

PONDICHERRY UNIVERSITY

DIRECTORATE OF DISTANCE EDUCATION



BUSINESS STATISTICS

Paper Code : BCOM 2002
BBA 2002

BACHELOR OF
COMMERCE
/BACHELOR OF BUSINESS
ADMINISTRATION
II - YEAR

DDE - Where Innovation is a Way of Life



DIRECTORATE OF DISTANCE EDUCATION

PONDICHERY UNIVERSITY

BACHELOR OF BUSINESS ADMINISTRATION (BBA)

Paper – VI BBA - 2002

BACHELOR OF COMMERCE (BCom)

Paper – VII BCom - 2002


SECOND YEAR

BUSINESS STATISTICS

UNITS: I – VIII

ht Reserved

ate Circulation Only



BBA Paper – VI / BCom Paper – VII

Business Statistics

BBA - 2002 / BCom - 2002

Units I – VIII written by

Ms. Malabika Deo

Professor
Dept. of Commerce

Pondicherry University

Pondicherry - 605 014

Printed by : Vijaay Offset Printer, Madurai -16.

August - 2016 Copies : 1200

DIRECTORATE OF DISTANCE EDUCATION

PONDICHERRY UNIVERSITY

BBA Paper VI BCom Paper VII BBA - 2002 / BCom - 2002

Business Statistics

TABLE OF CONTENTS

	TITLE	Page No.
UNIT - I		
1	MEANING, DEFINITION & SCOPE OF STATISTICS	1
2	FUNCTION, IMPORTANCE, MISUSES AND LIMITATIONS OF STATISTICS	21
UNIT - II		
1	STATISTICAL ENQUIRY	41
2	COLLECTION OF DATA	50
3	SAMPLING AND SAMPLING METHODS	74
4	CLASSIFICATION AND TABULATION OF DATA	92
UNIT - III		
1	MEASURES OF CENTRAL TENDENCIES	113
2	TYPES OF AVERAGES	124
UNIT - IV		
1	MEASURES OF DISPERSION	174
2	MEASUREMENT OF SKEWNESS AND KURTOSIS	214
UNIT - V	CORRELATION ANALYSIS	236
UNIT - VI	REGRESSION ANALYSIS	226
UNIT - VII		
1	INTERPOLATION AND EXTRAPOLATION	291
2	TIME SERIES ANALYSIS	311
3	BUSINESS FORECASTING	339
UNIT - VIII	PROBABILITY	346
	REVIEW QUESTIONS	371

PAPER VII- BUSINESS STATISTICS

UNIT I

Statistics – A Conceptual Framework – Meaning and Scope of Business Statistics – Definition – Function – Role of Statistics for Business Decisions – Importance - Limitations.

UNIT II

Statistical Enquiry and Methods of Sampling – purpose, Types, Collection of Data – Methods of Enumeration – Sampling Need – Method of Sampling – Merits and Demerits – Classification and Tabulation of Data.

UNIT III

Measures of Central Tendency – Average – Objectives of an Average – Types – Characteristics – Merits and Demerits – Mean, Median, Mode – Geometric Mean – Harmonic Mean – Quartiles – Deciles.

UNIT IV

Measures of Dispersion – objectives – Absolute and Relative Measures – Range – Quartile Deviation – Mean Deviation – Standard Deviation – Skewness – Kurtosis – Respective Merits and Demerits.

UNIT V

Correlation Analysis – Meaning - Uses – Types – Methods – Graphic – Scattered Diagrams – Algebraic Methods – Karl Pearson's Coefficient of Correlation – Merits and Demerits of Calculation – Concurrent Deviation method – Merits and Demerits.

UNIT VI

Regression Analysis – Difference between Correlation and Regression – Principles of Least Square Methods of Regression Analysis – Graphic, Algebraic – Regression Coefficients – Uses of Regression Analysis for Business Decision – Coefficient of Determination.

UNIT VII

Interpolation and Extrapolation and Time Series Analysis: Interpolation, Extrapolation – Meaning, uses, Measurements – Estimation Methods – Time series Analysis – Techniques of Measurements – Business Forecasting.

UNIT VIII

Probability – Concept of Probability – Types – Marginal, Joint Conditional Laws of Probability – Additional Theorem – Multiplication Theorem – Bayes Rule.

Note: Distribution of marks between problems and theory shall be 70% and 30%.

REFERENCES:

D.C.Sanchall & V.K. Kapoor, Statistics Theory, Methods & Applications,

S.P.Gupta, Statistical Methods,

Elhance D N, Fundamentals of Statistics

Gupta S P, Statistics for Commerce Students

UNIT - I

LESSON - 1

MEANING, DEFINITION & SCOPE OF STATISTICS

- ❑ INTRODUCTION
- ❑ DEFINITIONS
- ❑ DIVISION AND NATURE OF STATISTICS
- ❑ ORIGIN AND GROWTH
- ❑ SCOPE OF STATISTICS

INTRODUCTION

The word "Statistics" is a very commonly used word. But it conveys a variety of meanings to people. To some, statistics refer to a set of figures like profit made by a company, number of children born in a particular state in a year and their sex i.e., it is an information about a particular activity, or process nay it is production, population, national income, etc. that is expressed in numbers. Whenever numbers are collected and compiled regardless of what they represent, they become statistics. In other words the term statistics is considered synonymous with figures or data. Some people think of statistics as something used to reinforce a qualitative statement. Thus, when somebody says that the economic condition of Indian masses is improving in India since last 5 years, and gives the figures of rising per capita income of last five years, he is using statistics to support a qualitative statement. To many others, statistics is representation of a phenomenon with the help of figures, charts, diagrams, pictograms etc.

Many think statistics is a body of methods of obtaining and analysing numerical data in order to base decisions on them.

Thus the word statistics broadly refers to quantitative information or to a method of dealing with quantitative information. In the first reference, it is used as a plural noun i.e., statistics of production, import, export,

birth, death, etc. In this sense statistics are numerical description of the quantitative aspect of things. They take the form of counts or measurement. The use of the word statistics in this sense is always plural.

In the second reference, the word statistics is used to refer to the techniques and methods used in collection, analysis and interpretation of data. In this sense, the word is used in singular.

Thus statistics mainly is understood in two senses:

- 1) Statistical data
- 2) Statistical methods

Statistical Data

In this sense, the term statistics is applied to nearly any kind of factual information given in numbers. General statements often fail to acquaint the reader with the real situation and are vague or meaningless as they are not supported by figures. "Production of a company has increased". The statement is vague but when supplemented by figures i.e., production has increased from 20 crores to 50 crores in can be properly understood and analysed. Thus statements with statistical data are easier to follow, simpler to understand and analyse as compared to general statements. Such statements are clear, precise and meaningful as they are expressed quantitatively and the facts and figures speak for themselves.

However all facts numerically stated are not statistics. For example, a figure of Rs. 1,00,000 sales relating to a company is not statistics unless it is placed in relation to the sales of other companies or sales of the same company in different years. Thus figures relating to any phenomenon such as income, population export, money supply, input are called statistics.

Statistical Methods

When the term statistics is understood as statistical methods it refers to all those methods which are employed in collecting, presenting, analysing and interpreting numerical data. Thus it is nothing but the methods employed for drawing conclusion from quantitative data. It includes all the

necessary operations from the initial planning and assembling of data to the final presentation of conclusion. Statistical methods range from the most elementary descriptive device which may be understood by the common man to those complicated mathematical procedures which can be understood only by the expert theoreticians.

DEFINITION OF STATISTICS

The term statistics has been defined differently by different people. However the definition can be grouped into two categories depending on the two senses in which the term is generally used.

- 1) Definition of statistics as numerical data (Plural sense).
- 2) Definition of statistics as statistical method (singular sense).

Definition of Statistics as Numerical Data

In this section we illustrate the definitions given by two renowned statisticians who defined Statistics as numerical data. Webster defined statistics "as the classified facts representing the conditions of people in a state, especially those facts which can be stated in numbers or tables of numbers or in any tabular or classified arrangement".

This definition was proper when statistics was considered as a science of statecraft only. In olden days, statistics concerning population area under cultivation, wealth, land revenue etc. were collected by the government for administrative purpose or for nagging mass.

This definition is too narrow as it confines the scope of statistics to only such facts and figures which are related to the condition of the people in the state. In the modern world facts and figures are collected not only for studying the conditions of the people in a state but also for studying several other phenomena. Now-a-days statistics pervade all departments of enquiry. Thus this definition is not satisfactory.

Professor Horas Secrist gave the most comprehensive definition of statistics which include the scope of statistics. He defined statistics as follows:

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

This definition clearly points out certain characteristics which numeral data must possess in order that they may be called statistics.

Aggregate of Facts: Aggregate of facts are called statistics. Single and isolated figures are not called statistics for the simple reasons that such figures are unrelated and cannot be compared to derive any conclusion about the behaviour of the figure. The single age of 25 years or 50 years is not statistics but a series of ages of group of persons would be called statistics. A single figure relating to production, sales, birth, death, purchase, accident, similarly, can not be regarded statistics, though aggregated figures relating to production, sales, births, deaths, purchases, accidents etc. would be called statistics because they can be studied in relation to each other and are capable of comparison. It is possible to study them in relation to time, place and frequency of occurrence. Thus science of statistics is a science of aggregates and not of individuals.

Affected to a Marked Extent by Multiplicity of Causes: Statistics are aggregates of facts and figures which are affected by a variety of causes. The effect of various forces on a particular phenomenon cannot be studied separately because it is the result of the action and interaction of a number of forces. It is usually not possible to study the effect of any one of the factors separately as in the case of experimental methods. In statistical methods the effect of various factors affecting a particular phenomenon are generally studied in a combined form though attempts are also made to study the effects of different set of factors separately as well. Most of the statistics, however, are affected to a considerable degree by multiple causation. For example production of sugar is affected by seeds, fertilizer, rainfall, irrigation, climate quality of soil etc. The same is true of statistics of prices, imports, exports, sales, profits etc.

Numerically Expressed: Statistics are numerical statements of facts. Expressions of qualitative nature like good, bad, honesty, intelligences, young, old etc., do not constitute statistics. The general qualitative expressions like "India is a poor country", "Nehru was a great leader" "death rate in India has declined" are not called statistics. The statements must be expressed quantitatively if it is to be called statistics – for example, the density of population per square kilometre is 200 in India and 570 in Bangladesh.

Enumerated and Estimated According to Reasonable Standard of Accuracy: Facts and figures about any phenomenon can be derived in two ways, viz., by actual counting or measurement or estimate. Estimates cannot be as precise and accurate as actual counts and measurement. Where the scope of statistical enquiry is very wide or where the numbers are very large, enumeration is usually out of question, and in such a case figures can only be estimated. The degree of accuracy expected in such estimates depends to a large extent on the purpose for which statistics are collected. For example while measuring the distance between two places a few furlongs may be ignored; on the other hand in measuring the height of persons even 1/10th of a cm is material. Hence in many statistical studies mathematical accuracy cannot be attained. However it is important that reasonable standard of accuracy be attained; otherwise numbers may altogether be misleading.

Collected in a Systematic Manner: It is essential that statistics are collected in a systematic manner so that they may conform to reasonable standard of accuracy. Statistics employed for analysing of the behaviour of the variable, gives accurate results, provided they are collected in a systematic manner. Data collected in a haphazard manner would very likely lead to misleading conclusions.

Collected for a Predetermined Purpose: The purpose for collecting data must be decided in advance. It enables the investigator to distinguish between wanted and unwanted informations. If the purpose is not

determined in advance, then the investigator may collect irrelevant data which would do more harm than good. A general statement of purpose is not enough. For example if the objective is to collect price data, it would not serve any useful purpose unless one knows whether he wants to collect the data on wholesale prices, retail prices and what are the relevant commodities in view.

Placed in Relation to each Other: Statistical studies are undertaken mainly for the purpose of comparison. Statistical data may be compared either period-wise, class-wise, region-wise etc. Thus if the numerical facts are to be called statistics they should be comparable. For instance, population of India at a particular point of time may be compared with that of population of other countries. Valid comparisons can be made only when the data are homogeneous i.e., related to same phenomena or subjects. It is meaningless to compare the heights of men and animals.

In the absence of above characteristics, numerical data cannot be called statistics and hence "All statistics are numerical statements of facts but all numerical statements are not statistics".

Definitions of Statistics as Statistical Methods

The term statistics as statistical method too has been defined differently by different people. Some definitions are narrow and restrict the scope of statistics while others are comprehensive and widen the scope of statistics. A few important definitions are discussed below:

According to King, "The science of statistics is the method of judging collective natural or social phenomenon from the results obtained from the analysis or enumeration or collection of estimates". This definition limits the scope of statistics. The author himself admits the defect that it is not exhaustive but he feels that for the practical purpose, the definition is alright.

Al Bowley has given a series of definitions. At one place Bowley says "Statistics may be called the science of counting". At another place he is of the view that "Statistics may rightly be called the science of averages".

Both the definitions are defective as it has narrowed the scope of statistics to either counting or to averaging alone. There are, of course, some important methods but they do not cover the entire field of science of statistics. Yet another definition given by the same author characterizes statistics as "the science of measurement of social organism regarded as a whole in all its manifestations". Obviously, the definitions limit the application of statistical methods to the field of sociology only i.e., man and his activities. Bowley himself realised this when he remarked "statistics cannot be confined to any one science".

M.R. Spiegel has defined statistics as follows: "Statistics is concerned with scientific methods for collecting, organising, summarising, presenting and analysing data, as well as drawing valid conclusions and making reasonable decisions on the bases such analysis". This definition presents a wider scope of statistics and is comprehensive in nature. It takes all aspects of statistical enquiry into consideration thus bringing into light various phases of statistical investigations.

The definition given by Seligman is very short and simple, yet quite comprehensive. According to him, "Statistics is the science which deals with the methods of collecting, classifying, presenting, comparing and interpreting numerical data collected to throw some light on any sphere of enquiry".

Croseton and Cowden, two well known Statisticians have given a very simple and precise definition of statistics. In their view "Statistics may be defined as collection, presentation, analysis and interpretation of numerical data". This definition also points out various stages which are involved in a statistical investigation. This definition also is of wider scope and comprehensive nature.

From the analysis of above definitions, it is evident that the definitions of statistics given the Seligman Spiegel and Crouton and Cowden are scientific as they explain the various stages involved in statistical investigation. They are collection, organisation, presentation, analysis and interpretation of numerical data.

Stages of a Statistical Investigation

(1) Data Collection: Collection of data constitutes the first step in statistical investigation. The data may be collected either by primary method or by secondary method. The purpose of enquiry determines the method of data collection. However, utmost care must be exercised in collecting data because they form the foundation of statistical analysis and the conclusion drawn from the data can never be reliable if the data are faulty.

(2) Organisation: A Large mass of data collected from a survey frequently needs organisation though the data collected from published sources are generally in organised form. The collected data by primary method must be edited carefully so that omissions, inconsistency, irrelevant answers may be looked into, corrections and adjustments are made. The next step to editing the data is to classify the information. Classification is nothing but arranging the data based on some common characteristics. The last step in organisation is tabulation. The purpose of tabulation is to arrange the data in columns and rows so that there is absolute clarity in the data presented.

Presentation

After the collection and organisation of data, it must be presented in some suitable form for studying the salient features of the data. Data presented in an orderly manner facilitate statistical analysis. There are two methods which may be employed for the presentation of collected data.

1) Diagrams, and

2) Graphs

The tabulated data are transformed and presented in a compact manner either in the form of diagrams or graphs to make the raw data understandable.

Analysis

After collection, organisation and presentation, the next step is that of analysis. By analysis of data we mean, the study of the nature of the data. The nature of the data can be studied with the aid of several statistical tools which ranges from simple to complicated and sophisticated methods which can be handled by trained investigators or experts.

Interpretation

The final step in an investigation consists of interpreting the data which have been collected. Interpretation of data means the techniques of drawing conclusions from the critical study of the collected data. The interpretation of data is highly technical and difficult; therefore the task of drawing conclusion is done by specialised persons. If the data analysis are not properly interpreted the whole object of investigation may be defeated and fallacious conclusions may be drawn. Correct interpretation will lead to a valid conclusion of the study and this can aid one in taking suitable decision.

DIVISION AND NATURE OF STATISTICS

Statistics as a science can be divided into two main classes namely, inductive statistics and, descriptive statistics.

Inductive Statistics

Inductive statistics also called inferential statistics deals with all those devices, rules of procedure and general principles which are concerned with generalisation, prediction or estimation. Thus they include all the general principles and techniques which are commonly used in the collection, analysis and interpretation of data relating to any sphere of enquiry. Statistical methods are the tools in the hand of a statistical investigator. To infer about the nature of population, we select a random sample from the population and on the basis of the finding from a random

sample, we infer things about a population. This inferring about population on the basis of sample is known as statistical inference. The statistical tools or devices for achieving the desired end is the statistical inference. Since a method is always a means to an end, its accuracy and precision depends upon the object which is desired to be achieved and this in turn is considerably affected by the peculiar feature of the problem to which it is related. This is the reason why different statistical methods are used in different types of enquiries and no uniform standard of accuracy is desired to be achieved on different types of investigation.

Descriptive Statistics

Descriptive statistics deals with those statistical methods which are concerned with describing the characteristics of data. The type of statistics describe the phenomenon or condition that exists for describing the nature of the variable or variables under study. Statistical methods are employed for collection, classification and tabulation of data. For example, the net profit earned by a company is given below:

Year	Profit in Lakhs
1990	150
1991	170
1992	200
1993	250
1994	300

The above table shows the net profit earned by a company from 1990 to 1994. It shows that the profit of the company has increased and have got doubled in 5 years period. Here the behaviour of factors responsible for the increase or doubling of the profit are not analysed. Thus descriptive statistics deals with the activities such as classification of data, presentation of data, computation of mean, median, mode, standard deviation or an index number.

Statistics and Astronomy

The evolution of statistics dates back to the first collection of data by astronomers for the study of movement of stars and planets. The method of least square was first developed by an astronomer. Astronomers generally take large number of measurements and in most cases there is some difference between these several observations. To make better estimates and to bring more precision in the measurement, the application of statistical methods are made by assigning limits within which the true value of the phenomenon is expected to lie.

Statistics and Meteorology

Statistics is related to meteorology. In meteorology records are made of temperature, humidity of air and barometrical pressures etc. For the purpose of comparison and forecasting it becomes necessary to average these figures and to study their trend and fluctuations. A study of the significance of these deviations has also to be made for various purposes. All this cannot be done without the use of statistical methods. We thus find that the science of statistics helps meteorology in a large number of ways.

Is statistic a science or an art, is a question very often asked but not adequately answered. Science refers to a systematised body of knowledge. It studies cause and effect relationship and attempts to make generalisations in the form of principle or laws of the subject concerned. It describes objectively and avoids vague judgments. Art on the other hand, refers to the skill of collecting and handling of certain conclusion. The degree of precision in scientific laws is very high whereas the degree in the inference drawn with the skill of handling data are not as high as it is in the scientific laws.

Judged on the basis of the above definitions of the words science and art, statistics could lay claim on both. However statistics is not a science in the same sense of which the pure sciences like physics, chemistry, zoology and astronomy are. This is because statistical phenomena are

generally affected by multiplicity of causes which cannot always be measured accurately. The reason for this is that statistics deals with such variables whose individual effects cannot be studied in isolation as is possible in natural or physical science which are experimental in nature. In other words the science of statistics by its very nature is less precise than natural science. It is science only in a limited sense viz., as a specialised branch of knowledge. More appropriately statistics may be regarded as scientific methods than a science because it is a tool which can be used for scientific studies. Statistical methods can be and are applied in all sciences, natural or social and as such they should be regarded as indispensable tools for the study of any qualitative phenomenon.

Statistics is an art far as skill of handling facts and figures for the purpose of analysis, interpretation and policy formulation are concerned. Art signifies action. Looking from this angle statistics is regarded an art as it involves application of given methods to obtain facts, derive results and finally use them for appropriate action.

ORIGIN AND GROWTH OF STATISTICS

The word statistics seems to have been derived from the Latin word "status" or Italian word "statista" or German word "statistic" each of which means political state. In fact the origin of statistics was due to administrative requirement of the state. For the purpose of war or finance administration states required the collection and analysis of data relating to population and material wealth of the country. Even though collection of data for other purpose was not ruled out, the earliest form of statistical data mainly related to population and property. Perhaps one of the earliest census of population and wealth was held in Egypt as early as 3050 B.C. for erection of pyramids. In India about 2000 years ago we had an efficient system of collecting administrative statistics. Instances of efficient system of collecting vital statistics and of registration of births and deaths during the Maurya regime, detailed accounts of Statistical survey conducted during the reign of Emperor Akbar are there in the history of India. The history

of other countries of the world also clearly indicate that in ancient times statistics was regarded as a matter connected with the activities of the state and that is why it was known as a science of statistics and was the by-product of administrative activity of the state.

