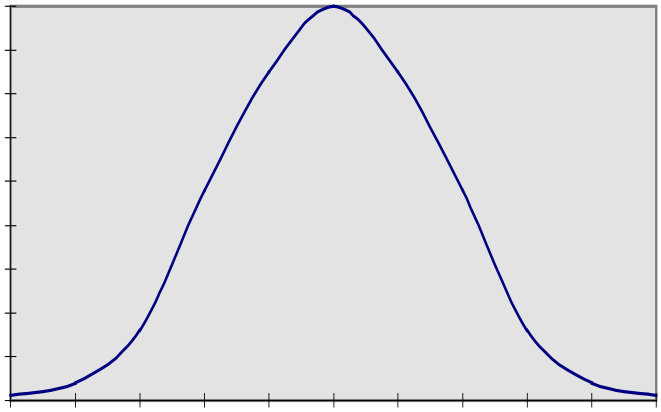
19.2 DEFINITIONS

Riggleman and Frisbee defined skewness as, “Skewness is lack of symmetry. When a frequency distribution is plotted on a chart, skewness present in the items teds to disperse the chart more on one side of the mean than on the other.”

In the words of Simpson and Kafka, “Measures of skewness tell us the direction and the extent of skewness. In a symmetrical distribution, he mean, median and mode are identical. The more the mean moves away from the mode, the larger the asymmetry or skewness.”

Thus we can conclude that the skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetrical if it looks the same to the left and right of the center point or value.

Pictorial display of the shape of a symmetrical curve is depicted in the Figure 19.1. It is bell shaped curve, and in it, there is no skewness. The value of mean, median and mode in such a distribution would coincide.



Mean = Median = Mode

Figure : 19.1 (Symmetrical disribution)

19.3 POSITIVE AND NEGATIVE SKEWNESS

We know that the three popular measures of Central tendency namely mean, median and mode are strongly affected by extreme values of the data set. Out of these three measures of central tendency, the mean is got influenced highly whereas the mode is not influenced. The median is affected by the numbers but not y the extreme values of the data set. Like mean, median is also shifted in the same direction of extreme valued data but not as far as the mean. Figure 19.2 and Figure 19.3 show this phenomenon.

For a positively skewed distribution:

Arithmetic Mean > Median > Mode;

Whereas for a negatively skewed distribution:

Arithmetic Mean < Median < Mode.

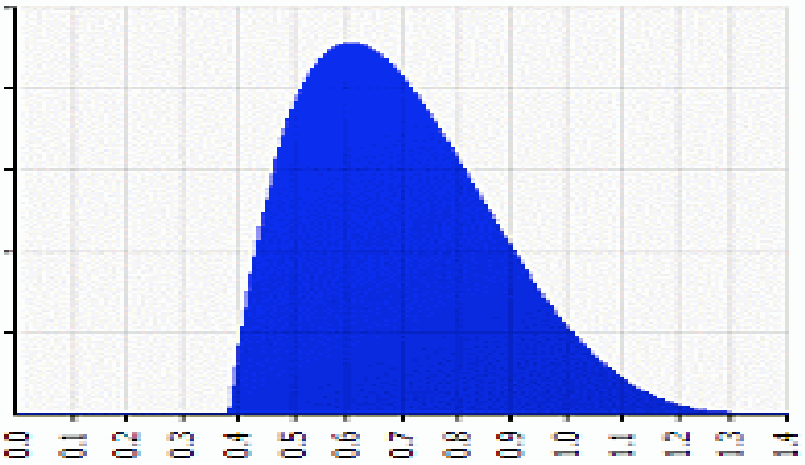


Figure : 19.2 (Positvily Skewed)