The science of statistics is said to have originated from two main sources.

1. Government records.

2. Mathematics.

Government Records: All cultures with a recorded history had recorded statistics and the recording as far as is known, was done by the agents of government for governmental purpose. Thus the government records are regarded as the earliest foundation of statistics. Since statistics were collected for governmental purpose. Statistics was then described as the "Science of Kings" or "Science of statecrafts". Many prominent people such as Captain John Graunt (1620-1674) William Petty (1623-1687) and Henry contributed a great deal to the development of statistics. Captain John Graunt was regarded as the Father of Vital statistics. William Petty was the author of "Essay on Political Arithmetic" (1690). He regarded statistics as Political Arithmetic.

Mathematics: Statistics is said to be a branch of applied mathematics. The present body of statistical methods, particularly those concerned with drawing of inference about population from a sample is based on the mathematical theory of probability. The theory emerged in seventeenth century as a result of gambling. Demoivre, Bernoulli, Laplace Gauss discovered and developed the theory of probability while estimating the chance of winning or losing in gamble.

The theory of probability indicates what will happen if very large number of trials are undertaken even though the actual outcome in any single trial is unknown. Abraham De Moivre (1667-1754) discovered normal curve which forms an important part of modern statistical theory. Laplace (1749-1827) and Gauss (1777-1855) independently aimed at the same

result as De Moivre. The principle of constancy of great numbers which is the basis of sampling was discovered by Quetlet (1796-1879). During the 19th century much of the development in statistical techniques has taken place. Sir Francis Galton (1822-1911) developed the concept of regression, Karl Pearson (1857-1936) developed the chi-square goodness of fit test. Sir Rosalind Fisher (1890-1962) made major contributions in the field of experimental designs. The contribution of such great personalities are immense to statistics because of which statistics has reached the present high as a body of knowledge. Though the relationship of Statistics and Mathematics is very old, it is only during the last 100 years or so that the two sciences have come very close to each other.

Growth of Statistics

In recent years, the domain of statistical methods has considerably widened and today there is hardly any science which does not make use of statistical methods. The science of statistics is now associated with all other sciences in some form or other. There is hardly any branch of science today that does not make use of statistics. Statistics is now regarded as one of the most important tools for taking decision in the midst of uncertainty. Of late there has been a remarkable and sustained growth in the use of statistics. This is because business, government and science - three fields in which application of statistics are so numerous and diverse - are growing in volume and complexity. It is also because of the technological revolution which has taken place in data handling, affecting especially computing and tabulating equipment and a scientific revolution in statistical theories and techniques.

Cause of recent growth of statistics: The tremendous growth in the use of statistics can be attributed mainly to two factors viz - increased demand of statistics and decreasing cost of statistics.

SCOPE OF STATISTICS

The scope of statistics is so vast and ever expanding that not only is it difficult to define it but also unwise to do so. All types of quantitative

analysis relating to any department of inquiry are included in it. This is so because statistics helps in drawing conclusion from the facts affected by multiplicity of causes in any department of inquiry. Whenever a study of quantitative phenomenon is required, statistics has entered as a tool in all branches of knowledge. In fact there is hardly any field whether it be business, economics, biology, physics, astronomy, meteorology, chemistry, medicine, sociology, psychology or education where use of statistics and statistical method is not made use of. The application of statistics are so numerous that it is often remarked "Statistics is what statistics with other subjects form which the scope of statistics would become more evident".

Statistics and the State

Since ancient times the ruling kings and chiefs have relied heavily on statistics in framing suitable military and fiscal policy. Most of the statistics such as that of crime, military strength, population, taxes etc. that were collected by them were a byproduct of administrative activity. With the increase in the function of welfare state, the latter collects statistics on several problems for the purpose of framing policies and taking corrective actions wherever necessary. Whether it be transport problem, or family planning problem, statistics is of immense help. All ministries and departments of government whether they be Finance, Transport, Defence, Railways, Food, Commerce, Post and Telegraph or Agriculture, depend heavily on factual data for their efficient functioning. Statistics are so significant to the state that the government in most countries is the highest collector and user of statistical data. Such data is of immense help to many institutions and research scholars, who further process it and arrive at useful conclusions which help in decision making.

Statistics and Business

With the growing size and ever increasing competition, the problems of business enterprises are becoming more complex. In order to be successful in decision making under the atmosphere of uncertainty concerning future operations, such as production, investment inventory and marketing of products, a businessman has to use statistical methods

to analyse and synthesize data relating to business. A businessman who has to deal in an atmosphere of uncertainty can no longer adopt the method of trial and error in taking decision. In recent years it has become increasingly evident that statistics and statistical methods have provided the businessman with one of the most valuable tools for decision making.

Business activities can be broadly grouped under the following heads:

- 1) Production
- 2) Sale
- 3) Purchase
- 4) Finance
- 5) Personnel
- 6) Accounting
- 7) Market and Product research and
- 8) Quality control

With the help of statistical methods in respect of each of the above areas, abundant quantitative informations can be obtained which can be of immense use in formulating suitable policies, the information might be records that are kept on letters and other books. An industrialist resorts to the statistical methods before setting up an industrial unit as they help him in taking decisions regarding the product and the products to be manufactured, deciding the location and size of the plant and conducting a survey of consumers with a view to make a study of their taste and preferences. For planning these proposed activities relevant informations are collected, analysed and applied. Also statistics help industrialists in instituting a control over quality of the products so as to maintain the standard of the quality. Statistics helps business manager in various fields. All business planning is based on forecast of sales. A forecast of sale is made by studying the time series of present and prospective conditions of the concern and related business, Statistics also is helpful in conducting market research by the businessman. Sampling methods are used by the

marketing researchers in making surveys of consumers' preference over certain brands of competitive merchandize. Statistical tables and charts are frequently used by the sales manager to present numerical facts of sales. Similarly in fixation of price of commodities, statistics is of great help. Thus it cannot be denied that modern business is organized around system of statistical analysis and control.

However it should be remembered that though statistical methods are extremely useful in taking decisions, they are not perfect substitutes of commonsense. A practitioner of business statistics must therefore combine the knowledge of business environment in which he operates and its technological characteristics with a heavy dose of commonsense and ability to interpret statistical methods to non statisticians.

Statistics and Economics

Economics is concerned with the production and distribution of wealth as well as with the complex institutional setup connected with the consumption savings and investment of income. Statistical data and statistical methods provide valuable assistance for the study and solution of economic problems as in formulation of economic policies. Most of the economic problems can be expressed numerically which help the economists in collecting data and applying statistical methods in formulating policies and arrive at certain decisions. In fact statistical data and statistical methods are the tools and appliances of an economist's laboratory. This led Prof. Marshall to say "Statistics are the straw out of which I, like every other economist, have to make bricks".

In the field of production the questions like what to produce, how to produce, for whom to produce, cannot be answered in the absence of statistical data. Statistics of production help in adjusting the supply to demand. Statistics of production of an area or country can be compared with another area or country. Census of production help in understanding the economy and taking corrective measures wherever necessary.

Statistics of consumption enable us to find out the way in which the people in different strata of society spend their income. Such statistics are

very helpful in knowing the standard of living and taxable capacity of people.

In the field of exchange, a systematic study of markets, law of prices based on supply and demand, cost of production, monopoly, banking, credit and credit investments etc. cannot be made without statistics. What shall be the price of a particular commodity if its supply increases or decreases, what price should the monopolist charge to reap the maximum profit, what should be the rate of interest that a bank should charge during a particular season – these are the questions which can be best answered with the help of statistics. In fact statistics are the very foundation stone of the theory of exchange.

In the field of distribution too, statistics has a vital role to play. How the national income is to be calculated and how it is to be distributed these are the questions which cannot be answered without statistics. In reducing disparity in the distribution of income and wealth, statistics are of immense help. Similarly in solving problems of rising price, growing population, unemployment, poverty etc. one has to rely heavily on statistics.

Thus it is not exaggeration to say that the science of economics is becoming statistical in its method. The study of statistics has become indispensable for description, comparison and correlation of economic data. This is so because these days economists are resorting to inductive method rather than to the deductive reasoning. So they need support either from large mass of data or from statistical method to pronounce certain economic laws. Not only that, statistics also can be useful for illustrating an economic phenomenon for proving the validity of certain economic theories.

Statistics and Econometrics

Econometrics is a subject, which consists of the application of statistical methods to the theoretical economic methods. The derivation of demand functions, production functions, cost function and the consumption function present many difficult problems in the analysis of which statistical tools are of immense use. Statistical methods of collecting basic data, statistical methodology to indicate the reliability of the data

and the significance attached to it for economic studies are, to name a few of the applications of statistical methods in basic economics. The aim of econometrics is to make economics more realistic and practical. Mathematical approach to economic theories make them more logical and similarly statistics gives a quantitative conclusion about the validity of purely theoretical concept, emerging with happy union between statistics, mathematics and economics.

Statistics and Biology and Natural Sciences

The development of biological theories has been found to be closely associated with statistical methods. Prof. Karl Pearson in his "Grammar of Sciences" says that the whole doctrine of heredity rests on statistical basis. The contention that tall fathers have generally tall sons can be proved only by the use of statistical data and statistical methods. The differences observed in various generations in different zoological species can be measured and studied only with the help of statistical technique. Thus we see that statistical methods help in formation of theory of development of human and animal life.

Statistics and Physical Sciences

Other than astronomy and meteorology, other physical sciences also use statistics to some extent. The physical sciences like geology, physics along with astronomy are among the fields in which statistical methods were first developed and applied. But until recently these sciences have not shared the recent developments of statistics to the extent it is being done in biological and social sciences. Currently however, the physical sciences especially chemistry, engineering, geology, certain branches of physics etc. seem to be making increased use of statistics. Conversion of heterogeneous data into homogeneous one, application of investigation and analysis of methods of interpretation and drawing inferences are some of the methods where physical sciences depend to a great extent on the science of statistics.

Statistical techniques have proved to be extremely useful in the study of all natural sciences like biology, medicine, meteorology, zoology, botany etc. For example in diagnosing the correct disease, the doctor has to rely heavily on factual data like temperature of body, pulse rate, blood pressure, similarly judging the efficacy of a particular drug for curing certain disease. Experiments have to be conducted and success and failure would depend on the number of people who are cured after using the drug. In botany one has to rely heavily on statistics in conducting experiments about the plants, effect on temperature, types of soil etc. In fact it is difficult to find any scientific activity where statistical data and statistical methods are not used.

Statistics and Research

Statistics is indispensable in research work. In fact there is hardly any research work today that one can find complete without statistical data and statistical methods. Also it is impossible to understand the meaning and implications of most of the research finding in various disciplines of knowledge without having acquaintance with the subject of statistics.

Statistics and other Fields

We have discussed above, the application of statistics in some important fields. Besides these, statistics are important and are applicable whenever quantitative studies are to be made. In political science, education auditing etc. also statistical methods are widely used. The list of statistical applications is not exhaustive. It only shows the diversity of application of the statistical methods. The statistics are useful even to bankers, brokers, insurance companies, social workers, trade unions, trade associations, and politicians. In fact, the application of statistics are so numerous that statistics today has risen from the science of statecraft to the science of universal applicability. The usefulness of statistics, however to a great extent depends on the ability of the person to use his imagination in applying a particular tool to his own particular situation.

LESSON - 2

**FUNCTION, IMPORTANCE, MISUSES AND LIMITATIONS
OF STATISTICS**

- ▣ FUNCTIONS OF STATISTICS
- ▣ IMPORTANCE OF STATISTICS
- ▣ MISUSES OF STATISTICS
- ▣ LIMITATIONS OF STATISTICS

Statistics plays an important part in every walk of life and has proved extremely useful in almost every line of scientific, economic and business enquiries. The human mind is not capable of keeping large number of distinct informations in mind. To bring the information into a condensed form statistics is required. Also without simplifying data in some fashion, the comparison between two groups of data cannot be done. Thus the use of statistical science which supplies various methods condensing, comparing and analysing data, is made for simplifying the unwieldy masses of facts. Without a statistical study most of our ideas are likely to be vague and indefinite. The use of statistics not only reduces the figures and gives a clear-cut form to hazy conception, it helps us to set objects in their proper perspective and relationship and also to analyse the law governing their movements and changes. Thus statistics helps the human mind in grasping the true significance of complex data. Figures are useful for making comparisons of similar objects with reference to time or place. The usefulness of figures also are for drawing inferences and testing hypothesis. That is the importance of statistics.

Despite the importance of statistics in various fields and important functions which it performs, the use of statistics is limited to numerical aspects only. In addition to it statistics as a tool of analysis fails to study the qualitative aspects. For example, how great was Mahatma Gandhi, cannot be stated statistically. The science of statistics may also be misused by certain interested and prejudiced persons. In order to make proper use

of statistics, it is necessary to keep in mind its limitations; otherwise there is every likelihood of drawing wrong conclusions.

In this lesson, we discuss functions, importance, misuses and limitations of statistics and statistical methods.

FUNCTIONS OF STATISTICS

At this stage, it is worthwhile to examine and identify the functions of statistics and the role played by statistics with regard to various aspects of quantitative data in various fields.

The following are important functions of science of statistics:

- 1) It presents fact in a definite form.
- 2) It simplifies the mass of figures or complex data.
- 3) It facilitates comparison.
- 4) It establishes relationship among variables.
- 5) It provides techniques for drawing inferences.
- 6) It helps in formulating and testing hypothesis.
- 7) It helps in prediction.
- 8) It helps in formulation of suitable policies
- 9) It adds to the knowledge.

1. Definiteness: Numerical expressions are usually more convincing than general expression. Thus one of the most important functions of statistics is to present general statements in a precise and definite form. Statistics presents facts in a precise and definite form and helps proper comprehension of what is stated. Consider for example a statement "the total exports of the country in 1997 is expected to be greater than last year i.e., 1996". The reader will not have a clear-cut idea of the situation from the statement. He would surely like to know what is the extent of increase the writer has in mind. On the other hand, if we quantify the statement, "the export in 1997 is expected to increase from 53978 crores in 1996 to 63813 crores in 1997" it conveys a definite information. Similarly the statements like "the inflation rate is declining", "The population of India

is growing at a very fast rate", "There is a lot of unemployment in India" etc. hardly convey any worthwhile informations as they do not specify the numerical dimensions involved.

2. Condensation: Not only statistics presents facts in a definite form but it also helps in condensing mass of data into few significant figures. The main function of statistics is to simplify the huge mass of data as it is not possible to make any analysis from the unorganised data. In a way statistics does a great service by reducing the raw data to totals or averages. Statistical methods present a meaningful overall information from the mass of data which enables us to know the salient features of the data. Thus it is impossible for us to form a precise idea about the wages of 800 workers in a firm from the pay roll of 800 workers. If we find out the average wage of a worker in the firm, it becomes easy to remember and it may also be used for comparison with the wages of other firms. Similarly the raw data may be translated into graphs, charts or frequency distribution to make it meaningful.

3. Comparison: Statistics facilitates comparison between two or more variables or objects or series relating to different times and places. Unconnected figures have no meaning unless they are being placed in relation to the other figures of the same type. Suppose we are told that BSE Index in Feb. 4 1997 was 3638.13. This single figure does not show whether the share prices were high or low; on the contrary if we are supplied with the BSE Index of any other period or day then we can compare the two indexes and opinions about the share prices can be given. To quote another example, if we are supplied with data relating to percapita income of different countries then it becomes very easy to compare these figures and draw conclusions about the standard of living of people in those countries. For making comparison, statistics provides measures such as rate, average, percentages, graphs, diagrams, etc. Thus by furnishing suitable device for comparison of data, statistics enables a better appreciation of the significance of a series of figures.

4. Relationship: Another function of statistics is to establish relationship between two or more variables relating to different fields. For example, in the field of agriculture, statistical methods are employed to study the effect of fertilizer or irrigation upon the crop yield. This relationship is very useful to farmers because they can administer the recommended doze of fertilizer to get desired amount of crop yield. For studying relationships statistics provides measures such as correlation, regression, coefficient of associations etc.

5. Inference: Statistics provides methods which are employed for drawing inferences about the nature of parent population. In most of the enquiries the characteristics of the population cannot be studied exhaustively. For example if one wants to study the mineral content of water of a particular city it is not feasible to study the whole water available at each and every point in that place. Thus it is not possible to adopt census method for studying the characteristics of the population, either due to lack of money or time or trained investigators. In such case sampling method is adopted and information is obtained about a small number of items. The results of sample findings are then generalized to reveal the main characteristics of the parent population. Therefore statistical inference has become the most important branch of statistics. It saves the investigator from the botheration of collecting huge mass of data and results in saving time, money and human energy.

6. Prediction: Forecasting of future events has become an important function of statistics. Plans and policies of organisations are formulated well in advance of the time of their implementation. A knowledge of future trends is very useful in framing suitable policies and plans. Statistical methods are employed for forecasting demand, production, population, exports etc. For example, if Videocon Ltd. has to decide how many T.Vs it should produce in 1997, it must know the expected demand for the year. It may use subjective judgment and take a guess. However a better method would be to analyse the sales data of past years or arrange a statistical survey of the market to obtain necessary data for estimating the sales

volume of next year; accordingly it can decide about the number of TVs to be produced. Likewise Government of India always makes estimates of domestic production of food grains and decides the quantum of imports, if necessary. Tools like regression, time series analysis, growth curves etc. are used for making forecasts.

7. Testing Hypothesis: Statistical methods are very useful in formulating and testing various types of hypothesis and to develop new theories. In the field of social sciences hypothesis formulation and their testing is very important. For example by the use of statistical methods we can test the hypothesis whether a rise in railway fare has affected the passenger traffic, whether Indian consumers are brand loyal, whether medicine A is effective in curing a disease or B, whether a particular coin is fair or not, whether credit squeeze is effective in checking price increase etc. In fact in most of the researches in social sciences some hypotheses are formulated and by collecting, analysing and interpreting empirical data the validity of the hypothesis is tested. The hypothesis will be either accepted or rejected. The acceptance of the hypothesis leads to the formulation of new theories.

8. Formulation of Policies: Statistics provide the basic material for framing suitable policies. For example it may be necessary to decide how much sugar should be imported in 1995. The decision would depend on the expected sugarcane production in the country and the likely demand for sugar in 1995. In the absence of the informations regarding the estimated domestic output and demand, the decision on imports cannot be made with reasonable accuracy.

9. Addition of Knowledge: The science of statistics enlarges human experience and knowledge by making it easier for man to understand, describe and measure the effects of his action on the action of others. Many fields of knowledge would have ever remained unknown to mankind but for the efficient and refined techniques and sound methodology provided by the science of statistics. It has provided such a master key to the mankind that we can use it any where and can study any problem in

its right perspective and on the right line. According the Prof. Bowley "The proper function of statistics is indeed to enlarge individual experience". Had there been no statistics it would have been impossible for us to digest the qualitative statements like "I have earned more this year compared to last year", "The inflation rate is falling" etc. Such statements are too vague to make any sense. For meaningful comparison the statements must be expressed numerically such as "I have earned Rs.10,000 this year compared to Rs.5,000 last year" Thus statistics broaden our horizon.

IMPORTANCE OF STATISTICS

Importance of statistics can hardly be overemphasised. The application of statistical techniques is widespread and influence of statistics on our lives and habits is great. Our knowledge would have remained limited and most of our problems would have remained unsolved in the absence of statistics and statistical methods. Now we discuss the major gains of the study of statistics under the following heads:

1. Importance for common man
2. Importance to government
3. Importance to economics
4. Importance to planning
5. Importance in business and commerce
6. Importance in research
7. Importance to bankers and insurance companies
8. Importance in management or administration
9. Importance in Public utility concerns
10. Universal applicability.

1. Common Man: The fact that in the modern world statistical methods are of universal applicability, is in itself enough to show how important the science of statistics is. As a matter of fact there may be millions of people who would not have heard a word about statistics and yet make a profuse use of statistical methods in their day-to-day decisions. When we

use a common proverb "as you sow, so you reap" we indirectly hint that there is a positive correlation between one's action and achievement. When a person wishes to purchase a refrigerator or a TV he goes through the price lists of various companies and makes to arrive at a decision. What he really aims at is to have an idea about the average level of price and the range between which the price varies and makes comparisons of prices, though he may not know a word about statistics or statistical methods. When a farmer expects a particular amount of yield from his land he has in fact the idea of correlation of the factors affecting the crop yield and the regression line of crop yield on rainfall or fertilizer etc.