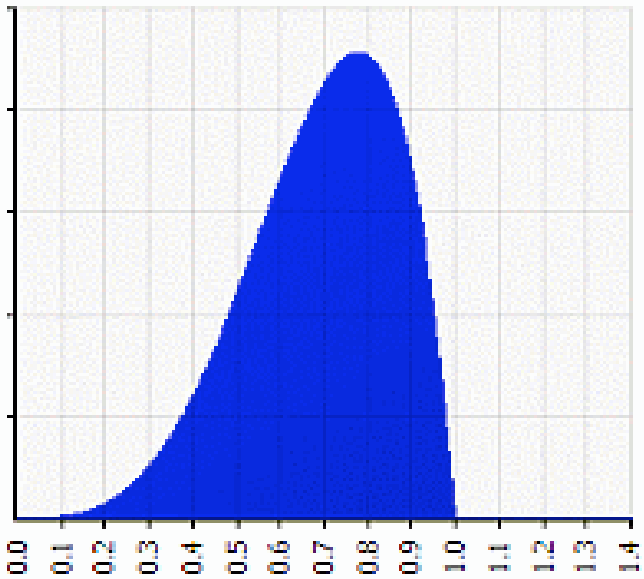


Figure : 19.3 (Negatively Skewed)

19.4 DIFFERENCE BETWEEN VARIATION AND SKEWNESS

(1) Variation or dispersion deals with the spread of values around the central value. Skewness, on the other hand, deals with symmetry of distribution around central value.

(2) Variation tells the amount of dispersion or spread of the individual values in a distribution around a central value, whereas Skewness hints about the direction of that spread or dispersion. Skewness explains how much it is near or away from symmetry.

(3) The extent of variation among individual values in a distribution is computed with the help of Variation, while the concentration of lower or higher values around the mean value is indicated by the Skewness.

19.5 QUESTIONS

1. What do you understand by skewness?
2. Explain the concept of skewness. Draw the sketch of a skewed frequency distribution and show the approximate position of the mean, median and mode.
3. Distinguish clearly between dispersion and skewness.

UNIT-20 MEASURES OF SKEWNESS – II

20.1 INTRODUCTION

In our previous unit, we understood that a frequency distribution of the data set that is not symmetrical or normal is called skewed (asymmetrical). In order to study a frequency distribution, it would be very significant to know whether it would follow normal distribution, and if not, to what extent it would deviate from a normal distribution. To get the true picture of the given frequency distribution, measure of central tendency and measures of dispersion should always be supplemented by measures of skewness. With the help of measures of skewness, we shall be able to know the direction and extent of skewness of the distribution in the real sense.

In this unit, we shall focus on the tests of skewness as well as popular methods available to measure skewness.

20.2 TESTS OF SKEWNESS

There are certain tests available to check whether the given data set or distribution is skewed or not. These tests are as follows:

(i) A skewed frequency distribution would not be a bell shaped curve, when it is plotted on the graph paper.

(ii) In a skewed distribution, values of mean, median and mode would never coincide. The mean and mode would be at a significant distance to each other and the median would be lying between them. For a distribution which is moderately asymmetrical

$$Mean - Mode = 3(Mean - Median)$$

⇨ 3.Median = 3. Mean + Mode – Mean = 2.Mean + Mode

⇨ = 2.Mean + Mode + 2Mode – 2Mode

⇨ = 3Mode + 2(Mean - Mode)

⇨ **Median = Mode + ⅔.(Mean – Mode)**

(c) As we know that in a symmetrical distribution, the two quartiles are at quidistant from the median. In other words, (Q_3 – Median) = (Median – Q_1). This property of symmetrical distribution may be used as a test for asymmetrical distribution. That is to say that for skewed distribution (Q_3 – Median) - (Median – Q_1) would not be zero.

(d) For a skewed distribution, frequencies are not equally distributed at various points which are equidistant from the Mode. In case of moderately aymmetrical distribution:

$$Mean - Mode = 3(Mean - Median)$$

(e) In a skewed distribution, the sum of positive deviations are not equal to the sum of negative deviations.