Examples are numerous to show that human behaviour and statistical methods have a lot in common. In fact statistical methods are so closely connected with human action and behaviour that practically all human activities can be explained by statistical methods. This shows how important statistics is for the mankind.

2. Government: Statistics is indispensable for government for the proper implementation of its policies. In fact no government can assess the working of its economic system in the absence of statistics relating to agricultural production, industrial production, export, import etc. To know the manpower and the occupational structure of population, government needs data on population. Government needs data of agricultural production, industrial production, consumptions, savings, unemployment, literacy etc. because they affect the policies of government in the sphere of taxation, money to be spent on public works etc. A state in the modern setup besides being an administrative body is a big commercial concern also. It carries on business of various kinds and has monopoly in many cases. It needs statistics, for carrying on these works on sales, purchases, personnel, prices etc. A Government has to maintain a large army and sophisticated military equipments for the safety of the country from external aggression. Therefore a government needs the records of army personnel and weapons.

In fact the government is always the most important single unit which not only collects the larger amount of statistics but also needs statistics in a very extensive scale. Probably this is the reason why official statistics occupy a very important place in the statistical literature of any country.

3. Economics: In the field of economics, it is almost impossible to find a problem which does not require an extensive use of statistical data. Important phenomena in all branches of economics can be described, compared and correlated with the help of the statistics. Thus statistics occupies a predominant place in the field of economics. In economics the study of consumption, production, distribution, exchange and public finance are incomplete without the help of statistics. Statistics of consumption tell us of the relative strength of the desire of a certain section of the people and its variation from time to time. By statistical analysis we can study the manner in which people spend their income over various items of expenditure, namely food, clothing, shelter etc. Statistics of production describe the wealth of a nation and compare it year after year showing thereby the effect of changing economic policies and other factors of production. Exchange statistics throw light on the commercial development of a nation. The volume of business done by the country, the amount of money in circulation etc. are known from the various classes of people is disclosed by distribution statistics. Thus we see, for all kinds of economic problems statistical approach is essential and statistical analysis is useful.

4. Planning: Modern age is an age of planning. The days of laissez fair are gone and state's role in all aspects of life has become universal in character. As planning of some sort has become popular in all walks of life so economic planning has become popular in communist as well as capitalist countries of the world. The main objective of economic planning is to attain maximum production with available resources. To achieve the desired objective in implementing the plan, proper assessment of material and human resource is essential. For this the planning authority of the country needs statistics of population, agricultural as well as industrial production

prices, national income, employment before the beginning and the end of planning period to assess the success of plans. The planning authority always keeps an eye on the economic trends by collecting data so that if need arises adjustments can be made in the set targets. Thus not only plans of economic developments are constructed on the basis of statistical data but the success that a plan achieves is also measured best by the use of statistical apparatus. Thus we infer statistics are indispensable in economic planning and economic planning, is inconceivable without statistics.

5. Business and Commerce: The importance of statistics in business and commerce has been expanding rapidly because the methods of production has changed from manufacturing to order, to anticipatory mass production. A producer has to make a number of decisions such as where to produce, how to produce, where to sell and at what price to sell. As a consequence the success or failure in business to a large extent depends upon the accuracy of forecasting and, statistical methods serve the purpose of forecasting and, modern devices have made business forecasting more definite and precise. Economic barometers are the gift of statistical methods and businessmen all over the world make extensive use of them. A producer estimates probable demand of his good, analyses the effect of trade cycle and seasonal variations, changes in habits of customers in the demand of his products and after taking all test factors into consideration finally takes decision about the quantum of production. A businessman who ignores the effects of booms and depressions can never succeed. All types of businessmen have to make use of statistics in one form or other in order to succeed. Various branches of commerce utilise the services of statistics in different forms. Promoters of new business make extensive use of statistical data to arrive at conclusions which are vital from the point of view of starting a new concern. In the absence of statistical data business failure ought to be more than success. Again, cost accounting is entirely statistical in out-look and it is with the help this technique that producers are in a position to decide about the prices of various commodities. We

thus find that for a balanced growth of business and commerce, science of statistics is of utmost importance.

6. Research: The methodology of statistics is constantly used in research conducted in various fields of natural and social sciences. Had there been no statistics or statistical methods it would have been impossible to reach at any conclusions about the result of the research. In agricultural research experiments about crop yields with different types of fertilizers and different types of soils are very often designed and analysed according to statistical methods. In medical research for testing the efficacy of new medicines or methods of treatment statistical methods and analysis are used. In the field of trade and commerce statisticians carry on different types of researches. They try to find out the cause of variations of different products from their standard quality. The technique of quality control is entirely statistical in nature. Market researches are carried on by making extensive use of statistical methods. In fact for a research worker in any field which is concerned with numerical results a study of statistical methods is a basic necessity.

7. Bankers and Insurance Companies: Bankers, stock exchange, broker, investor and insurance companies all make extensive use of statistical data. A banker needs information not only about the bank's operations but also regarding general economic conditions, business cycle to forecast a probable boom or depression and has to study in detail the seasonal variations on demand for call money from its clients. It is after a study of these factors that banker decides about the amount of reserves that should be kept. All such decisions are made on the available, statistical data i.e. money in circulation, borrowed money, etc. In India, for example, there is always heavy rush for borrowing during the months of Oct/Nov and April/May for purchase of agricultural produce. If banks fail to meet the money requirements of public, people will lose confidence in the banking system. In the past many banks failed because they did not study or analyse general business conditions. They must have full information about the amount deposited under various heads, the number of depositors, amount

advanced, interest charged, interest paid on deposits etc. Thus statistics are essential for the smooth running of banking system. Insurance companies cannot carry on their business in the absence of statistical data relating to life tables and premium rate. In fact insurance has been one of the pioneer branches of commerce and business which has been making use of statistics from the very beginning. Generally insurance companies carry on their business in terms of probability. On the basis of past experience and statistics relating to death of persons at different ages for many years, they determine probability of dying of a newly insured person by accident. Had there been no knowledge of probability, they would have suffered losses by insuring such persons who are prone to die of accident. Thus the science of statistics is indispensable for Banking and Insurance companies.

8. Management and Administration: Statistics provide the administration information for planning, checking effectiveness of regulations that have been issued. Statistics also provide the administration a powerful instrument of control of its own activity. One element common to all business managers is making decision in face of uncertainties and the essence of modern statistics lies in development of general principles for dealing wisely with uncertainties. Modern statistical tools of collection, classification, tabulation, analysis and interpretation of data have been found to be an important aid in making wise decisions in various levels of managerial functions. There are many areas where the use of statistical methods are made to arrive at correct decisions like production programming, sales, quality and inventory control etc.

The service of production programming both in the short as well as long periods depend to a great extent on quality of sales forecast and projections. A company's regional offices make estimates of their respective areas taking into account the seasonal variations. These forecasts may further be recast in the light of statistically estimated variations in aggregate market variations under the economic programming at national level.

Basing on the sales variation due to seasonal effect, the production schedule may be disturbed. Basing on the statistical methods, upper and lower zone of sales fluctuations can be set to statistic production. If this fluctuation of sales goes outside the zone, set changes in production schedules are made.

Market research on consumer preference studies, trade channel studies and leadership surveys are some methods of sales control which make extensive use of statistical tools.

Statistical methods also come to the aid of quality control. Statistical methods can be used to avoid cumbersome process of quality checking. By proper sampling and testing the sample for the required quality, the purpose of quality control is served with least botheration. Proper sampling and their testing take the help of statistical methods. Quality control in inventory is not only facilitated but also made more accurate with the aid of statistics. When the size of inventory is very large it is not possible to meticulously inspect each and every item. The problem is tackled by random sampling. Some items are selected on this basis and subjected to close quality inspection and if these are found according to specification the whole lot is accepted.

9. Public Utility Concerns: Statistics is extensively employed in public utility concerns such as railways, electric supply companies, water supply, post offices, schools, hospitals etc. The public utility concerns are started by the government for providing basic amenities to the people for which government needs statistics about size of population, its geographical divisions. In the absence of statistics regarding the geographical distribution of people it would have been difficult to plan for a school or a post office or a hospital.

10. Universal Use: There is hardly any branch of knowledge where statistics is not used. Statistics is used in any area, be it geology, psychology, sociology, political science, library science in which decisions are made on the basis of quantitative information. We thus find that

statistical methods are of wide, almost universal applicability. Government needs them, economists need them, businessmen need them. In fact all types of persons – astrologer, astronomer, biologist, botanist, meteorologist – make use of statistics and statistical methods. Statistics when used effectively become so intertwined in the whole fabric of the subject to which it is applied that it becomes an integral part of the subject. The universality of statistics is enough to indicate its importance, utility and indispensability.

MISUSES OF STATISTICS

Despite of its importance and usefulness the science of statistics is looked upon with suspicious eyes and is quite often condemned as a tissue of falsehood. The science of statistics is a very useful servant but only of great value to those who understand its proper use. Thus it is believed, the statistics is like clay of which one can make a god or a devil as he pleases. It is not out of place to give some quotations regarding misuse of statistics. According to Mark Twain there are three kinds of lies namely lies, damned lies and statistics wicked in order of their naming.

Another criticism levelled against statistics are of type "People lean on statistics like a drunken men on lamp post for support rather than illumination". "Statistics are lie of first order". Such comments and criticisms refer to misuses of statistics. Thus the persons who use statistics and apply statistical methods should be expert in handling them objectively. If statistics and statistical methods are used properly they can help arriving correct decision. But if they are misused, fallacious conclusions, may be arrived at. It is like a knife which is used for cutting vegetables but it can cut somebodies nose also. It is not the knife that is to be blame but the person who uses it for cutting nose. the fault lies with the way in which statistics and statistical methods are used and the person or persons using it.

The following are the main misuses of statistics:

1. Defective data.

2. Unrepresentative sample
3. Insufficient or incomplete data
4. Non comparable data
5. Inadequate sample
6. Misinterpretation of findings
7. Misleading results
8. False assumptions
9. Undefined units

1. Defective Data: The behaviour of a variable is studied on the basis of collected data. In case the data are collected in unscientific method it is difficult to make proper analysis of the nature of the problems under study. The conclusion drawn from the defective data are always misleading which defeats the very purpose of the enquiry. Thus it is essential to judge the data properly before its use.

2. unrepresentative Sample: In statistical analysis, mostly the conclusions on the characteristics of population are drawn from the study of the sample taken from the population. However if the sample drawn is not the true representative of population from which it comes, then conclusions drawn about the population would be wrong. Such findings may do more harm than good. Unrepresentative sample may result, when scientific method of drawing sample is not adapted like a method in which all the items don't have chances of being selected or the data have been collected by such a method that certain classes fail to respond.

3. Insufficient Data: It is always dangerous to draw conclusions from insufficient data. The data must be sufficient to study the salient features of the variable under study. For example it is not fair to announce the effectiveness of a new medicine after testing it on a few patients say two or three. For reliable results it should be tried with large number of patients. Then only the doctors should place high degree of confidence on the effectiveness of a medicine.

4. Non Comparable Data: The main function of statistics is to facilitate comparison between two or more objects. No fair comparisons can be made and the conclusion drawn would be wrong in case the data are non comparable. For example note the following statement: The profit of Firm A are Rs. 40,000 in 1995-96 and that of B Rs. 6,70,000. On the basis of this information we would form the opinion that firm B is decidedly better than firm A. However when we examine the amount of capital employed we might reach a different conclusion. Hence while making use of statistics one should not only avoid outright falsehood but also be alert to detect possible distortion of truth.

5. Inadequate Sample: Sampling techniques have become very popular with research workers for making inference about the nature of population on the basis of the findings of the sample. The conclusion drawn about the characteristics of the population from a very small sample unscientifically selected would be unreliable. If one wants to study the performance of public sector units in India and takes one public sector unit as a sample, the generalization about the public sector units may prove to be utterly fallacious. Thus conclusion drawn from a very small sample about a very high population may be entirely misleading.

6. Misinterpretation of Findings: One of the main functions of statistics is to establish relationship between the variables. But without carefully studying the data there should not be any attempt to establish relationship which may be misleading. Sometimes a certain degree of correlation may be apparent from a set of figures when actually no such correlation exists. The correlation may be because of mere chance, like there may be a positive correlation found between income and weight of persons by chance. But in fact the conclusion drawn from this finding, like those who earn more are more fat and heavy is fallacious.

7. Misleading Results: To gain knowledge about the behaviour of variables, people mostly make use of totals or average. In many cases totals of figures mislead and do not give a clear picture of the whole story. For

example it was reported in newspaper that 5 million people watched Asian Games at Beigng. This figure was arrived at by adding the number of people watching each game. There is every likelihood many people would have watched all games and counted each time. Thus it does not mean 5 million sports fans but a specified number of admissions and many individuals having attended more than one game. Similarly the arithmetic mean may be misused and its misuse may lead the user to trouble. There is an old saying that a statistician along with his family members had to cross a stream. He computed the mean depth of stream and concluded that they can cross the water. To his surprise, he and his family members got drowned because the stream was too deep in the middle.

8. False Assumption: In statistics mostly inferences are made on the basis of certain assumptions. Two such assumptions, which are involved in statistical reasoning are

- (a) the population is normal.
- (b) the sample is a random sample.

However in many cases the population does not confirm the normality requirement but when the assumptions are not met conclusion drawn would be wrong. For example while studying the intelligence of students we take students population to be normal. In fact student population is not normal because the intelligence of the students varies with the educational level of the parents, income level of the parents, the atmosphere where they live and study.

9. Undefined Units: The science of statistics deals with numbers and numbers presuppose units. Different investigators would interpret the statistical unit in different manner if the units are not properly defined. The collected data would be heterogeneous. Suppose we are asked to collect data in price of commodities the question crops up whether retail price or wholesale price of commodities. If we are to collect the share price index numbers unless it is rightly defined it may mean Bombay Stock Exchange Index, Bombay Stock Exchange Sensitive Index, RBI Index, Madras Stock

Exchange Index Economic Times Index, Financial Express Index etc. Thus there arises a need that statistical units employed in the collection of data must be defined and defined clearly so that there is no doubt about its meaning.

A layman has therefore to be cautious while making use of statistics or statistical inferences. If figures have been given without the context in which they are collected or if they are not complete or if they relate to phenomenon different from the one under investigation or even if the figures are correct and complete but a faulty or biased logic is applied to them the conclusion arrived at are bound to be wrong and it strengthens the belief that statistics are lies of first order. Unfortunately a set of figures cannot by itself, disclose whether it is dependable or not. All figures appear to be correct and innocent, figures do not lie but liars figure. It is therefore necessary that whenever we use statistics we should first of all make sure that they were properly collected and are suitable for the problem under investigation. In fact statistical methods are delicate tools and since they are liable to be misused easily they are very dangerous as well. The results of misuse of statistical methods and statistical data should not be used to discredit the science. Statistics are tools and can be used in any way we like, and it is in our own interest that we use them in a proper manner.

LIMITATIONS OF THE SCIENCE OF STATISTICS

Despite its universal applicability, the science of statistics has some limitations. It should not give an impression that statistics are magical devices which always provide correct solution to problems. Unless the data are properly collected and critically interpreted there is every likelihood of drawing wrong conclusions. The science of statistics may also be misused by prejudiced persons. In order to make proper use of statistics one has to keep in mind its limitations.

The following are the main limitations of the science of statistics:

- 1) Statistics does not study qualitative phenomena.
- 2) Statistics does not study individual measurements.

- 3) Statistical results are true only on averages.
- 4) Statistics is liable to be misused.
- 5) Statistics does not reveal the entire story.
- 6) Statistics is only one of the methods of studying the problem.
- 7) Statistics cannot prove anything.

1. Qualitative Phenomena not Studied: Despite of the universal application of science of statistics, it suffers from the limitation that it can only be applied to the problems which can be expressed in figures. Thus the qualitative phenomena have very little use of statistical methods. For example, the study of honesty, misery, intelligence, sincerity, etc. can not be done by the help of statistics. Thus the scope of statistics becomes limited because statistical methods cannot be applied to the study of qualitative nature.

Though data which are qualitatively expressed are outside the purview of statistics, attempts are made to convert qualitative data into quantitative forms. For example honesty itself may not be capable of quantitative analysis but may factors which are related to this phenomenon are capable of being expressed in figures and as such, can throw some light on the study of the problem. A study of number of thefts, cases of cheating, or swindling can indirectly tell something of the problem under study. Intelligence itself is not capable of quantitative analysis, indirectly it can be given numerical expressions because intelligence of students can be measured on the basis of marks secured by them in the examination. In fact now-a-days there is hardly any field where statistics is not applied. Also a small branch of statistics – the theory of attributes, deals with qualitative data.

2. Does not Study Individuals as it deals with Aggregates: Statistics deals with aggregates and these aggregates are obtained by adding up the individual items which are different from each other. Thus statistical methods have no place for an individual item in the series. Statistics though deals with aggregates for the purpose of analysis, very often the aggregates

are reduced to single figure. A statistical series is condensed into an average for purpose of comparison though individual item in the series has no specific recognition. Statistics cannot be of help for making a study of change which may have taken place in individual case. For example national income statistics would not show how the income of different persons has been affected by its rise or fall or who has amassed fortune and who has become poorer. They would only show whether the nation as a whole has grown richer or poorer.

3. True only on Averages: The conclusion obtained statistically are not universally true, they are true only under certain conditions. This is because statistics as a science is less exact as compared to natural science like mathematics, physics or astronomy. The statistical results are not exactly observed in the individual observation because the results are derived by taking a majority cases. Thus statistical inferences are uncertain. Due to this limitation in the statistical method, the conclusion arrived are not perfectly accurate and consequently the same conclusions cannot be arrived at under similar conditions all the times.

4. Is Liable to be Misused: The main limitation of statistics is that they are liable to be misused and mishandled. Any person can misuse statistics and draw any type of conclusion he likes. In reality statistical methods can be properly used only by trained people and their use by less expert hands is sure to give inaccurate results. To quote Prof. Bowley "statistics only furnishes a tool, necessary, though imperfect in the hands of those who do not know its uses and deficiencies". The figures may be stated without their context and thus fallacious conclusions may be reached. The argument that "in a country 20,000 vaccinated person died of cholera, therefore vaccination is useless" is statistically defective since we are not told what percentage of persons were not vaccinated and died. The data may be inaccurately compiled, deliberately manipulated and unscientifically interpreted and made to produce a false statistical argument.

5. Does not Reveal the Entire Story: Another limitation of statistics is that it cannot reveal the entire story of a problems. Since many problem, are affected by such factors, which are incapable of statistical analysis it is not always possible to examine a problem in all its manifestation only by statistical approach. Many problems have to be examined in the background of country's culture, philosophy or region. All these things do not come under the orbit of statistics.

6. Only one of the Methods of Studying a Problem: Statistical tools do not provide the best solution under all circumstances. Statistical evidences should be supplemented with other evidence.

7. Statistics Cannot Prove Anything: Some statisticians claim that statistics can prove any thing. On the other hand it is also claimed by others that statistics cannot prove anything. In fact they only describe a phenomenon quantitatively, classify it into parts, summarise a fact relating to each part and prepare the ground for logical inference. It may lead to unwarranted conclusion if the inferences are not drawn in appropriate surrounding which may be personal, political or social. Statistics does help in throwing light on problems but sometimes either ignorantly or dishonestly it may be misinterpreted to support a fallacious proposition. Statistics cannot give a result that can be an end in itself. Thus it is obvious that statistics plays an auxiliary role and not a basic one.

UNIT - II**LESSON - 1**

STATISTICAL ENQUIRY

- ❑ INTRODUCTION
- ❑ MEANING AND PURPOSE
- ❑ PLANNING AN ENQUIRY
- ❑ EXECUTING THE ENQUIRY

INTRODUCTION

Data constitute the foundation of statistical analysis. In social sciences where the use of statistical methods is indispensable, data can be collected through a statistical enquiry (survey or investigation). Thus a statistical enquiry is the most popular device of obtaining the desired data. It is a process of collecting data from existing population units with no particular control over factors that may affect the population characteristics of interest in the study. For example in the study of salary of workers' of a factory, the salary may be affected by a number of factors such as sex, educational level, nature of job etc. and as we get information about the workers' salary we have no control over these factors. They happen to be existing attributes of the workers. It should be noted that population is the totality of units under study and the population characteristics are the attributes of a population unit.