20.3 MEASURES OF SKEWNESS

The degree of in a distribution of a given data set may be measured in two ways:

- Absolute Measure of Skewness
- Relative Measure of Skewness

Absolute Measure of Skewness

As we know that the mean is equal to mode for a symmetrical distribution. Hence, the distance between mean and mode is measured to get the degree of skewness for an asymmetrical distribution.

Thus,

$$\text{Absolute } S_k = \text{Mean} - \text{Mode}$$

$$= Q_3 - Q_1 - 2.\ \text{Median}$$

For a positively skewed distribution, Mode < Mean, therefore S_k comes out a positive value.

And for a negatively skewed distribution, Mode > Mean, therefore S_k comes out a negative value.

However, this absolute measure of skewness is not very suitable and desirable due to under mentioned reasons:

(a) When the given distributions are having different measuring units, this absolute measure of skewness cannot be used for making comparison the skewness of these distributions;

(b) When the difference between mean and mode be different or not same (i.e. this difference may be small in one distribution and large in another distribution), although the shape of these frequency curves is not different or same.

Hence, these two limitations of absolute measure of skewness, is removed with the help of relative measure of skewness.

Relative Measure of Skewness

Relative measure of skewness removes the shortcomings of absolute measure of skewness. It enables us to make valid comparison between skewness of two or more distributions. Relative measures of skewness are

obtained by dividing the absolute measures by any one of the measures of dispersion. Relative measures of skewness are known as coefficient of skewness.

There are three significant ways to compute the relative measure of skewness:

(1) Karl Pearson's Coefficient of Skewness

(2) Bowley's Coefficient of Skewness

(3) Kelly's Coefficient of Skewness

We discuss these measures of skewness one by one.

(1) Karl Pearson's Coefficient of Skewness

The method suggested by Karl Pearson to measure coefficient of skewness is based on the difference between mean and mode divided by standard deviation. It is given by:

$$Skp = \frac{Mean - Mode}{Standard\ deviation} \qquad (eqn\ 20.1)$$

Where, Sk_p is the Karl Pearson's coefficient of skewness.

For a distribution which is moderately skewed, the following relationship holds good:

$$Mean - Mode = 3(Mean - Median)$$

$$\text{or,} \quad Mode = (3\ Median - 2\ Mean)$$

therefore,

$$Skp = \frac{3(Mean - Median)}{Standard\ deviation} \qquad (eqn\ 20.2)$$

The value of the above coefficient of skewness would be 0 for a symmetrical distribution and for an asymmetrical or skewed distribution, it would be pure number. For a moderately skewed distribution, value of Sk_p varies between +1 to -1.

(2) Bowley's Coefficient of Skewness

According to Prof. Bowley, if the given distribution is symmetrical, Q_1 and Q_3 would be at equal distances from the Q_2 or Median value, that is

$$Median - Q1 = Q3 - Median$$

$$\text{or,} \quad Median = \frac{Q1+Q3}{2}$$

Hence, the value of the median is the mean value of Q_1 and Q_3.

Therefore, for a symmetrical distribution, the coefficient will be zero.

For an asymmetrical distribution, quartiles do not fall on the equal distances. When the distribution is positively skewed, (Q_1 – Median) > (Q_3 - Median).

Similarly for a negatively skewed distribution, (Q_1 – Median) < (Q_3 - Median).

Prof. Bowley converted the absolute value of Skewness into relative value by dividing the inter quartile range. That is,

$$Relative\ Skb = \frac{Q3 + Q1 - 2.Median}{Q3 - Q1} \qquad (eqn\ 20.3)$$

The value of Sk_b varies between ± 1 for the moderately skewed distribution. This coefficient is also a pure number.

(3) Kelly's Coefficient of Skewness

Kelly's Coefficient of Skewness is a slight modification of Bowley's Coefficient of skewness. According to Kelley, instead of quartiles, deciles and percentiles should be considered in the calculation of skewness. Any two deciles or any two percentiles which are equidistant from the Median may be used. However, he emphasized the use of first and ninth decile or tenth and ninetieth percentile for measuring skewness. Thus, Kelly's Coefficient of Skewness is given by:

$$Skk = \frac{D9 + D1 - 2.Median}{D9 - D1} \qquad (eqn\ 20.4)$$

or in terms of percentiles,

$$Skk = \frac{P90 + P10 - 2.Median}{P90 - P10} \qquad (eqn\ 20.5)$$

This method suggested by Kelley is not a popular method to calculate coefficient of skewness.

Example 20.1: For a moderately skewed distribution, Mean = 150, Mode = 140, S.D. = 45.

Find(i)Coefficient of variation;

(ii)Pearson's coefficient of skewness;

(iii)Median.

Solution:

$$Coefficient\ of\ Variation\ (C.V.) = \frac{Standard\ deviation}{Mean} * 100$$

$$C.V. = \frac{45}{150} * 100 \ = 30\%$$

Now, Pearson's coefficient of skewness is calculated by;

$$Skp = \frac{Mean - Mode}{Standard\ deviation} \ = \frac{10}{45} = 0.22$$

Median may be computed as:

$$Mean - Mode = 3(Mean - Median)$$

or, (150 – 140) = 3.(150 - Median)

⇨ Median = 150 – 10/3 = 146.7 **Answer**.

Example 20.2: For a negatively skewed distribution, the measures of central tendency are as follows: Mean = 26, Median = 27, Mode = 23. Do you agree with these data?

Solution: We know that, for a negatively skewed distribution, the following condition should be fulfilled:

Mode > Median > Mean

Which is not true for the given data, therefore, we may conclude that the given data is not correct.

Example 20.3: For the data given in the following table

Class Interval	Frequency
130 - 134	3
135 - 139	12
140 - 144	21
145 - 149	28
150 - 154	19
155 - 159	12
160 - 164	5

Calculate the following:

(i) Arithmetic Mean;

(ii) Standard Deviation;

(iii) Karl Pearson's coefficient of skewness.

Solution:

Class Interval	Mid Value X	Frequency (f)	(x – 147)/5 (dx)	dx^2	(f.dx)	$f.dx^2$
130 - 134	132	3	-3	9	-9	27
135 - 139	137	12	-2	4	-24	48
140 - 144	142	21	-1	1	-21	21
145 - 149	147	28	0	0	0	0
150 - 154	152	19	+1	1	+19	19
155 - 159	157	12	+2	4	+24	48
160 - 164	162	5	+3	9	+15	45
		N = 100			$\sum$(f.dx)=+4	$\sum(f.dx^2) = 208$

(i) **Arithmetic Mean** = $A + \left[\frac{\sum(f.dx)}{N} \times i\right]$

= 147 + (4/100) ×5 = 147.2

(ii) **Standard Deviation** = $\sqrt{\frac{\sum(f.dx2)}{N} - \{\frac{\sum(f.dx)}{N}\}^2 \times i}$

$= \sqrt{\frac{208}{100} - (\frac{4}{100})^2 \times 5} = 7.21$

Modal Class is 145–149 (Class limits would be 144.5 – 149.5)