MEANING AND PURPOSE

The term 'enquiry' (or survey or investigation) means search for information or knowledge or truth. Statistical enquiry therefore implies search for knowledge with the help of statistical methods of collection, compilation, analysis, interpretation etc. In order to apply statistical methods to any problems, it is necessary that numerical facts are collected as statistical analysis is not possible without them.

A statistical enquiry may be either a general purpose enquiry or a special purpose enquiry. A special purpose enquiry is that in which data obtained may be useful in analysing a particular problem. An enquiry into the expenditure habits of working women residing in a working women's hostel of Pondicherry is a special purpose enquiry. On the other hand, in a general purpose enquiry, data obtained can be used for diverse purposes. For example data collected by the population censuses not only help in knowing the total population but also provide useful information about the structure of the population, male, female, literates, illiterates, employed, unemployed etc.

A statistical enquiry whether it be into an economic, sociological or political problem, passes through several stages before completion starting from planning and ending with writing the final report.

The following are the two steps which are to be taken to achieve purposeful result in statistical enquiry:

(1) Planning the enquiry

(2) Executing the enquiry.

PLANNING THE ENQUIRY

Planning the enquiry is the first step in collection of data. Proper planning of the survey is of paramount importance because the quality of survey results depends considerably on the preparation made before the investigation is conducted. Following are the matters which require careful consideration at the planning stage of statistical investigation:

1. Nature of the problem

2. Object of enquiry

3. Scope of the enquiry

4. Unit of data collection

5. Source of data (i.e. primary, secondary or both)

6. Technique of data collection (sample or census if sample method of sampling).

7. Nature and type of enquiry.
8. Choice of a format or construction of a format if none is available.
9. Degrees of accuracy required.

1. Nature of the Problem: A statistical enquiry is always undertaken to answer some questions which emerge from any important problem. But all types of problems cannot be statistically answered. Therefore the first thing to be observed by a statistical investigator is whether the problem, or more particularly, the question arising out of it can be quantitatively expressed. Such questions like how virtuous Sri Aurobindo was cannot be expressed quantitatively but what is the population of Pondicherry, what is the national income of India etc. can be quantitatively expressed.

2. Object of Enquiry: The determination of object of enquiry is a very important step in statistical investigation. This is so because the whole designing of the enquiry will be based on the object of enquiry. This will help in determining the scope of enquiry, source of data and lot other matters. If the object of enquiry is properly determined and defined many difficulties of collection and analysis of data are automatically removed. The knowledge of the purpose of enquiry helps as a guide in the statistical investigation. Thus failure to set out clearly the purpose of enquiry is bound to lead to confusion and waste of resources. For example if the object of enquiry is to find out the cost of living figures of certain families, a very simple procedure of family budget enquiry may be adopted, but if a detailed account of population living in a country is to be given relating to sex, income, literacy etc. a detailed census has to be taken up. Therefore care should be taken to clearly define the objective which will help to discriminate relevant data from irrelevant ones.

3. Scope of Enquiry: Once the purpose of enquiry has been clearly defined, the next important step is to decide the scope of the enquiry i.e. its coverage with regard to type of information, the subject matter, geographical area, time etc. If a very large quantity of data are collected they are likely to become unmanageable and it may not be easy to draw correct inference

from them. On the other hand if the quantum of statistical data collected is inadequate there is every possibility of incorrect conclusion. It is therefore essential that efforts should be made to come to a correct decision about the exact quantum of data that have to be collected.

The scope of a given statistical enquiry is influenced by the following three factors:

1. Object of Enquiry.
2. Availability of Time
3. Availability of Resources

The object of enquiry would decide the number of items to be covered. The enquiry must be completed within a reasonable period of time; otherwise the condition might change and the data collected may become useless and obsolete. The scope of enquiry also get influenced by the available resource like money, manpower etc.

The scope of enquiry fixes the limits of the enquiry and therefore careful consideration must be given while fixing the scope of enquiry.

4. Unit of Data Collection: Another important step before the collection of data is to define a statistical unit or units in which the statistical data are to be collected. As statistics are figures of measurement the data must be represented in units which must be defined carefully. The definition of a statistical unit depends upon the purpose of enquiry. The unit of measurement or counting when applied to statistical data is statistical unit. The definition of statistical unit is not as simple as it appears to be. For example prices may be wholesale, retail; wage may refer to money wage, real wage, task wage or kind wage. Thus it is very essential to give a clear and precise definition of the units to be employed in data collection.

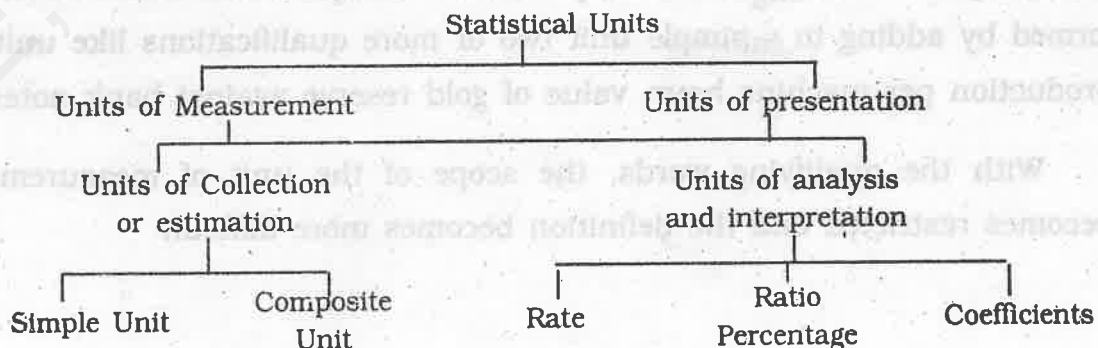
The essential requisites of statistical unit are:

- 1) The unit must suit the purpose of the enquiry.
- 2) It should be simple to understand i.e. it should be free from ambiguity so that the risk of collecting fallacious data is avoided.

- 3) It should be definite, specific, concise, and unmistakable. Such a unit should be interpreted in the same sense by different persons, so that collected data would be homogeneous.
- 4) It should be stable in character; if unit changes its characteristics every time, the measurement and counts would be misleading and would lead to wrong or even absurd conclusion. If the unit of measurement is yard one time and meter other times, the utility of data is lost. Thus the unit must be stable for meaningful compression.
- 5) The definition of the unit of data collection should be same throughout the investigation. Sometimes a social or economic concept is statistically comprehended in several forms and the variable used to measure it may admit of several definitions. That is, the concept of statistical unit may differ from enquiry to enquiry; for example capital may mean authorised capital, subscribed capital, paid up capital. The price may mean wholesale price, retail price. But when the general price level are studied we shall consider wholesale price whereas if the consumer price index is being constructed the retail price should be taken into account.

A statistical unit may be (a) arbitrary or (b) conventional. In other words it may be used in a special sense or in the sense prevalent in common usage. But whatever may be the sense in which the unit is used it is essential that the meaning should be clear and unambiguous.

Types of statistical units: The functional classification of units may be given as follows:



As the units of measurement would be similar to units of presentation, virtually the statistical units can be studied under the two heads (1) units of collection (2) units of analysis and interpretation.

(1) Units of Collection

The units of collection are those in terms of which data are collected. Counting is done in one of physical items and measurement is done in case of qualitative attributes. Examples of units of counting are number of persons in a family, number of children in a family, and units of measurement are like production in quintals, height in inches, weight in kilograms etc.

The units of collection may be (1) simple or (2) complex or composite units.

Simple Unit: A simple unit is one which represents a single condition without qualification such as wage, weight, ton yard etc. These units are not difficult to define and are in common use. But care must be taken in interpreting some of them. For example a wage may be defined as money wage or kind wage.

Composit Unit: A compound or complex unit is formed by adding qualifying words to a simple unit. For example electric power is measured in ton miles and ton miles are equal to number of tons multiplied by number of miles carried. Other examples of composite units are educated unemployed, factory building, working women, skilled worker, man hours, passenger miles etc. These are called compound or composite units where one qualifying word is added to a simple unit. A complex unit is one which is formed by adding to a simple unit two or more qualifications like unit of production per machine hour, value of gold reserve against bank notes.

With the qualifying words, the scope of the unit of measurement becomes restricted and the definition becomes more difficult.

(2) Units of Analysis and Interpretation

Statistical data are generally collected for making comparison. Comparison can be made with regard to time or space. These units include (i) rate (ii) rates and percentage (iii) coefficients.

A ratio is the relation between two quantities of the same kind in respect of magnitude. The relation is ascertained by expressing how many times or parts is the former of the latter. In other words if two quantities are similar, their ratio will be one divided by the other i.e. the ratio of A to B is expressed as $A:B$ or A/B e.g. the ratio of literates to illiterates is 100:400 or 10:40 or 1:4. All the ratios point to the same conclusion and are identical i.e. illiterates are 4 times of literates.

Rates are used in those cases where the comparisons are made between quantities of different kinds i.e. where the numerator and denominator are not of same kind, such as birth rates and death rates. Rate per unit are called coefficients. For example if it is stated death rate in India is 1.6 per cent or 16 per thousand, it means coefficient of death is 0.016. If this coefficient is multiplied by the total population we obtain the total number of deaths.

(3) Source of Data

Once the object and scope of enquiry has been laid down the next step is to decide about the source of data. The source of information may either be primary or secondary. The data collected for the first time by the investigator as original data are known as primary data. On the other hand if he obtains the data already collected, from published or unpublished sources, it is called secondary data. Quite often it is necessary to make use of both primary and secondary data. National income data collected by government is primary data but they become secondary data for those who use them.

(4) Techniques of Data Collection

The following are the two methods by which the required information in any statistical enquiry can be obtained: (i) Census method (ii) Sampling method.

A census is a complete enumeration of each and every unit of the population. Whereas, in a sample only a part of universe is studied and conclusion about the population is drawn from the sample study. The census type of enquiry requires a great deal of time, money and energy. In practice in most of the cases it is not possible to examine each and every item in the population. The investigator must decide which method of collection he will adopt. For example if the data about consumption pattern of people of Delhi are to be collected, the investigator has to decide whether heads of each family are to be contacted or whether heads of a few family are to be contacted. The former is the census method and the latter is sampling method. The choice of selecting between the two methods depends on the factors like (i) the availability of resources (ii) the times factor (iii) degree of accuracy desired (iv) nature and scope in the problem.

(5) Nature and Type Of Enquiry

The investigator also has to determine the nature and type of enquiry. Statistical enquiries may be of different types as follows:

- (a) Official, semi official or non-official.
- (b) Confidential or non-confidential
- (c) Regular and adhoc
- (d) Initial or original and repetitive.
- (e) Direct or indirect.
- (f) Extensive or limited inquiry
- (h) Main or subsidiary enquiry.

(a) Official, Semi-official or Non-official: An official enquiry is one which is conducted on behalf of Government. A semi-official enquiry is one which is conducted by bodies like RBI which enjoy Government patronage. Non-official enquiries are those which are carried out by private agencies,

institutions or individuals. The facilities available for collection of data differ according to the nature of organisation conducting the enquiry. Government may compel to supply information through legislation, semi-official organisation; may request for information and a private agency has to beg for the information.

(b) Confidential or Open: A statistical enquiry is termed confidential if the findings of the enquiry are to be kept secret and are not disclosed to general public. The results may be made known to the members of the organisation conducting the enquiry. For example chamber of commerce or a trade association may collect information for their use only. On the other hand open enquiry is one, the result of which are not kept secret but are made known to general public. Government enquiries are, in general, open enquiries.

(c) Regular or Adhoc: When the data are collected at regular intervals over a period of time, such an enquiry is called regular enquiry. On the other hand in an adhoc enquiry data are collected as and when necessary without any regularity. For example prices of different commodities may be collected on a weekly basis for years at a stretch, but enquiry about rural indebtedness in India was made once by All India Rural Credit Survey Committee.

(d) Initial or Original and Repetitive: If an enquiry is conducted for the first time it is initial original. The repetitive enquiry is one that is conducted in continuation of previous enquiries.

(e) Direct or Indirect: When the data are capable of direct quantitative measurement such as height, weight, income it is called direct enquiry. On the other hand when the subject matter of enquiry is not capable of direct quantitative measurement, for example honesty, intelligence, efficiency etc. they are called indirect enquiry. In case of indirect enquiry the investigator has to convert the qualitative information into some objective measurable phenomena before he can proceed to collect data. For

example intelligence can be measured by collective marks for various subjects which are taught to the members of the group.

(f) Extensive or Limited: When the number of persons involved are many it is called extensive enquiry. If we are making a management study for accommodation of the delegates coming to Pondicherry University for a conference it will be a case of extensive enquiry. If the sources are few in relation to the size of the enquiry then the enquiry is limited. Whether the enquiry will be extensive or limited depends on the availability of sources.

(g) Main and Subsidiary Enquiry: A main enquiry is that in which collection of data is the sole objective. On the other hand, in subsidiary enquiry, collection of data is not the main purpose. For example – collection of income tax return by Income-Tax Department is for administrative purpose. But the data in return also serve as statistical raw material.

(6) Choice of Frame or Construction of a Frame, if none is Available

The term 'frame' or population frame refers to the listing of all units in the population under study. The identification of the units in a population under study is often a difficult task. If we want to know the capital invested in small scale industries in Pondicherry we must have a complete list of all small scale firms. The list of names and addresses will be called a frame. The whole structure of enquiry is to a considerable extent determined by the frame. The method of survey suitable for one type of material may not be suitable for another type because of difference in the frame that is used. Consequently until the nature and accuracy of available frames are known, detailed planning of survey cannot be done. Various types of defects may exist in the available frames like (a) inaccurate, (b) incomplete (c) subject to duplication (d) inadequate (e) out of date etc. The investigator has to choose a frame from amongst the frames available. It is therefore essential to carry out careful investigation of the frame that is proposed to be adopted since the defects in the frames are not apparent until a detailed investigation has been made. If no frame for the particular enquiry exists, the investigator will have to make one himself.

(7) Degree of Accuracy Desired

The degree of accuracy which is desired must be decided in advance because the nature and quantum of data collected depends upon the desired level of accuracy. Of course as discussed earlier absolute accuracy is neither possible nor desirable in statistics. The degree of accuracy also depends upon the purpose, scope, measuring rod and the attitude of investigators. The reasonable degree of accuracy would also depend on the circumstances of each case. For example in weighing rice or wheat difference of a few grains may not be materially significant. Riggelman and Frisbee rightly pointed out "The necessary degree of accuracy in relation to its cost". However it does not mean that one should sacrifice accuracy to keep the cost low. It would ultimately depend upon the purpose of investigation. Quite often the investigator considers desirable to have approximate result in a shorter time than to have a slightly higher degree of accuracy at the cost of longer time or more money. The investigation has to prescribe a degree of accuracy first and then obtain the possible accuracy in the result. While determining the standard of accuracy needed in the collection of data two factors are to be kept in mind (a) the accuracy which is normally possible (b) the degree of accuracy that is considered necessary in the circumstances of a particular investigation.

EXECUTING OF ENQUIRY

After the plan of collection of data is prepared, the next step is to put the plan into operation. This is called executing the survey. The various phases of work subsequent to planning stage may be enumerated below.

- (1) Setting up an administrative organisation.
- (2) Designing the forms.
- (3) Selection, Training and Supervision of field investigators.
- (4) Control over accuracy of the field work.
- (5) Arrangement to follow up measures in case of poor or non response
- (6) Editing of the data.
- (7) Presentation of informations.

(8) Processing data and statistical analysis.

(9) Preparation of Report.

(1) Setting up an Administrative Organisation: This is the first step in the execution of an enquiry. The statistical data are collected through the medium of an administrative setup. The administrative setup required for an enquiry depends very much upon the nature and scope of enquiry. If an administrative and office organisation is already in existence, then the services of the persons therein may be utilised. In case of extensive enquiry regional offices may be set up in addition of a central office but if the scope of enquiry is limited a small office alone would be needed to do the job. The direction and control of the survey or investigation are the main responsibility of the administrative organisation. This organisation prepares the format on which the informations are to be collected.

(2) Design of Forms: During the course of enquiry many forms including questionnaire are required. As already stated the administrative organisation prepares the format or questionnaire in the light of the nature and scope of enquiry. However, careful attention should be given to the designing of various forms that will be used in the course of enquiry, specially the forms of questionnaire.

(3) Selection, Training and Supervision of Field Investigators: The execution of enquiry is in the hands of a large number of field workers called enumerators or investigators. The nature of the enumerators' job is such that great care has to be exercised in their selection. In fact, since the very success of investigation depends largely upon the investigators it is essential that they are properly selected, thoroughly trained and their work closely supervised. The enumerators or investigators should know the purpose of the enquiry, the manner in which the data are to be collected, the definitions of various terms used and the manner in which they are to conduct interviews of the respondents. The source from which field investigators may be drawn are (a) they may be specially appointed (b) they may be drawn from the existing staff in the case of a research institute (c) they may be individuals asked to undertake the work on voluntary basis

or on a small honorarium like school teachers, government employees engaged during census work.

The enumerators should be honest, intelligent, hardworking and should know the language of the respondents. They must have an amiable temperament, patience and knack to get correct information without in anyway annoying the respondents. After selecting the enumerators through suitable tests, proper arrangement should be made for training them. They should undergo a rigorous training programme through lectures and training manuals and should also be asked to conduct model interviews. It is also necessary to watch carefully the work of enumerators. From time to time the supervisory staff should themselves undertake field work in order to appreciate the difficulties involved in the enumeration work and to suitably guide the enumerators as regards those difficulties.

(4) Control over the Accuracy of Field Work: As soon as the investigators have started their work of collection of data, the supervisor should prepare a plan for checking the accuracy of the data collected by them. Steps must be taken to ensure that the survey is under statistical control, that the errors are random and no assignable cause of variation are present. This can be done by introducing field checks and should be conducted in such a manner that the investigators do not have prior knowledge of the work going to be checked.

(5) Follow up Measures in Case of Poor or Non Response: In spite of best efforts some of the respondents may not supply the desired information. A suitable machinery for dealing with such cases should be set up. The supervisor should chalk out a plan of follow up measures in case where desired informations are not collected. This plan may consist of contacting the persons on telephone, by telegram and if need be, by in-person visit. It is important to see that enumerators are not allowed to make substitutions for those not found. If this practice is followed the enumerators will not take pain to persuade the non respondents to cooperate and there will be a tendency to substitute for any one who is

not considered to be a good respondent. This may introduce bias in the survey result.

(6) Editing of Data: After the work of collection of data is complete and the questionnaire or schedules are handed over by the enumerators to the supervisors the supervisors should scrutinise these to check omission, inconsistency and inaccuracies.

(7) Presentation of Information: After the data have been edited and scrutinised, the next step is to present them in a simplified manner. The informations are classified and presented with the help of some statistical devices, like tables, diagrams, and graphs.

(8) Processing of Data and Statistical Analysis: After classification and tabulation, the data are analysed by the help of statistical tools such as average, dispersion, correlation etc. A great deal of survey works these days are tabulated and processed by computers. Computers not only save time but also make it possible to study large number of variables affecting a problem simultaneously. By the help of sophisticated techniques like computer simulation, it is possible to simulate and analyse the operation of extremely complex system which cannot be studied economically by other means.

(9) Preparation of Report: After the data have been analysed a report is drafted containing the findings of the survey. This is the final step in the execution of a survey. Two kinds of reports may be presented, either a general report giving the description of the survey and its result for those who are only interested on result, or as technical report giving details of sample design, computational procedure accuracy and allied aspects.

In a general report the following aspect of the survey or enquiry should be highlighted:

- (a) Statement of the purpose of survey.
- (b) Description of the coverage.
- (c) Collection of information

- (d) Numerical results.
- (e) Accuracy attained etc.

As regards the technical report the following aspects should be clearly brought to light:

- (a) Specifications of the frame.
- (b) Design of the survey.
- (c) Personnel and equipment.
- (d) Statistical analysis and computational procedure.
- (e) Comparison with other sources of information.
- (f) Observation of technicians

It should be noted that a statistical enquiry demands utmost care at each phase of the activity; poor work in one phase may defeat the purpose of the survey in which everything else is done well.