Mode = $L1 + \frac{(f1-f0)}{(2.f1-f0-f2)}\ (L2 - L1)$

$= 144.5 + \frac{(28-21)}{(56-21-19)}\ (149.5 - 144.5) = 146.69$

(iii) **Karl Pearson's coefficient of skewness:**

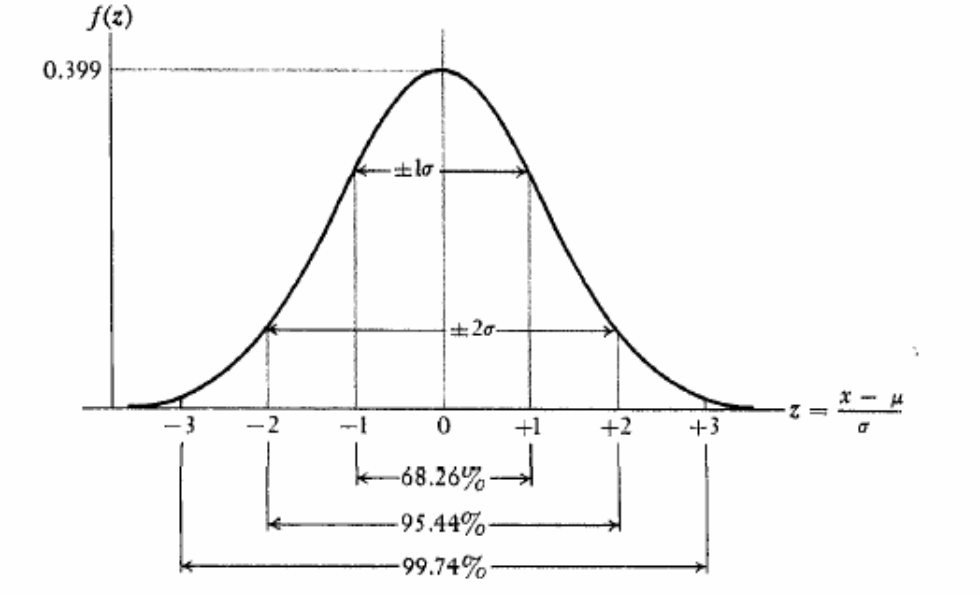
$$Skp = \frac{Mean - Mode}{Standard\ deviation} = \frac{(147.2 - 146.69)}{7.21}$$
$$= +0.07$$

Answer.

20.4 NORMAL CURVE AND NORMAL DISTRIBUTION

The normal probability distribution occupies a place of central importance in *Modern Statistical Theory*. This distribution was first observed as the *normal law of errors* by the statisticians of the eighteenth century. They found that each observation X involves an error term which is affected by a large number of small but independent chance factors. This implies that an observed value of X is the sum of its true value and

the net effect of a large number of independent errors which may be positive or negative each with equal probability. The observed distribution of such a random variable was found to be in close conformity with a continuous curve, which was termed as the *normal curve of errors* or simply the *normal curve.*



When a set of data is collected, the data forms a distribution, meaning that each data point has a particular value along some dimension. Together the data points can form a visual shape of the distribution.Since Gauss used this curve to describe the theory of accidental errors of measurements involved in the calculation of orbits of heavenly bodies, it is also called as Gaussian curve.

Random variables observed in many phenomena related to economics, business and other social as well as physical sciences are often found to be distributed normally. For example, observations relating to the life of an electrical component, weight of packages, height of persons, income of the inhabitants of certain area, diameter of wire, etc., are affected by a large number of factors and hence, tend to follow a pattern that is very similar to the normal curve.

20.4.1 THE CONDITIONS OF NORMALITY

In order that the distribution of a random variable X is normal, the factors affecting its observations must satisfy the following conditions:

(i) ***A large number of chance factors :*** The factors, affecting the observations of a random variable, should be numerous and equally probable so that the occurrence or non-occurrence of any one of them is not predictable.

(ii) **Homogeneity :**The factors must be similar over the relevant population although, their incidence may vary from observation to observation.

(iii) **Independence :**The factors, affecting observations, must act independently of each other.

(iv) **Symmetry :**Various factors operate in such a way that the deviations of observations above and below mean are balanced with regard to their magnitude as well as their number.

20.4.2 PROPERTIES OF NORMAL PROBABILITY CURVE

A normal probability curve or normal curve (depicted in Figure 20.2) has the following properties:

1. It is a bell shaped symmetrical curve about the Mean. There is no skewness in it.
2. It is unimodal curve and its tails extend infinitely in both directions, i.e., the curve is asymptotic to X axis in both directions.
3. All the three measures of central tendency coincide, i.e., mean = median = mode

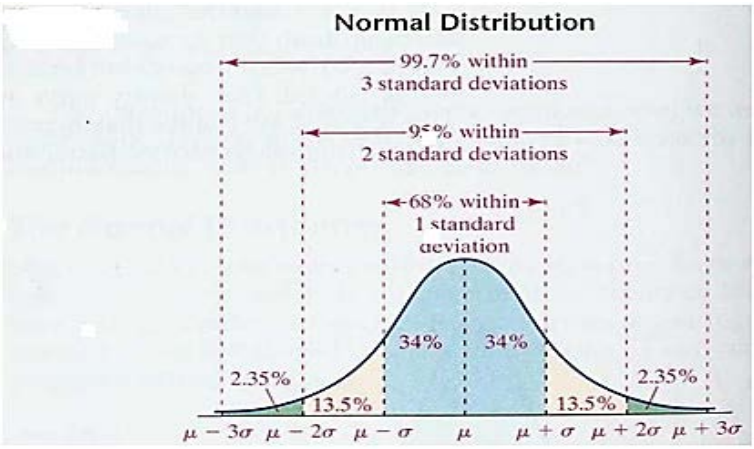


Figure 20.2

4. The height of the normal curve is maximum at the mean value. This ordinate divides the normal curve in two equal and identical parts.

5. The total area under the curve gives the total probability of the random variable taking values between -□□ to □□

6. The value of p(X) is always non-negative for all values of X, i.e., the whole curve lies above X axis.

7. 7The points of inflection (the point at which curve changes its direction) of the curve are each at a distance of one standard deviation from the mean.

8. The quartiles are equidistant from median, i.e., (Median - Q1) = (Q3 – Median) , by virtue of symmetry.

9. Semi-inter quartile range is equal to the probable error which is 0.6745 of standard deviation.

10. 68.27% of values fall within 1 SD of the mean; 9545% of values fall within 2 SD of the mean; and 99.73% of values fall within 3 SD of the mean..

11. Since the distribution is symmetrical, all odd ordered central moments are zero.

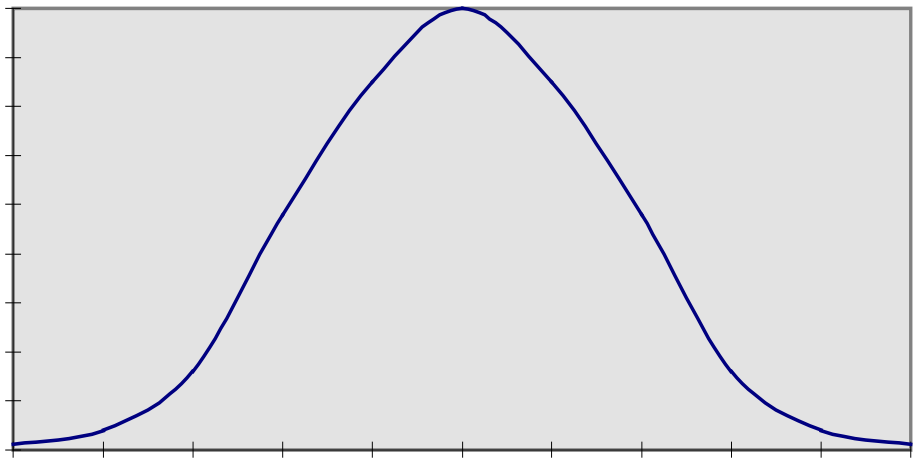


Figure : 20. Normal Curve

20.5 QUESTIONS

(1) Explain the different tests for skewness.

(2) What do you understand by skewness? What are the various methods of measuring skewness?

(3) For a distribution, Karl Pearson's coefficient of skewness is 0.32. Its standard deviation is 6.5 and mean is 29.6. Find the values of Mode and Median.

(4) Calculate Karl Pearson's coefficient of skewness from the following data:

Class Interval	**Frequency**
70 – 80	11
60 – 70	22
50 – 60	30
40 – 50	35
30 – 40	21
20 – 30	11
10 – 20	6
0 - 10	5

(6) Describe the normal distribution and explain its importance in statistics.

Notes

Notes