LESSON - 2

COLLECTION OF DATA

- ❑ INTRODUCTION
- ❑ DATA SOURCE
- ❑ PRIMARY AND SECONDARY DATA
- ❑ METHODS OF COLLECTING PRIMARY DATA
- ❑ SOURCE OF SECONDARY DATA
- ❑ PRECAUTIONS IN USE OF SECONDARY DATA
- ❑ CHOICE BETWEEN PRIMARY AND SECONDARY DATA

INTRODUCTION

Data constitute the foundation on which the superstructure of the statistical analysis is based. The policy decisions are taken on the basis of the interpretation of the result of the analysis of those data. Thus utmost care must be exercised while collecting the data, as if the data are inaccurate and inadequate the whole analysis may be faulty and the decision taken will be misleading. Hence the most difficult task for an investigator is to collect the desired information.

Data in statistics are as indispensable as raw material in production. In case, good quality raw material is not available for feeding the machine, quality production is not possible; similarly in the absence of accurate data it is not possible to analyse the true behaviour of variables. Thus the accuracy and reliability of the findings of an enquiry depend upon the nature of collected data in the same manner as the quality of manufactured product depends on the quality of material. Therefore data should be collected in a systematic manner to draw the right conclusion about the characteristic behaviour of the variable. The success of any enquiry depends upon the availability of accurate and reliable data. Reliability, in turn, depends on the appropriateness of the method chosen for data collection. Therefore data collection is a very basic activity in decision making.

DATA SOURCE

Data may be obtained either from the primary source or the secondary source. The source, by which data are collected originally for a certain purpose is called primary source. Thus a primary source is one that itself collects the data, a secondary source is one that makes available data which are collected by some other agency. A primary source usually has more detailed information particularly on the procedure followed in collecting and compiling the data.

It is preferable to make use of primary source where ever possible for the following reasons:

- (i) Primary source often includes a copy of the schedule and a description of the procedure used in selecting sample and in collecting data.
- (ii) Primary source usually shows the data in greater detail.
- (iii) Primary source normally includes the definitions of the terms and units used.
- (iv) The secondary source may contain mistakes due to errors of transcription made at the time of copying the figures from the primary source.

PRIMARY AND SECONDARY DATA

Based on the source of availability of data the data is generally classified to be

(a) Primary data.

(b) Secondary data.

The collection of data refers to purposive gathering of information relevant to the requirement of decision making. When data are collected for an investigation by actual observation, measurements and direct recording it is called 'Primary data'. Data which are not collected originally by the investigator but are obtained from other sources published or unpublished then the data is called "secondary data". The secondary data constitute the chief material on which statistical work is carried out in

many investigations. In fact before collecting primary data it is desirable that one should go through the existing literatures and learn what is already known of the general area in which the specific problem falls. This can help in getting an idea about possible pitfalls and avoiding duplication of efforts and waste of resources. The process of gathering primary data is called 'collection' of statistics and the process of gathering secondary data from different published sources is known as compilation of statistics.

The difference between primary and secondary data is only of degree. Data which are primary in the hands of one become secondary in the hands of others. Thus data are primary in the hands of collecting agency whereas for the rest of the world they are secondary. Secondary data offers the following advantages (i) It is highly convenient to use information which someone else has collected. Thus there is no need to set up the organisation of data collection and editing, tabulating the data collected. (ii) Secondary data can be obtained with less effort and less cost. (iii) On some subjects it may not be possible to collect primary data like, census data cannot be collected by an individual or research organisation, but can only be collected from government publication.

However three major problems of secondary data are (i) The difficulty in finding suitable data which would exactly fit to the requirement of the statistical enquiry. (ii) Problem of finding the data which is sufficiently accurate. (iii) The secondary data may contain the errors of transcription.

Source of Primary Data

Most of the organisations generate a good amount of data in their ordinary course of operation. Financial reports, data on manufacturing cost, personnel file constitute internal data. All these records are prepared by the direct observation by different levels of management within the organisation and hence can be treated as primary data. The primary data also can be collected with the help of specially designed questionnaire or schedule for interview.

Sources of Secondary Data

When the data are sorted from external published or unpublished materials, they are called as 'secondary sources'. Mostly the reports published by government, and non-government agencies constitute the secondary data. To mention some – The Economic Survey, Reserve Bank of India Bulletin, Statistical abstracts of Central Statistical Organisation, Public Enterprises Survey of Bureau of Public Enterprises, Bombay Stock Exchange Official Directory and so on.

METHODS OF COLLECTING PRIMARY DATA

Primary data may be obtained by applying any of the following methods.

- i. Direct Personal interview.
- ii. Indirect Oral investigation.
- iii. Information from correspondents.
- iv. Mailed questionnaire method.
- v. Schedules sent through enumerators.

1. Direct Personal Interviews: As the name suggests the investigator collects the information personally from the source concerned. It is necessary that in such case the investigator has a sense of observation, polite, courteous and has the knowledge of the language the respondent speaks.

2. Indirect Oral Investigation: In this method of collecting data the investigator contacts third parties or witnesses capable of supplying them necessary informations.

This method is generally used in those cases where the information to be obtained is of complex nature, the area to be covered is vast or the informants are reluctant to respond if approached directly. For example in an enquiry regarding addiction to drugs, the addicts may not be inclined to respond. In such case the investigator has to approach their friends, neighbours, relatives, dealers of drugs etc. In a similar manner clues about

thefts and murders are obtained by the police by interrogating third parties who are supposed to have knowledge about the problem under investigation. Enquiry committees and commissions appointed by government generally adopt this method. This success of this method depends upon the character of persons who are interviewed and of efficiency of the investigator who interviews such persons. The correctness of the information obtained depends on a number of factors like:

1. The knowledge of the informants. If the people who are contacted do not know the full fact of the problem under investigation, the purpose of investigator would not be served by this method.

2. Nature of the informant. If the informants are biased or prejudiced also correct information cannot be collected.

3. The honesty of the interviewers who are collecting the information. The informations may be twisted because of bribery, nepotism, threatenings etc. As a result correct conclusion cannot be arrived at.

4. The capability of the respondent expressing himself correctly and giving the true account.

When this method is employed for obtaining information, the investigator should not depend upon the information supplied by one man but a number of persons should be interrogated.

The major merits of this method are:

1. **Wide Scope:** This method may be employed for collecting data, where the scope of enquiry is wide. Most commissions and committees adopt this method as they have to conduct an extensive enquiry.

2. **Saving of Time and Money:** This method results in savings of time and cost as those persons only are interviewed who know the full fact.

3. **Secret Informations:** This method is suitable in those cases where informants hesitate to supply informations when contacted directly.

The main limitations are.

1) *It Consumes Time*: This method is quite time-consuming if the people refuse to cooperate or supply needed information and the investigator has to convince the persons to supply informations.

2) *Personal Bias*: This method is exposed to bias of both sides of the interview. The informant may not be able to express himself as well as the interviewer may be biased and he may interview those people who may be known to him.

3) *Poor Investigators*: The success of this method depends upon the qualities of investigator. If the investigators are ill trained they cannot do justice to their job.

Suitability: This method is suitable when the enquiry is exhaustive in nature and where the indirect source of information are required to be tapped either because direct source does not exist or cannot be relied upon.

The success of this method depends to a large extent on the character and efficiency of the investigator. The investigator should be polite, tactful and must have full knowledge of the time of enumeration, the area to be covered, the persons to be interviewed and the meaning of the questions to be asked to the informants.

The major advantage of this method are:

1) *Uneducated Respondents*: It is the most satisfactory method of data collection where the respondents are uneducated. The investigator can explain the purpose of the enquiry and the meaning of the questions to illiterate persons.

2) *Accurate Information*: The information received is more reliable, as the accuracy of the statements can be checked by the enumerator with the help of supplementary questions, whenever necessary.

3) *More Response*: This method eliminates to a great extent the problem of non response, as the enumerators go personally to obtain information. The respondents respond well as they are interviewed at their residence and also according to their own convenience.

4) *Supplementary Information*: The investigator may also collect supplementary informations which may be employed in the analysis of related problems.

This method is exposed to various limitations as follows:

1. *Expensive*: This method is quite expensive because a large number of investigators are to be employed for data collection. Usually the enumerators are paid persons.

2. *Untrained and Careless Investigators*: If the investigators are not properly trained or not trained at all and are careless, then they would collect unwanted data which may do more harm than good to the enquiry.

3. *Time Consuming*: This method is time consuming as the investigators have to go to the informants personally and if they are not available at their place the investigator has to make repeated visits.

4. *Biased Investigators*: The investigator might be a biased one and may not enter the answers given by the respondents truthfully. He may twist or suppress the information provided by the informants.

5. *Variation in the Answer Obtained*: Where there are many enumerators they may interpret various terms in the questionnaire according to their own understanding. Also the variation in the personalities of the interviewers will cause variation in the answer obtained.

Suitability: This method is most suitable where finance and trained enumerators are available to cover a wide field and where some significance is attached to the accuracy of results obtained.

In order to get accurate data by this method it is advisable to keep the following points in mind:

1. The questionnaire should be pretested to find out if the questions are proper and will be answered in the desired manner.
2. The enumerators should be properly trained, they should know the exact scope of the enquiry. They should be courteous and explain the object of enquiry with patience.
3. The questionnaire or schedules duly filled in should be scrutinised to detect any apparent inconsistency in the information provided by the respondent.

3. Information from Correspondents: The investigator collects the data through local agents or correspondents in different parts of the field of enquiry under this method. The correspondents submit the informations, collected through the agents, to the central office where the data are processed. This method is generally adopted by newspapers or periodicals and also various departments of government in those cases where regular information is to be collected from a wide area. For example in construction of wholesale price index regular prices of different commodities are regularly obtained from correspondents appointed in different areas. This method is particularly suitable for crop estimates.

The main advantages of this method are:

- 1) Cheap and appropriate for extensive investigation.
- 2) Quick results – It gives rough and approximate results very quickly where high degree of precision is not necessary.
- 3) Saves from botheration – This saves from the botherations usually associated with statistical investigation of other types.

The important limitations are:

- 1) Personal bias – This method is susceptible to personal bias. The data are collected by the correspondents in their own fashion and according to their own likings.

- 2) Not reliable - As the data collected by the correspondents may be prejudiced, it may not always ensure accurate results.
- 3) Suitability - This method is generally suitable for the cases where information needed is of regular nature and is collected from a wide area. In addition it suits where the rough and approximate estimates are desired.

4. Mailed Questionnaire: Collection of data through questionnaire is one of the most popular methods used these days. Under this method a list of questions pertaining to the survey (known as questionnaire) are prepared and is sent by post to persons from whom the informations are to be obtained. The questionnaire contains questions and provides space for answer. The informants send back the duly filled questionnaire within the stipulated time mentioned in the covering letter sent with the questionnaire. The success of this method depends upon the skill with which the questionnaire is drafted and the extent to which the informants are willing to co-operate. Since the questions' answers are sought through correspondence it lacks personal contact. Thus the form and tone of the questionnaire should be so designed to supply as far as possible the missing personal element. Because of the legal or administrative sanction, information required by the government departments are obtained through this method comparatively easily. In other cases it is necessary to take informants into confidence so that they furnish correct informations. This method of data collection is satisfactory only in the cases like:

- 1) When the respondent has interest in the enquiry.
- 2) When the respondent is under legal compulsion to supply the information.
- 3) When the questionnaires are sent by an association to its members.
- 4) When the questionnaire is accompanied with some gifts.

To make the method more effective the following points can be considered:

- 1) The questionnaire should be so prepared that the respondents should not feel it as a burden or taxing affair.

- 2) The sample should be large to neutralise the chances of non-response.
- 3) Prepaid postage stamps should be affixed so that respondents can mail the questionnaire without spending anything.
- 4) It should be adopted in the cases where there is under legal compulsion to supply the information.
- 5) It should be adopted in such enquiries where it is expected that the respondents would return the questionnaire because of their own interest in the enquiry.

The principle merits of this method are:

1. *Wide Scope*: It can be employed where the scope of enquiry is very wide.
2. *Convenient*: The person giving the information can fill-in the questionnaire at his convenience without being disturbed by the investigator at an inconvenient time.
3. *Less Expensive*: It is the least expensive method of data collection. Therefore most research workers and private organisations adopt this method.
4. *Confidential Informations*: Confidential informations may be given on a postal questionnaire which the informant may show hesitation to respond when he is contacted personally.
5. *Expeditious*: This method of collecting data is much more expeditious provided the respondents are willing to part with information and respond timely.

This method is exposed to the following limitations:

1. This method presumes that the informants are literate and understand the language of the questionnaire. This limits the scope of this method to certain investigations only.
2. Non response is a more serious problem in case of postal questionnaire. A large part of sample taken may not answer the

questionnaire. Thus it involves uncertainty about the response and cooperation on the part of respondents may be difficult to presume.

3. There are every possibility that respondents may not understand the meaning of the questions and may supply wrong answers and it may be difficult to verify.

4. **Suitability:** This method is appropriate where informants are spread over a wide area i.e. in case of extensive survey and when the respondents are educated and trustworthy.

5. Schedule Sent through Enumerators: Another method of collecting information is that of sending schedules through the enumerator or interviewer. Schedule is the name usually applied to a set of questions which are asked and filled in a face to face situation with another person i.e. by the interviewer. The essential difference between the mailed questionnaire method and this method is that in the former case the questionnaire is sent to the informants by post and in the latter the enumerator carries the schedule personally to the informant. The success of this method depends upon the character and the efficiency of the investigator. To get maximum information about the problem under study, the investigator needs acquaint himself with local customs, conditions and tradition so that he is in a position to identify himself with the persons from whom the information is sought.

Following are the advantages of Direct personal interview

(i) **Encouraging Response:** Response is more encouraging as the investigator approaches the informants personally and most people hesitate least to part with informations when approached personally.

(ii) **Accurate Data:** As the nature of enquiry is intensive and conducted personally, results obtained by this method are generally accurate and reliable. The investigator can clear up doubts of the informants about certain questions. In case the interviewers apprehends that informant is not giving accurate information, he may cross-examine him and thereby try to obtain the information.

(iii) *Supplementary informations*: It is possible through personal interview to collect supplementary informations about the informants' personal characteristics and environment which may be employed in the analysis and interpretation of results.

(iv) *Handling Delicate Situations*: The sensitive informations can be collected carefully and tactfully and handled effectively by a personal interview than by any other method of investigation.

The important Limitations of this method are:

(1) *Limited Seope*: If the number of persons to be interviewed is large and they are spread over a wide area this method cannot be useful as it requires personal attention of the investigator.

(2) *Expensive*: This method is very expensive because a large number of investigators have to be employed for collecting data. Individuals and small institutions cannot employ this method for data collection for want of funds.

(3) *Personal Bias*: The chances of personal bias and prejudice are greater in this method which may do a lot of harm to the investigation. The personal likes and dislikes of the investigator may defeat the purposes of the plan.

(4) *Time Consuming*: This method usually consumes more time as the interview can be held only at the convenience of the informants. In order to familiarise with the local conditions, customs and language in order to observe the phenomena properly investigator has to spend more time in the preparatory work.

(5) *Untrained Investigators*: The interviewers have to be thoroughly trained and supervised, otherwise they may not be able to obtain informations. If the investigators are poorly trained or untrained they may include unwanted materials and omit required materials, thereby spoiling the entire work.

Suitability: Despite the above mentioned disadvantages this method is favoured for obtaining information in intensive investigations. It is also recommended in case where it is required to get the correct and reliable information. This method gives very satisfactory results if the scope of enquiry is narrow and intensive at the same time and the investigators are dependable and unbiased.

Characteristics of a Good Schedule or Questionnaire

The constructions of a good schedule or questionnaire is essential for collection of primary data. The value of results depend greatly on the adquancy of questionnaire. The drafting of a questionnaire is the most difficult task and this job should be done by the experts as it is a specialised job and needs lot of skill and experience. Though there is no hard and fast rule for preparing a questionnaire, yet some broad characteristics are essentially to be followed for construction of any questionnaire.

1. Get-up of the Questionnaire and a Covering Letter: The questionnaire should be made attractive and interesting through proper presentation and layout. Every questionnaire should contain a covering letter containing the aims and objectives of the enquiry and the use that would be made of the information collected along with an appeal for seeking help and co-operation of the persons who are in a position to supply the information. The covering the letter also should contain the assurance as regards to the confidentiality of the responses of the respondents.

2. Size of the Questionnaire should be Small: Unnecessary details should be avoided and only relevant questions should be asked to keep the number of questions to a minimum. However the investigator should keep in mind that there are sufficient number of questions to cover the scope of study comprehensively. There is no hard and fast rule about the number of questions in a questionnaire. The precise number of questions would depend upon the object and scope of the investigation. Fifteen to twenty five maybe regarded as a fair number.

3. Questions Should be Short and Simple to Understand: There should be no ambiguity in the questions. Thus the Questions should not be confusing and should be capable of straight answer.

4. Personal Questions Should be Avoided: It is advisable as far as possible to avoid questions, which an informant may be unwilling or reluctant to answer as they may involve disclosure of private confidential or personal informations.

5. Questions should be in a Logical Order: The questions in a questionnaire should be so arranged that while reading through the questions the respondent should be able to understand the object of the enquiry. The questions if arranged logically helps tabulation and classification of data. The questions should not slip back and forth from one topic to other. Thus it is undesirable to ask a man what brand of toothpaste he uses before asking whether he uses toothpaste are not. Questions supplying identification and description of the respondents should come first followed by major information questions. Two different questions worded differently may be included in the same subject to provide cross check on important points.

6. Instructions to the Informants: The questionnaire should provide necessary instruction to the informants. An instruction sheet containing the operational definitions of various terms and concepts used in the questionnaire should be attached to the questionnaire. Instruction as regards the unit of measurement also should be given. Also instruction about the time within which the questionnaire duly filled in should be sent and to which address should be given.

7. Questions should be Capable of Objective Answers: The questionnaire should be so designed that wherever possible questions about opinions should be avoided. A set of possible answers may be given against each question and respondent may be asked to tick the correct answer.

8. Questions Regarding Calculations should be Avoided: Questions should not require calculations to be made. Like questions necessitating calculation of ratio, percentages etc. to answer a question should be avoided.

9. Pretesting of Questionnaire: The questionnaire should be pretested with a group before going for collection of data. The advantages of pretesting is that the shortcomings of the questionnaire can be discovered and it can be revised.

SOURCES OF SECONDARY DATA

In most of the studies the investigator finds it impracticable to collect the first-hand information on all related issues either due to shortage of money or lack of money. In such cases one can use secondary data for analysing the behaviour of the variables under the study. The sources of secondary data maybe divided into two categories published and unpublished.

1) Published Source are Usually:

- (a) Official Publications of international bodies like IMF, IFCI, United Nations Organisations.
- (b) Official publications of Central, State or local governments.
- (c) Reports and publications of trade associations, banks, co-operative societies, stock exchanges, trade unions, chambers of commerce etc
- (d) Technical trade journals like Commerce, Capital, Economic, India/ journal of Economics, and books and newspapers.

Unpublished Sources: All statistical data need not be published. majority of materials may be found with scholars, research workers, trad associations, chambers of commerce, government and private offices research institutions etc. Such sources can be used where necessary.

PRECAUTIONS IN USE OF SECONDARY DATA

Since secondary data have already been obtained it is highly desirab that before the investigator uses it he should make a proper scrutiny

the data. "Secondary data should not be accepted at their face value". Statistics collected by others cannot be depended fully as they may contain some pitfalls or limitations and unless they have been thoroughly scrutinised they should not be used. Thus before using the secondary data the investigator should confirm whether the following characteristics are present with the secondary data that is desired to be used.

1) Suitability to the Purpose of Investigation: It is essential for the analyst to satisfy whether the secondary data conforms to the purpose of study while using the same.

If any doubt develops in the mind of the analyst about the nature and scope of secondary data then such data should not be used. The suitability of data can be judged in the light of nature and scope of investigation.

2) Adequacy of Data for the Investigation: If the data are found suitable for the investigation they should be tested for adequacy. Adequacy of data is to be judged in the light of the requirements of the survey and the geographical area covered by the available data. The adequacy of the data also should be judged as regard to the degree of accuracy achieved in the data.

3) Reliability of Data: In order to determine the reliability of the published data it is better to enquire about the collecting agency, the methodology adopted in collecting and compiling the data, sampling procedure followed, degree of accuracy achieved etc.

4) Accuracy of Data: Analyst should also ensure about the accuracy of the secondary data as the accuracy of conclusions drawn depends mainly on the accuracy of collected data. For judging the accuracy of collected data the analyst should see

- (a) Whether the data collecting agency was unbiased, or biased.
- (b) Whether questionnaire were properly designed or not.
- (c) Whether the investigators were properly trained or not.

Editing of Primary and Secondary Data

Once the data have been collected either from primary source or secondary source they need to be scrutinised or edited to detect possible errors or irregularities. The task of editing is a highly specialised one and require great care and attention. However it should be remembered, data collected from internal records or published source is relatively simple than the data collected from a survey. While editing the data the following considerations need attention:

- 1) Completeness
- 2) Consistency
- 3) Accuracy
- 4) Homogeneity.

The secondary data also should be scrutinised because the data may be inadequate, inaccurate or unsuitable. However, it is quite difficult to verify secondary data to find inconsistency, probable errors or commissions.

The investigator must decide which source of data he can use. When one is deputed for collection of data one is tempted to go for secondary data because of its obvious easiness. But for more reliability of data one prefers a primary data to a secondary are.

CHOICE BETWEEN PRIMARY AND SECONDARY DATA

The choice between the two depends on the following consideration:

(a) Availability of Time: The Availability of time at the disposal of investigator affects the choice of the method. If data is needed immediately, choice is definitely for secondary source whereas if time is sufficient primary data is preferred because of its accuracy and reliability.

(b) Availability of Finance: The Availability of finance also influences the method to be adopted in collection of data. If the data collecting agency has vast financial resources at its disposal, it is better to adopt primary method. However shortage of funds forces one to go for secondary data even though secondary data are not as reliable as primary data.

(c) Availability of Trained Investigators: Availability of trained investigator also affect the choice of the source. In case of non availability of trained investigations secondary method is adopted in preference to primary one.

(d) Objective and Scope of Enquiry: The objective and scope of enquiry indicates the suitability of a particular source, thus the source is adopted accordingly.

(e) Degree of Accuracy: If the degree of accuracy is very high primary source invariably is to be adopted in preference to the secondary one.

LESSON - 3

SAMPLING AND SAMPLING METHODS

- ❑ INTRODUCTION
- ❑ OBJECTS OF SAMPLING
- ❑ PRINCIPLES OF SAMPLING
- ❑ METHODS OF SAMPLING
- ❑ MERITS AND LIMITATIONS OF SAMPLING

INTRODUCTION

Sampling is a procedure of making decisions by studying a few items regarding the characteristics of items in a population. By population we mean the totality of items. When the secondary data are not available for the problem under study, primary data are collected using any of the methods of collection discussed earlier. The required informations may be obtained either by census method or the sample method. A complete enumeration of all items under consideration in any field of enquiry which is regarded as the universe or the population, is known as census method of collecting data. In practice sometimes it may not be possible to examine each individual item in the population. may be because of its vastness or lack of time, money etc. or the complete enumeration of all items in the population may not be necessary. It may also be possible to obtain sufficiently accurate result by studying only a part of population. In such cases a few items are selected from the population in such a way that they represent the population and can be studied to draw inferences about the population. Such a section of population is called sample and the process of selection is sampling.

Thus following are the two ways by which the required information in any statistical enquiry can be obtained.

- 1) Census method or complete enumeration
- 2) Sampling technique

In spite of several advantages like highest accuracy, maintaining of the characteristics of universe in original, more reliability, completeness, free from sampling errors etc., the census method is not very popularly used in practice. The cost of complete enumeration is generally very large and in many cases is so prohibitive as regards to effort, time as well as cost that the very idea of collection of information may have to be dropped. In case the population is infinite or it destroys the population unit, the census method cannot be adopted.

Thus sampling as a method of learning about population on the basis of sample drawn from it has become indispensable in almost all fields of studies. Thus process of sampling involves three elements:

- a) Selecting the sample,
- b) Collecting the information,
- c) Making inference about the population.

These three elements are so closely interwoven that they cannot be considered in isolation.

Even though the theory of sampling has been developed only in recent years the idea of sampling is quite old. A housewife examines only two or three grains of boiling rice to know whether the pot of rice is ready or a doctor draws a few drops of blood and draws conclusion about the blood constitution of the whole body; people examining a handful of grain to ascertain the quality of the entire lot. In fact consciously or unconsciously sampling technique is used in every field of study. The characteristics observed in the sample are known as 'statistic' and the same is referred as 'parameters' in case of population.

OBJECTS OF SAMPLING

The most important aim of sampling study is to obtain maximum information about the phenomena under study with least sacrifice of money, time and energy. By the study of sample, one tries to have an idea about the whole population; thus sampling aims at obtaining the best possible values of the parameters. The aim is best achieved if the sample

studies are made in such a way that they disclose a mathematical relationship between the value of distribution.

PRINCIPLES OF SAMPLING

The possibility of reaching valid conclusions concerning a population on the basis of sample are based on two main important principles. They are:

- 1) Law of "Statistical Regularity"
- 2) Law of "Inertia of Large Numbers"
- 3) Principle of Persistence
- 4) Principle of Optimism
- 5) Principle of Validity

Though the principles of sampling are often referred to as Laws of sampling, in the stricter sense of the term they are just the tendencies which universally operate.

(i) Law of Statistical Regularity

This law is derived from the mathematical law of probability. L.R. Conner states that the law of statistical regularity lays down that a group of objects chosen at random from a particular group tends to possess characteristic of that group (universe). In the words of King "The law of statistical regularity lays down that a large number of items taken at random from each group are almost sure on an average to possess characteristic of the group". In other words the law points out that if the sample is taken at random from the population, it is likely to possess almost same characteristics as that of the population. The above two quotations emphasise two points namely,

- i) That the sample size should be large; as the size of the sample increases it becomes more and more representative of the universe and exhibit its characteristics. Thus larger is the sample the better would be the results. In actual practice very large samples create their own problem and become more expensive. A balance is desired to be struck between the

sample size the degree of accuracy required and the availability of financial and other resources.

ii) The second point emphasised is that the sample on the basis of which inference is drawn about the population must be selected at random. By random selection we mean a selection where each and every item of the population has an equal chance of being selected in the sample which means selection must not be made by deliberate exercise of one's discretion. A sample selected in thus manner would be representative of the population. Thus the law of statistical regularity makes possible a considerable reduction of work before any conclusion is drawn regarding a large universe and is of great practical significance.

(ii) Law of Inertia of Large Numbers

This principle is derived from the principle of statistical regularity. It is of great significance in the theory of sampling. It states that other things being equal, larger the size of the sample more accurate the results are likely to be. In comparison to small numbers, large numbers are relatively more stable in their characteristics. The difference in the aggregate results are likely to be insignificant in a large sample which is not so in case of small sample. For example if a coin is tossed 10 times though we expect to get equal number of heads and tails it is quite likely that we may get 8 heads and 2 tails. But if we toss it 100 times the result would be more dependable as we may get 60 heads and 40 tails. Similarly if the coin is tossed 1000 times the number of heads and tails would be very closer to each other. The basic reason of such likelihood is that the experiment has been carried out sufficiently large number of times and the possibility of variation on one direction getting compensated by other in different direction is more.

Other than this two main principles the other principles are as follows:

(iii) Principle of Persistence

If some items of the population possess some specific characteristics these characteristics will be found in the sample also and even if the sample

size increases the characteristics are supposed to be reflected in the same manner as before. For example if on an average 10% students pass the Chartered Accountants examination in first attempt, even if the population is increased, the Percentage would more or less remain the same.

(iv) Principle of Optimism

According to this principle size of the sample is maintained in such a way that the results are optimised in terms of cost and efficiency. Larger sample would give more accuracy but at a high cost. But this principle aims at obtaining a desired level of efficiency at the minimum cost and obtaining maximum possible efficiency with a given level of cost.

(v) Principle of Validity

A sample design is called valid only if the inferences drawn from it about the population are valid. For a valid conclusion the sample drawn has to be random, size to be adequate and data collection and analysis have to be scientific.

METHODS OF SAMPLING

When it is decided to take sample from the population, it is necessary to choose some methods of sampling. There are many methods of choosing a sample from the population. The choice of the sampling method depends upon the nature of data and the purpose of the enquiry. The various methods of Sampling also called "sampling design" can be grouped into two broad heads (1) Random sampling (2) Non-random sampling.

1. Random Sampling

It is otherwise called as probability sampling. In probability sampling all the items in the population have a chance of being chosen in the sample. However 'random sampling' does not mean haphazard selection or the term 'random sample' is not used to describe the data in the sample but the process employed to select the sample. Randomness is thus a property of sampling procedure instead of individual sample.

Advantages of Random Sampling: The following are the basic advantages of random sampling:

- 1) Random sampling provides estimates which are not biased.
- 2) Existence of detailed information about universe is not required.
- 3) Helps in evaluating the relative efficiency of various sample designs.

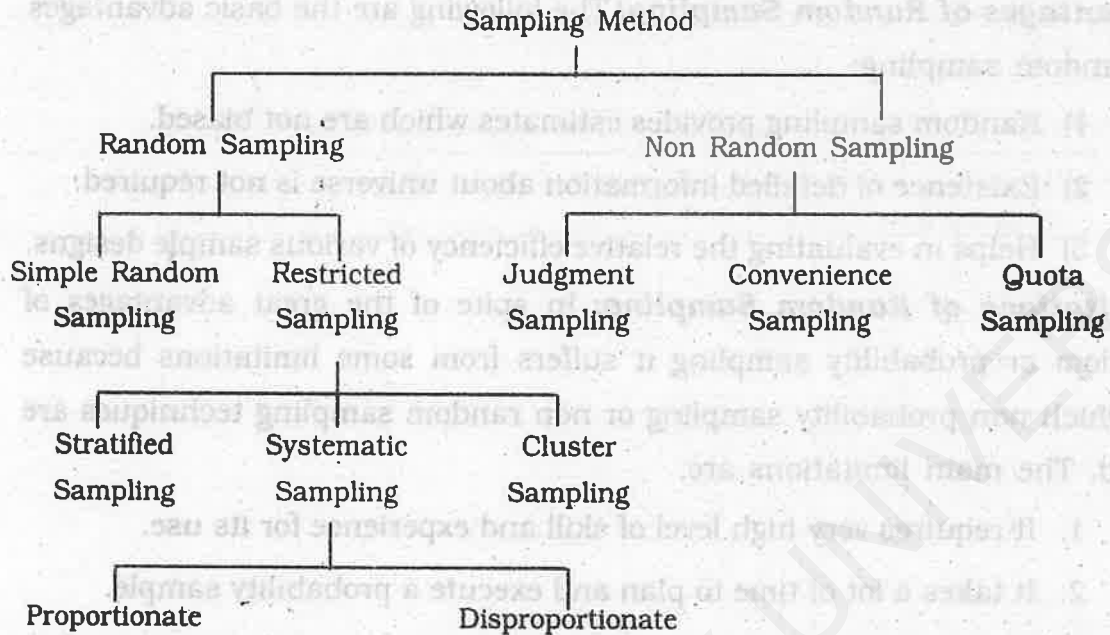
Limitations of Random Sampling: In spite of the great advantages of random or probability sampling it suffers from some limitations because of which non probability sampling or non random sampling techniques are used. The main limitations are:

1. It requires very high level of skill and experience for its use.
2. It takes a lot of time to plan and execute a probability sample.
3. The cost involved are generally high as compared to non random sampling.

2. Non-random Sampling

Non-random sampling is a process of sampling without use of randomisation. A non random sample is selected on the basis of some one's experience about the population or convenience. That is why it is otherwise called as 'Purposive sampling' where the basis of selection is other than probability consideration. Purposive sampling means selecting the items of the sample in accordance with some purposive principle. In this method the criterion for selection is laid down first and items are selected in accordance with it. Thus the probability of inclusion of some units are very high, while the probability of inclusion of others very low. In this sampling procedure personal element has a greater chance of entering into selection of sample. Sometimes the investigator may select a sample to yield favourable results – in results bias on sampling technique. In small enquiries this sampling design may be helpful with its relative advantage of time and cost but samples so selected do not possess statistical characteristics to make inference about the population.

The sampling methods can be classified further as follows:



A discussion of these methods is given below:

(a) Simple or Unrestricted Random Sampling

Simple Random sampling is the technique in which samples are so drawn that each and every unit of the population has an equal and independent chance of being selected in the sample. In simple random sampling the influence of personal bias is eliminated altogether. No factors other than pure chance affects the likelihood of an item being included in or excluded from the sample. In the selection of a random sample conscious effort should be made to ensure the operation of chance factor, so that the resulting sample is a random one. In order to eliminate bias in sample selection various mechanical devices also are often used for this purpose. To ensure randomness in the selection one may adopt either (1) lottery method or (2) consult table of random numbers.

1. Lottery Method: In this method all the items in the population are numbered. The numbers are written in small chits of papers or on cards which are homogeneous in all respects. These numbered chits of paper must be identical in size, colour and shape. All the numbered slips are then folded and placed in a drum or container and well shuffled. A blind folded selection is then made of the number of slips required to constitute

the desired size of the sample. The numbers corresponding to the slip drawn will constitute the sample. The selection of item depends entirely on chance. However the possibility of personal prejudice or bias cannot be ruled out if the slips are not of uniform size or shape. This method works well with small population. There is the added problem too of not being certain that the chits are not mixed properly.

2. Table of Random Numbers: The lottery method becomes quite cumbersome to use with a large population. Personal bias also cannot be excluded altogether. At the same time randomness in the drum is not as simple as it appears to be because slips may stick together or to the sides. If the slips are too much they do not get mixed thoroughly. When the population is too large it takes too much time to number individual members of the population. Since lottery method is not always practicable an alternative method of selection of random sample is employed. It is that of using the table of random number. Several standard random number tables are available of which any one can be used.

They are

- (i) Tippet's table of random numbers.
- (ii) Fisher and Yates table of random numbers.
- (iii) Kendall and Babington Smith numbers.

Among these tables Tippet's table of random numbers is most popular and widely employed for selecting random sample. The table consists 41600 random digits grouped in to 10,400 sets of four digit random numbers. These numbers have been put to various tests from time to time and their randomness have been proved. The digits in each column and in each row are in random number. It makes no difference from where one starts and in which direction one proceeds. The column arrangement is considered most convenient and the number of columns to be used depends on the size of the sample.

The method of drawing random sample comprises of the following steps:

1. Suppose we had a population of 600 units, identify all the units with a serial number ranging from 000 to 599.
2. Then select at random any page of random number table and pick up numbers given in it in any row, column or diagonal at random and also specify the sequence of selection of numbers.
3. Suppose we have specified to select first column and planning to move down to pick up further numbers then choose first three digits of the first number in random table. Verify whether it falls within our required range of 000 to 599. If it falls we accept the same as a number for our random sample; otherwise we discard it and proceed further.
4. Listing this procedure we select the numbers to the extent of required sample size.

Use of Random Number Table

Suppose we want 10 sample units out of 600 population size, using random numbers given in Appendix, if we follow 1st column and downward the following numbers would result: 126, 547, 371, 519, 171, 105, 421, 347, 041, 536.

Thus it may be easy to draw good random simple from Random number table but it is more suitable when the population is finite and it is possible to number them.

Merits and Limitations

Merits:

- 1) As the selection of items in sample depend entirely on chance sampling is not affected by personal judgment or bias.
- 2) The investigator can assess the accuracy of his results because the sampling error is inversely proportional to square root of number of items in the sample.

- 3) As compared to judgment sampling a random sample represents the universe in a better way. As the sample size increases it become increasingly representative of population.
- 4) It is relatively inexpensive and needs much less time and energy.

Limitations:

This method suffers from limitations like:

- 1) The use of simple random sampling necessitates numbering of the whole population. In case of very large population numbering the entire population is extremely costly and time consuming. In many cases, it is often impossible to number each member of the population.
- 2) The size of sample required to ensure statistical reliability is usually larger under random sampling.
- 3) In cases of heterogeneous population, the random sample will fail to depict the true characteristics of a population as some of the groups may not be represented at all in the sample.

b) Restricted Random Sampling

1. Stratified Random Sampling: When the population from which the sample is to be drawn is heterogeneous in nature having several segments or natural information simple random sampling fails to fetch a sample truly representing the population. Thus stratified sampling method is employed. This method attempts to design a more efficient sample than obtained by simple random procedure. This is done by dividing the population into different 'types' or 'strata' and then are combined to form a single sample of the universe. A Stratified sample is thus equivalent to a set of random samples of a number of population, each representing a single type or stratum. The purpose of stratification is to increase the efficiency of sampling by dividing heterogeneous population that (i) there is a great homogeneity within each stratum (ii) a marked difference between the strata.

Proportionate and Disproportionate Stratified Random Sample

A stratified sample may either be proportionate or disproportionate. In a proportional stratified sampling plan an equal proportion of units are drawn from each stratum, for example if the population consists of 5 groups their proportions are 5, 15, 20, 25 and 35 percent of the population and a sample of 1000 is drawn the desired proportional sample may be obtained in the following manner.

From stratum one	$1000 \times 0.05 =$	50
From stratum two	$1000 \times 0.15 =$	150
From stratum three	$1000 \times 0.20 =$	200
From stratum four	$1000 \times 0.25 =$	250
From stratum five	$1000 \times 0.35 =$	350
Size of entire sample		<u>1000</u>

In disproportionate stratified sample an equal number of cases is drawn from each stratum regardless of how the stratum is represented. Thus in the above sample 200 from each stratum may be drawn. In case of proportional plan the total number of samples would properly represent all the strata. This eliminates the difference between the strata and reduces sampling error. Such a method adds to the precisions of the sample estimate when within strata variability is least.

Limitations:

- 1) It needs complete and up to date list of finite population. Such lists are not generally available.
- 2) Utmost care must be exercised in dividing the population into various strata. Each stratum must contain, as far as possible homogeneous items, as otherwise results may not be reliable. In the absence of proper stratification the sample may have effect of bias.
- 3) It is tedious and time consuming to stratify the population; also in many cases information needed to set up groups may not be available.

- 4) The items from each stratum should be selected at random which needs skilled sampling supervision.

Systematic Random Sampling

The simple random sampling method is time consuming as the whole population is to be numbered and stratified random sampling involves division of population into homogeneous groups which is too technical. Thus another sampling technique can be adopted to avoid the above limitations of random sampling. In this form of sampling the first unit of the sample is selected at random from the universe and other units are drawn at a specified interval from the selected unit. This method can be used when the population is finite and the units in the universe are arranged on the basis of any system like alphabetical arrangement, numerical arrangement, geographical arrangement etc. The units of universe are then arranged in serial order. First the space interval has to be calculated taking into account the size of universe and size of sample required. The space interval represented would be N/n where 'N' is total size of the universe and 'n' is the sample size. Thus if out of 1,000 units 200 are to be selected k will be $1000/200 = 5$. Systematic sampling is relatively simple. This sampling technique is employed in those areas where complete list of items of population are available.

Merits and Limitations

Merits:

- 1) The systematic sampling method is simple and convenient to adopt.
- 2) The time and work involved in this type of sampling is relatively smaller.
- 3) If population is larger, systematic sampling gives us the result similar to that obtained by proportionate stratified sampling at much ease.

Limitations:

- 1) Systematic random sample is not the true representative when we are dealing with population having hidden periodicity.

- 2) If the population is ordered in a systematic way with respect to the characteristic the investigator is interested in, then it is possible that only certain types of items will be included in the sample.
- 3) In systematic sampling there is possibility of introducing error in the sampling process. For example we want to study the prices of the shares and randomly choose Friday as the day on which the price is to be recorded. We may get the lowest prices of the week, because it being the last day in the week the prices may be usually low.

Cluster Sampling

This is sometimes called as multistage sampling. As the name indicates there are several stages in which the sampling process is carried out. This sampling method is employed for selecting sample when populations are very vast and scattered over a wide area. Multistage sampling is an improved version of cluster sampling. In cluster sampling the universe is divided in to some recognisable subgroups which are called clusters. After this a simple random sample of these clusters is drawn and all the units belonging to the selected clusters constitute the sample.

For example we have to conduct an opinion poll in the city of Delhi. Then the city may be divided into say 60 blocks and out of these 60 blocks 6 blocks are picked up randomly and the inhabitants in these six blocks can be interviewed to give their opinion on a particular issue. While using this method, it should be seen that the clusters are as small as possible and the number of sample units in each cluster should be more or less same. The technique of cluster sampling is based on the principle that there is greater heterogeneity within clusters and greater homogeneity between the clusters. However in many cases this condition is rarely satisfied because well to do families live in the same locality and poor families in some other locality.

An improved version of cluster sampling is multistage sampling. In multistage sampling at first the first stage units are sampled by some suitable method such as simple random sampling. Then a sample of second

stage units is selected from each of the selected first stage units again by some suitable method which may be different or same as the method employed for the first stage units. Further stages may be added as required. The procedure may be illustrated as follows:

Suppose we want to take 5000 households from the state of Tamil Nadu. At the first stage the state may be divided into number of districts and a few districts are selected at random. At the second stage each district may be subdivided into a number of villages. A sample of villages may be taken at random. At third stage a number of households may be selected from each of the villages selected in the second stage.

Merits and Limitations

Merits:

- 1) The cluster or multistage sampling introduces flexibility in the sampling method which is lacking in the other methods.
- 2) It enables existing division and subdivisions of population to be used as units at various stages and permit the field work to be concentrated and yet large area is covered.
- 3) This sampling procedure is such that we need the second stage samples only for limited number of units i.e. those which are selected in the first stage. This leads to greater saving in cost.
- 4) This sampling method is administered with ease and convenience.

Limitations:

- 1) The principal demerit of the system is that the quantum of error in it may be large and variability of the estimates in this method would depend on the some position of primary and secondary units.
- 2) This sampling method assumes more homogeneity among clusters; however in reality there may be more heterogeneity among clusters. For example rich people travel by first class or air conditioned cars whereas poor people travel in ordinary compartments. Thus there is

greater heterogeneity among people traveling in different classes and more homogeneity within the class of travel.

- 3) The statistical theory of multistage sampling is more complicated in selecting sample by a single stage process.

NON-RANDOM SAMPLING METHODS

Judgment Sampling

In judgment sampling as the name means, selection of items to be included in the sample depends on the judgment discretion of the investigator. In other words the investigator exercises his judgment to choose the sample most typical of the universe with regard to the characteristics. In this sampling technique that character or qualities of the universe about which information is to be collected forms the basis of the judgment in selecting the sample. For example if investigation has to be done about the expenditure of students in a hostel, then under this system the investigator will pick up such students who are neither miserly nor extravagant.

Merits and Limitations

Merits:

- 1) It is mainly employed when the population to be sampled is very small. By the help of simple random selection chances of missing more important elements of the small population is there; whereas judgment sampling would certainly include them.
- 2) Judgment sampling is employed to conduct pilot studies to pretest the questionnaire which is to be used in general survey.
- 3) In solving everyday business problems of urgent nature this method is adopted by businessmen, executives, govt officials etc.

Limitations:

- 1) This method though simple, is unscientific in nature, where chances of personal bias in sample selection is quite high.

- 2) A poor knowledge of the population on the part of investigator may result in selecting unrepresentative samples for the study.
- 3) The principle of probability cannot be applied in judgment sampling. Thus there is no objective method for determining the extent and likelihood of sampling error.

Convenience Sampling

Convenience sampling is one in which a sample is obtained by selecting such units of the universe which may be conveniently located and contacted. Such a sample is selected neither on the basis of rules of probability nor even on the basis of judgment of the investigator.

Merits and Limitations: This method is usually employed in public opinion polls and the investigators generally interview people at railway stations, bus stand etc. Convenience sampling are subject to bias by their very nature. The results obtained following convenience sampling can hardly be representative of the population and thus are unsatisfactory. However this type of sampling can be used for making pilot study.

Quota Sampling

In quota sampling quotas are set up according to some specified characteristics such as so many of each of several income group, so many in each age etc. Each investigator, in this method, is assigned a fixed number or quota of persons from within certain well defined categories like sex, occupation, age, income etc. and within the quotas selection of sample items depends on personal judgment. The quota sampling is nothing but stratified purposive sampling.

Merits and Limitations: This method involves less money, energy and time compared to any other sampling technique. But because this method allows bias and prejudice to enter into the process of selection, the quota sample is not very popular.

Quota sampling is often used in public opinion poll. It occasionally provides satisfactory result if the interviewers are carefully trained and if they follow instructions closely.

CHOICE OF APPROPRIATE SAMPLING TECHNIQUE

It is very difficult to say that any one of the technique discussed above would always be the best. As each method has got its speciality no one method is regarded the best in all circumstances. Factors likes size of the sample, size of universe, availability of finance, time, nature of the universe etc., would influence the selection of a particular method of sampling.

Merits and Limitation of Sampling

Merits:

- (1) *Less Time Consuming:* Since sample is a study of a part of universe considerable time and efforts are saved not only in collecting data but also in its processing.
- (2) *Economical:* Although the amount of effort and expenses involved in collecting informations is always greater per unit of sample than the complete census the total financial burden is less than that of census.
- (3) *More Reliable Result:* Inaccuracy of informations, incompleteness of returns are likely to be more serious in census method than sampling method, as more effective precautions can be exercised in sample survey to ensure that information is accurate and complete. At the same time although sampling techniques involve certain inaccuracies owing to sampling error the result obtained is more reliable than that obtained from a complete count. It is always possible to count the sampling error, at the same time service of experts to impart thorough training to investigators in a sample survey can reduce the possibility of errors.
- (4) *More Detailed Information:* Since sampling technique saves time and money it is possible to collect more detailed information in sample survey.
- (5) *Non Applicability of Census Method in all the Studies:* There are some cases in which census method is inapplicable. If the population is infinite

or the investigation destroys the population unit, sampling method only can be adopted.

Limitations: Despite of several advantages, sampling is not altogether free from limitations. The followings are the main limitations of sampling:

- 1) Unless the sample survey is properly planned the results obtained would be inaccurate and misleading.
- 2) In the absence of qualified, experienced and unbiased persons the results, information from sample survey cannot be relied upon.
- 3) If the information is required for each and every unit in a population we need to depend upon census than sample survey.
- 4) There is every likelihood of sampling and non sampling errors.
- 5) There is possibility of bias on the part of investigator regarding inclusion or exclusion of items in the sample.

LESSON - 4

CLASSIFICATION AND TABULATION OF DATA

- ▣ INTRODUCTION
- ▣ MEANING AND OBJECTIVE OF CLASSIFICATION
- ▣ TYPES OF CLASSIFICATION
- ▣ FORMATION OF DISCRETE FREQUENCY DISTRIBUTION
- ▣ FORMATION OF CONTINUOUS FREQUENCY DISTRIBUTION
- ▣ TABULATION OF DATA, MEANING AND ROLE, PARTS OF A TABLE
- ▣ RULES OF TABULATION
- ▣ TYPES OF TABLE
- ▣ MACHINE TABULATION

INTRODUCTION

When the purpose of enquiry is defined, unit of data collection is decided and data are collected the next step in the statistical investigation is to classify type data. The data contained in schedules or questionnaires 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 as well. For the purpose of comparison, analysis, interpretation it is essential that data are to be put in condensed form. As long as collected data remain unorganised it is not possible to analyse the behaviour of data. Thus there arises the need to condense and simplify the raw data into such a form that the features of the data may be brought out. The procedure employed to reduce and to simplify raw data is called classification and tabulation. For the purpose of analysis and interpretation data have to be divided into homogeneous groups and are presented in a condensed form by the help of classification and tabulation of data.

MEANING AND OBJECTIVE OF CLASSIFICATION

Meaning

Classification ordinarily means grouping of related facts into class. Classification is a process of arranging a huge mass of heterogeneous data into homogeneous groups. In classification, items possessing common characteristics are separated from dissimilar items and placed in one class.

According to Connor, "Classification is a process of arranging things into different classes according to their resemblance and affinities and give expression to the unity of attributes that may subsist among a diversity of features". Classification of statistical data is comparable to the sorting operation in a post office. In classification, data units having a common characteristic are placed in one class, like the letters are classified according to their destination and dispatched in one bag, and in this fashion the whole data is divided into a number of classes

Objectives of Classification

The main objectives of classifying data are:

1. *Simplification of Raw Data*: The data in the raw shape are difficult to understand. Through classification homogeneous figures are grouped together, thus helping in understanding the data.
2. *Facilitate Comparison*: The technique of classification facilitates comparison of data within the class between the classes of the same series as well as of different series.
3. *Depicts Salient Features of the Data*: The technique of classification throws light upon the significant features of the data and one can understand the significance of data at a glance.
4. *Makes Data more Intelligible*: The process of classification differentiates the homogeneous figures from heterogeneous ones and also simplifies the statistical calculation like mean, median, mode, standard deviations etc.

5. *Saves Space and Time:* As the data are condensed and presented in a compact form it saves time and space.

6. *Eliminates Unnecessary Details:* it eliminates the unnecessary details found in the raw data and gives prominence to the important information gathered.

7. *Easy to Interpret:* It enables the statistical treatment of the material collected and help in interpreting the data.

A classification to be regarded as an ideal one, should have the following characteristics:

- 1) It should be unambiguous.
- 2) It should be stable to facilitate comparison.
- 3) It should be flexible so as to adjust to change condition.
- 4) It should cover the whole data.

TYPES OF CLASSIFICATION

Basically though classification of data depends on the nature of data, but more specifically it depends on the purpose for which the data is being processed. For example the data on the consumption of a particular variety of fast food can be classified based on regions to find out geographical popularity. The same data when classified based on income levels of consumers, it reveals the acceptance of fast food and economic status.

Broadly the data can be classified on the following four bases:

- 1) Geographical i.e area-wise or region-wise.
- 2) Chronological i.e. on the basis of time.
- 3) Qualitative i.e. on the basis of attributes.
- 4) Quantitative i.e in terms of magnitude.

1) Geographical Classification: Under this classification data are arranged according to geographical area, for example the production of wheat in India may be presented state-wise in the following manner:

Production of Wheat for the Year 1996 (Imaginary Figures)

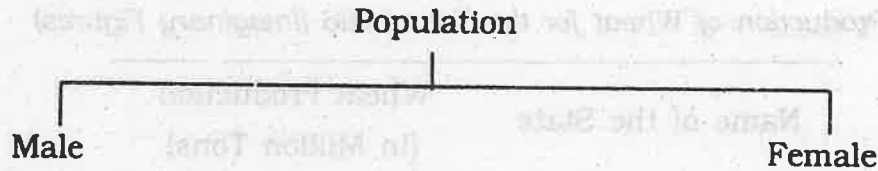
Name of the State	Wheat Production (In Million Tons)
Punjab	58
Haryana	40
U.P.	32
Bihar	20
Maharashtra	19
Others	10

2) Chronological Classification: When the data are classified in order of time, it is named as chronological classification; for example we may present the figures of production (population, sales, profit etc.) as follows.

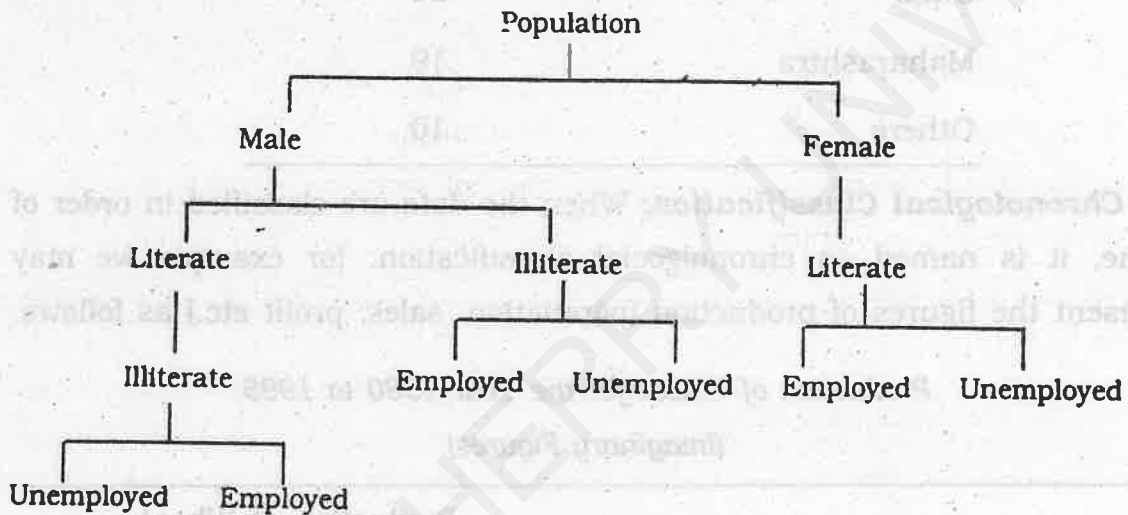
*Production of Wheat for the Year 1980 to 1995
(Imaginary Figures)*

Year	Production of Wheat (In Million Tons)
1980	200
1985	290
1990	350
1995	450

3) Qualitative Classification: In qualitative classification data are classified on the basis of some attributes or qualities like sex, colour of hair, literacy, religion, skill, etc. The point to note in this type of classification is that one can find out whether the quality is present or absent in the units of population under study. When only one attribute is studied two classes are formed; this is called as simple classification. For example the population under study may be divided into two categories as follows:



If instead of forming only two classes we further divide the data on the basis of some other attributes so as to form several classes, the classification is called manifold classification. An example of such classification is given below:



4) Quantitative Classification: When the data is classified according to some characteristics that can be measured quantitatively such as height, weight, income, etc. it is called quantitative classification. For example the following table shows the number of persons classified according to income.

Income	Number of Persons
50 - 99	17
99 - 149	20
150 - 199	25
199 - 249	35
250 - 299	19
299 - 349	14
	<u>130</u>

Such a distribution is called a frequency distribution. In this type of classification we generally encounter two terms. (i) variable, (ii) Frequency. In the above illustration income is the variable and the number of persons–frequency.

A frequency distribution refers to data classified on the basis of some variable that can be measured such as price, wage, age, number of units produced and consumed etc.

The term *Variable* refers to the characteristic that varies in amount or magnitude in a frequency distribution. A variable is a statistical terminology which stands for any measurable quantity. Examples of variables are height, of students, income of workers, production of food grains, weight of babies, price of commodities etc. In brief variable is any quality which can be quantitatively measured. Variables are of two types.

(i) Discrete Variable

(ii) Continuous Variable

(i) Discrete Variable refers to quantitative data which assumes only discrete values. Counting gives rise to discrete variable. Examples of discrete variable are number of children, number of books, number of students etc. Thus discrete variable is limited to certain numerical value of variable.

On the other hand a continuous variable is capable of manifesting every conceivable fractional value within the range of possibilities. Such as height weight etc. Thus continuous data is obtained by numerical measurement rather than counting. Series represented by continuous variables is called continuous series and series represented by discrete variables are called discrete series.

Frequency may be defined as the number of times a value appears in the series. Suppose 50 workers of a particular firm are paid Rs. 300 each as monthly wage the frequency of this wage is 50.

Frequency Distribution

When data are arranged into groups or classes according to conveniently established categories of range of observations, such arrangement in tabular form is called frequency distribution. The data which is represented by distinct groups is called class and the number of observations which fall into each class is known as frequency. When data is described by a continuous variable, the data is called continuous data distribution and when described by discrete variable, it is called discrete data distribution. The following are two examples of discrete and continuous frequency distribution.

Discrete Frequency Distribution		Continuous Frequency Distribution	
Size of Ready Made Dress	Number Sold	Wages per month	No. of Workers
6.5	40	0 - 1000	20
9.0	45	1000 - 1500	31
10.5	20	1500 - 2000	27
11.0	50	2000 - 2500	25
12.0	41	2500 - 3000	5
14.0	30	Above - 3000	2

Formation of a Discrete Frequency Distribution

The process of preparing this type of distribution is very simple. In case of discrete frequency distribution we count the number of times each value of variable is repeated, which is called frequency of that class. The task of counting usually is done with tally bar. A tally bar is simply one single indication of count, which would be helpful to count the occurrence of the observation in that class of distribution. To facilitate counting blocks of 5 tally bars are prepared and some space is left between each block. After counting we put bar opposite to particular value to which it relates. We finally count the number of bars corresponding to each value of variable and place it in the column entitled frequency.

To construct a discrete frequency distribution let us consider a sample in which 40 customers of a large departmental store purchasing packets of fast food noodles. The number of packets of noodles purchased by them are.

2 4 2 3 1 3 5 2 3 1 4 3 3 3 4 4 5 5 1 2
2 1 3 1 5 6 1 4 3 2 5 3 1 2 4 2 3 4 6 1

The data would emerge as a frequency distribution of following type.

No. of Packets Purchased	Tally Bars	Frequency
1		8
2		8
3		10
4		7
5		5
6		2

Formation of Continuous Frequency Distribution

If a variable can assume within a particular range, then continuous frequency distribution is prepared. For this purpose the entire data are divided into number of classes and then frequency of each class is obtained. The following technical terms are important to be known when a continuous frequency distribution is formed or data are classified according to class interval.

(1) Class Boundaries: The highest and lowest values that can be included in a class are known as class boundaries or class limits. For example take a class of 5 – 10. The two boundaries of class are known as upper limit and lower limit of the class which indicates that no item can belong to the class if its value is less than 5 or more than 10.

(2) Class Intervals: Class interval represents the width of the class. This is nothing but the difference between upper and lower limit of the class. To decide about the width of the class while constructing a frequency distribution a simple formula is used

$$i = \frac{L - S}{K}$$

i = width of class or class intervals

l = largest item

s = smallest item

k = number of classes.

(3) Class Frequency: The number of observations falling within a particular class is called its class interval. Total frequency of all classes would indicate total number of observations in the data set. In the following example frequency of the class of Income Rs. 100 – 200 is 30 which implies that there are 30 persons having income between Rs. 100 – 200.

Class Mark or Class Mid-Point: It is the value lying half way between lower and upper class limit of a class interval. Mid-point of a class can be ascertained as follows:

$$\text{Mid-point} = \frac{\text{Upper limit} + \text{Lower limit}}{2}$$

The mid-point of class is taken to represent the class for the purpose of statistical calculation. There are two methods of classifying data according to class interval.

(1) Exclusive Method

(2) Inclusive Method

(1) Exclusive Method: When the class intervals are so fixed that the upper limit of one class is the lower limit of next class, it is known as exclusive classification.

The following data are classified on this basis.

Income (Rs.)	No. of Persons
500 – 600	100
600 – 700	200
700 – 800	150
800 – 900	40
900 – 1000	10
Total	500

This method ensures the upper limit of one class is the lower limit of the next class. In the above example 100 persons have income between Rs.500 to 599.99. A person whose income is 600 would be included in the class 600 – 700.

(2) Inclusive Method: Under the inclusive method the upper limit of one class is included in that class itself.

The following example illustrates the method.

Income (Rs.)	No. of Persons
500 – 599	100
600 – 699	200
700 – 799	150
800 – 899	40
900 – 999	10
Total	500

In the class 500 – 599 we include persons whose income is between 500 to 599. If the income is 600 he is included in the next class.

To decide whether to use the inclusive or exclusive method it is important to determine whether the variable under observation is continuous or discrete one. In case of continuous, variable exclusive method must be used.

1) The number should not be too small or too large. Though there is no hard and fast rule the number of class should preferably be between 5 to 20.

3) The starting point i.e. the lower limit of the first class should be either zero or multiples of 5.

Let us consider an illustration to construct continuous frequency distribution.

[illegible]

In order to form the frequency distribution for the above data we have to consider the maximum and minimum values and their difference and if we divide by 10 it forms 5 class intervals.

Marks	Tally Bars	Frequency
20 - 30	II	7
30 - 40	I	16
40 - 50		15
50 - 60		9
60 - 70		3

TABULATION OF DATA

Meaning: Tabulation may be defined as systematic arrangement of data in columns and rows. It simplifies the presentation of data for the purpose of analysis and statistical inferences. The main purpose of tabulating the data is to depict the salient features of data and to facilitate comparison. According to H. M. Walkers and J. Lev "A well planned table is a unified, coherent, and in a sense, complete story about some aspects of set of data. Elements of the data are set up in rows and columns so as to indicate important relationship". It should be taken note of that rows are the horizontal arrangements whereas columns are vertical ones.

However, classification and tabulation should not be understood as two different processes. Classification is the first step of tabulation, thus they go together. After the data are classified they are displayed under different columns and rows so that their relationship can be properly understood.

Statistical tables are constructed in many ways. Their selection depends upon the use to which they are to be put and the purpose which the data have got to fulfil.

Role of Tabulation

Tabulation is a process of presenting data in condensed form so as to make the data processing convenient. Table makes it possible for the

analyst to present a huge mass of data in a detailed, orderly manner within a minimum space.

The significance of tabulation can be properly understood from the following points:

1. *It simplifies complex data and presents it in a compact form:* A systematic presentation of data in row and columns through tabulation gives a clear idea of what the table presents. Economy of time and space is affected by avoiding the repetition and all unnecessary details. Tables present the data in a compact and simplified manner, thus considered a better form in comparison to textual presentation.
2. *It facilitates Comparison:* As the data are presented in compact and organised form with the help of tables, comparison of figures is facilitated. It enables us to compare the values falling in different classes of the same table and also value falling in a particular class of different tables. Thus it permits intertable comparison.
3. *It gives Identity to the Data:* When the data are arranged in a table with a title and a number they can be identified distinctively and can be used as a source of reference in interpretation of a problem.
4. *It reveals Pattern:* Tabulation shows pattern within figures, which cannot be seen in the narrative form. By examining the table we can get information whether the frequencies are concentrated in the centre or at the ends or are equally distributed.
5. *It simplifies the Computation of Statistical Measures:* Tabulation facilitates the work of further analysis. Computation of statistical measure from organised data is simple and convenient in comparison to unorganised data which involves too much of time and is liable to wrong calculation.
6. *Error of Omission:* It is easy to detect errors and omission from tabulated data compared to untabulated data.

Parts of a Table

The number of parts of a table varies from case to case depending upon the data:

1. Table number
2. Title of the table
3. Captions or column
4. Stubs or row heading
5. Body of the table
6. Head note or prefatory note
7. Averages and totals
8. Foot note and reference
9. Sources

The following is a format of a table indicating the above parts:

Table No.

Title of the Table

(Head note)

Sub (Row Headings)	CAPTION		Total
	Column Heading	Column Heading	
STUB	Layout of Figures	Layout of Figures	Row Total
(Row Entries)	Layout of Figures	Layout of Figures	Row Total
Total			

Foot Notes:

Source:

1. Table Number: Each table should be numbered for easy identification. The number of the table should be given at the top, above the title of the table so that it may be easily noticed.

2. The Title: The title is placed on the top of the table below the table number. The title should indicate

(a) the nature of data (Population, production, sales etc.)

- (b) The area (country, state, locality) covered
- (c) The time period included (eg. 1990-95, Year ending 31st Dec etc.) The title should be clear, brief and self-explanatory. The title should be so worded it permits one and only one interpretation.

3. Caption: Caption refers to column heading. It explains what the column represents. Under column headings there may be sub-headings. The caption should be carefully worded and written in the centre at the top of the column. If different columns are expressed in different units the units should be mentioned with the caption.

4. Stub: The designation of the horizontal row or the data in the table are called stub. The function of horizontal row of numbers in the table is same as the column headings do for the vertical column numbers.

5. Head Note: It is placed just below the title in a smaller or less prominent type and usually put in brackets. It is used to explain points relating to the whole table that have not been included in the title or in caption and stubs. For example unit measurement is frequently written as a head note such as "in thousands", "in million ton" or "all figures in crores of rupees" etc.

6. Averages and Totals: Averages are placed either at the bottom or the right of the number averaged. The table should contain sub-total for each separate class and grand total for all classes.

7. Foot Notes: The foot note is a phrase or a statement which makes clear some specific items which the reader may find difficult to understand from the title. Caption and stubs should be explained in footnotes. If footnotes are needed they are placed directly below the body of the table. There are various systems of identifying the footnotes. One is numbering them consecutively with small number 1, 2, 3, 4 or better a, b, c, d. Another system of footnote is putting (*) stars. Sometimes instead of stars another sign is used.

8. Source: It is always placed below the table and helps in the authority for the data and as a reference for additional data if needed. In the case of original data the 'source' is omitted.

GENERAL RULES FOR TABULATION

It is difficult to lay down any hard and fast rule for tabulating data as it depends on the nature of data and its purpose. At the most some general rules may be given. The following are some general conditions which should be kept in mind while tabulating the data.

1. The table should be accurate, attractive and neat.
2. The table should suit the size of the paper usually with more rows and column.
3. The table should be constructed in such a way that it is easily read, easily understood and the data in it can easily be compared.
4. The captions, stubs should be arranged in some systematic order like alphabetical, chronological, geographical, etc.
5. The title should be written at the top in the centre.
6. The title should be self explanatory of material included in the table.
7. Heading of the columns and rows should be clear without any ambiguity.
8. The point of measurement should be clearly defined and given in the table.
9. Figures should be rounded off to avoid unnecessary details in the table and footnote to this effect should be given.
10. The figures which are compared should be placed near to each other preferably in vertical fashion.
11. A column entitled miscellaneous should be added for data which don't fit to any classification.
12. Percentage and ratios should be computed and shown if necessary.

13. Indicating a zero quantity as zero and not using zero to indicate information not available. If it is not available it can be indicated with letter NA or by dash (-).
14. Abbreviations should be avoided especially in title and headings for example 'Yr' should not be used for year.
15. It need to be explicit. The expression 'etc.' should not be used in the table.
16. Ditto marks should not be used. If a figure is repeated it should be shown each time, as ditto marks may be mistaken for the number (eleven).
17. The convenience of the person who needs the table may be consulted.

TYPES OF TABLES

The main purpose of tabulation is to condense the huge mass of data thereby facilitate the comparison of data. The type of table to be employed varies according to the nature of data and the requirement of the survey. In this section we discuss four main types of tables.

1. General purpose table
2. Special purpose table
3. Simple table
4. Complex table

1. General Purpose Table: The general purpose table is also called as primary or repository or reference table. Such table contains detailed or summed up information and are not constructed for specific discussion. These tables are not used for analytical purpose, rather they serve as repository of information and are arranged for easy reference. These tables are generally given in appendix if included in the report. Table published by governmental agencies are of this type. The main characteristic of such table is that they contain actual figures instead of rounded or percentage value.

2. Special Purpose Table: Special purpose tables are known as summary table or text or analytical table or derivative table. These tables are employed in the analysis of special problem. A special purpose table should be designed in such way that the reader may easily refer to this table for comparison, analysis or emphasis concerning the particular discussion. In case, they are embodied in the report they are always found in the body of the text. A special purpose table is always smaller in comparison to a general purpose table.

3. Simple Table: A Table which presents the measurement of a single set of items is called simple table. In simple table only one characteristic is shown. Hence this type of table is called one way table. Illustration of a simple table is given below:

Number of Employees in Pondicherry University According to age Groups	
Age in years	No. of employees
Below 25	-
25 - 35	-
35 - 45	-
45 - 55	-
Above 55	-
	Total

The above takes only one characteristic that is age into consideration.

4. Complex Table: A table which presents the number of measurements in more than one group of items is called a complex table. Such a table is set out with additional rows and columns. It shows the relationship of one set of data to other. Such table facilitates comparison among related facts. In such tables either stubs or captions are divided into subgroups. The number of characteristics represented in the table should not be more than four; otherwise the table would become too complex to understand

and the very purpose of tabulation would be lost. The following are examples of two way, three way and four way tables.

Two Way Table

NUMBER OF EMPLOYEES IN PONDICHERRY UNIVERSITY
IN DIFFERENT AGE GROUPS ACCORDING TO SEX

Age in years	Employees		Total
	Male	Female	
Below 25	-	-	-
25 - 35	-	-	-
35 - 45	-	-	-
45 - 55	-	-	-
Above 55	-	-	-
Total	-	-	-

Three Way Table

NUMBER OF EMPLOYEES IN PONDICHERRY UNIVERSITY
ACCORDING TO AGE GROUP, SEX AND NATURE OF JOB

Age in years	Nature of Job						Total		
	Teaching			Non-Teaching					
	M	F	Total	M	F	Total	M	F	Total
Below 25	-	-	-	-	-	-	-	-	-
25 - 35	-	-	-	-	-	-	-	-	-
35 - 45	-	-	-	-	-	-	-	-	-
45 - 55	-	-	-	-	-	-	-	-	-
Above 55	-	-	-	-	-	-	-	-	-
Total	-	-	-	-	-	-	-	-	-

Note - M indicates male & F indicates Female.

Four Way Table

**NUMBER OF EMPLOYEES IN PONDICHERRY UNIVERSITY ACCORDING
TO CATEGORY, AGE GROUP, SEX AND NATURE OF JOB**

Category	Age in years	Nature of Job						Total		
		Teaching			Non-Teaching					
		M	F	Total	M	F	Total	M	F	Total
FC	Below 25	-	-	-	-	-	-	-	-	-
	25 - 35	-	-	-	-	-	-	-	-	-
	35 - 45	-	-	-	-	-	-	-	-	-
	45 - 55	-	-	-	-	-	-	-	-	-
	Above 55	-	-	-	-	-	-	-	-	-
OBC	Below 25	-	-	-	-	-	-	-	-	-
	25 - 35	-	-	-	-	-	-	-	-	-
	35 - 45	-	-	-	-	-	-	-	-	-
	45 - 55	-	-	-	-	-	-	-	-	-
	Above 55	-	-	-	-	-	-	-	-	-
SC/ST	Below 25	-	-	-	-	-	-	-	-	-
	25 - 35	-	-	-	-	-	-	-	-	-
	35 - 45	-	-	-	-	-	-	-	-	-
	45 - 55	-	-	-	-	-	-	-	-	-
	Above 55	-	-	-	-	-	-	-	-	-
-	Total	-	-	-	-	-	-	-	-	-

Note - M indicates male and F indicates Female

FC denotes Forward caste

OBC denotes Other backward castes

SC/ST denotes Scheduled caste and Scheduled tribe.

MACHINE TABULATION

The importance of machine tabulation cannot be over-emphasised in case of data collected through survey conducted on large scale. The sorting, tabulating and analysis of data if carried out manually are time consuming. If all these three processes are done mechanically the process is known as machine tabulation. Mechanical sorting and tabulating are done with the help of cards known as 'punch cards', 'tabulating card' or 'cord card'. The data from a source is recorded by punching holes at appropriate location of the card. The other machines then sort and count the cards and print or otherwise record the result as well as check the accuracy of punching.

Advantages of Mechanical Tabulation

1. Data tabulation work is done in a much shorter time, which if done manually would be very time consuming.
2. Surveys conducted extensively on a large scale can be handled conveniently.
3. It ensures grater accuracy of the result as after the cards have been punched the accuracy of punching is verified by another machine called verifier.
4. Monotonous and routine nature of work are transferred to machine.
5. Because of the advantages mentioned above, the cost of collection of data are also reduced considerably.

UNIT - III**LESSON - 1**

MEASURES OF CENTRAL TENDENCIES

- ❑ INTRODUCTION
- ❑ DEFINITION
- ❑ OBJECTIVES OF AVERAGING
- ❑ REQUISITES OF A GOOD AVERAGE
- ❑ TYPES OF AVERAGES
- ❑ CHOICE OF AN AVERAGE
- ❑ LIMITATIONS OF AVERAGES

INTRODUCTION

Classifying and tabulating the data and then presenting it in the form of frequency distribution is the first step in making a long series of data comprehensible. Once the frequency distribution are formed, the next step is to study the characteristics of data to make them comparable. It leads us to summarise and condense them in such a way that their distinguishing feature can be expressed in an intelligible manner. If for example one has to compare between the marks obtained by a group of 200 students of Pondicherry (Central) University with another group of 200 students belonging to Hyderabad (Central) University it would be impossible to arrive at any conclusion if the two series relating to these marks are directly compared. On the other hand if each of these series is represented by one figure, understanding the standard of students in each university as well as the comparison would become an extremely easy affair. It is obvious that a figure which is used to represent a whole series should neither have the highest value of the series nor the lowest value but a value somewhere between these two limits, possibly in the centre where most of the items in the series concentrate. Such 'Measures of Central Tendency' are averages. The average represents a whole series and its value

always lies between the minimum and maximum values and is generally located in the centre or middle of the distribution. Thus the most popular statistical measure which helps to get a single value that describes the characteristic of an entire mass of data is the 'measure of central tendency' or 'average'.

DEFINITIONS

The word 'average' or the term 'measures of central tendency' have been defined by various authors. Some important definitions are given below:

Simpson and Kafka observe that "A measure of central tendency is a typical value around which other figures congregate".

Clark says that Average is an attempt to find one single figure to describe whole of figures.

As per A.E. Waugh " An average is a single value selected from a group of values to represent them in some way a value which is supposed to stand for whole group of which it is a part, as typical of all the values in the groups".

According to Ya-lun-Chou "An average is a typical value in the sense that it is sometimes employed to represent all the individual values in a series or of a variable".

"An average value is a single value within the range of data, that is used to represent all of the values in the series. Since the average is somewhere within the range of the data, it is some thing called a measure of central value" say Crouton and Cowden.

It is thus clear from the above definitions that an average is a single value which represents a whole series and is supposed to contain its major characteristics.

OBJECTIVES OF AVERAGING

There are two main objectives of the study of average.

1) To get a single value which is representative of the characteristic of the entire group of data: Measures of central value enable us to get a bird's eye view of the entire data which ordinarily are not easily intelligible. Measures of central tendency are the device to help human mind to understand the true significance of larger aggregate of facts. They put the concise picture of a mass of data setting aside the unnecessary details. Thus one value can represent thousands, lakhs and millions of values. For example it is impossible to remember the individual income of the millions of earning people of India and even if one could do so, it is hardly of any use. But if the average income is obtained we get one single value that represents the entire population. This figure can throw light on the standard of living of an average Indian.

2) To facilitate comparison: Measures of central tendency or averages reduce the mass of statistical data to a single figure. Thus they help in making comparisons. For example, the average marks obtained by two colleges in B. Com. examination affiliated to Pondicherry University would give reasonably a clear picture about the level of performance or the standard of students of the two colleges which would not be possible from the two full series of marks of individual students of the two colleges. However when such a comparison is made we have to be careful in drawing inference, as the marks of students in one college may vary within a small range and in the other college some students may have got very high marks and others very few marks. The comparison of averages in such cases may be misleading.

REQUISITES OF A GOOD AVERAGE

There are several measures of central tendency. Thus different types of averages may be employed to study the central position of a frequency distribution. Each type has got its own advantages and disadvantages. Therefore the question rises, what criteria should be prescribed for an ideal average? Since the average is a single value representing a group of values, it is desired that such a value satisfies the following properties:

- 1) *It should be easy to understand:* Since statistical methods are designed to simplify complexity, an average should be such that it can be readily understood.
- 2) *It should be rigidly defined:* The average should be rigidly defined so that different investigators interpret it in an identical manner. If an average is left to the estimation of an observer and if it is not a definite or fixed value it can be representative of the series as the observer may manipulate the meaning to his way of thinking. Such an average if it is rigidly defined, there would not be instability in value, and thus can be used for comparison.
- 3) *It should be based on all the observation of the series:* An ideal average should be based upon all observation. If some of the observations are not taken into account in its calculation, the average can not be said to be a representative one.
- 4) *It should be easy to calculate:* Not only the average should be easy to understand but also it should be simple to compute, so that it can be widely used. If complex mathematical formula are involved in this calculation its uses would be limited. But care should be taken not to sacrifice accuracy for the sake of simplicity in calculation.
- 5) *It should not be unduly affected by extreme observation:* Although each and every item should influence the value of average, none of them should unduly influence it. If one or two very small or very large items unduly affect the average i.e., either increase or decrease its value, the average can not be regarded as the typical of the series.
- 6) *It should not be affected by fluctuations of Sampling:* An ideal average should not be affected unduly by fluctuation of sampling, i.e., when different samples are drawn from the same field, an average should not differ significantly from one sample to the other. No doubt when two separate enquiries are made there is bound to be difference in the average

value i.e. average calculated from different samples will rarely be the same. But smaller the difference, better is the average.

7. *It should be capable of further algebraic treatment:* The most important property of an ideal average is that it must lend itself readily to further statistical computation. If the average cannot be manipulated algebraically, the average is of limited utility. For example if we are given two or more series of the same variables one should be able to compute the combined average of the said series.

TYPES OF AVERAGES

The following are the important type of averages.

A. Arithmetic mean

C. Mode

B. Median

E. Harmonic Mean

D. Geometric mean

F. Quartile

G. Deciles

Besides these there are less important averages like moving average, progressive average etc. These averages are not so popular as their applicability is limited to certain fields only.

Arithmetic Mean

The arithmetic mean is the most widely used measure of central tendency in statistical work. It is the value obtained by adding together the value of all the items and by dividing this total by the number of items. It is generally denoted by the symbol \bar{x} read as x bar. Arithmetic mean may either be (1) Simple Arithmetic mean or (2) Weighted Arithmetic mean.

Median

Median refers to the middle value of the distribution. As distinct from the arithmetic mean which is calculated from the value of every item in the series, the median is what is called a positional average. Median is the central value of the distribution or the value that divides the distribution into two equal parts. Fifty percent of the data in the distribution would be above the value of median and fifty percent would be below median.

Mode

The mode or the modal value is that value in a series of observation which occurs with the greatest frequency. The mode is often said to be the value which occurs most often, that is with highest frequency. Mode is defined as the value at the point around which the items tend to be most heavily concentrated in a frequency distribution. It is regarded as the most typical and fashionable value of a distribution. As per Alva M. Tuttle, mode is the value about which the items are most closely concerned. It is the value which has greatest frequency distribution in its immediate neighbourhood. There are many situations where mean or median fail to reveal the true characteristic of data. For example when we talk of most common wage, most common height, most common size of ready-made garments etc., mean often fails to represent the distribution due to presence of extreme item, and median owing to uneven distribution. Whereas mode refers to the value which occurs most frequently in distribution. Mode is the easiest to compute since it is the value corresponding to highest frequency.

Geometric Mean

Geometric mean is defined as the n_{th} root of the product of n items. Geometric mean would be helpful in averaging ratios, rates of changes. It is said to be appropriate in computing the average rate of growth of population, or average increase in the rate of sales, profit, production, gross national product etc.

Harmonic Mean

The harmonic mean is a measure of central tendency expressed as rates such as kilometres per hour, tonnes per day, kilometres per litre etc. This average is more suitable for finding out average speed of a vehicle average speed at which a work is done. The harmonic mean is defined as the reciprocal of arithmetic mean of the reciprocal of individual observations.

Quartiles

The quartiles is a measure which divides the series into four equal parts when the series is being arranged in the order of magnitude. There are three quartiles which divide a series into four equal parts. The first quartile is the value of the variable which divides the series into 25 percent of items below it and 75 per cent above it. The second quartile which divides the series into two equal parts which is also the median. The third quartile is the value of the variable that divides the series in a way so that 75 percent of items are below it and 25 per cent of items exceed the value.

Deciles

Decile is a measure which divides the series into 10 equal parts. Thus there are 9 deciles in a series.

CHOICE OF AN AVERAGE

The choice of an average is an important and difficult problem which a statistician has to face. If a wrong average is chosen it will lead to inaccurate conclusion. Thus the selection of average must be cautiously made. The selection of a particular average should be done after careful consideration of the following points.

- i) *The object or the purpose for which an average is being calculated:* If all the values in a series are to be equally treated, mean would be the suitable average. If most common item of a series has to be located, mode should be used. When the purpose is to determine rank of various values and to find the average rank, median is chosen. Geometric mean is the only average which suits to study the relative changes. Where smaller items are to be given more weightage, harmonic mean is the best average.
- ii) *Whether the average would be used for further statistical computation:* If further statistical computation is required only means like arithmetic, geometric or harmonic are suitable. Which of these three averages should be chosen depends upon the object for which the average has to be used. However median and mode are not suitable for further computation.

iii) *The nature and type of data available:* (a) In symmetrical or nearly symmetrical distributions arithmetic mean, median or mode can be used. Their value would be more or less identical in such distribution.

(b) Median or mode would prove to be good averages when the data is skewed or asymmetrical. Mean cannot be regarded a good average in skewed distribution.

(c) If the data are of unequal class intervals median is a better choice as it takes into account the class interval of median class only.

(d) If the data have open end class intervals median or mode would be better than any other average.

(e) If the data comprises rates, ratios and percentages geometric mean would be the most suitable average.

Sometimes it may be worthwhile to have more than one average and study them in accordance with their characteristics.

To sum up the above discussion, we can say the following:

Arithmetic Mean: Should be used (a) when the distribution is not skewed (b) the distribution does not have open-end intervals (c) when the distribution is symmetrical or evenly spread (d) rates, ratio or percentages are not studied (e) when the distribution does not have very large or very small items, as in such cases extreme items would adversely affect the average. (f) when further algebraic manipulation is necessary.

The Median: The median is most suitable average when a large number of items in a series have small values and one or two items have large values. The median is not affected by presence of high or low values like mean. This property makes it the most suitable average to represent observations in a skewed distribution. The median is the most appropriate measure in case of open end distribution. The characteristics which are not capable of direct measurement like intelligence, honesty, beauty in the case of enquiries to these characteristic median is the best average.

The Mode: The mode is used when we deal with qualitative data and where we have to find out preference of people. Consumer preferences are best studied with the help of mode as it gives the most common or fashionable items of the series.

In discrete series like average size of shirts, average size of shoes, average number of children per couple, average number of rooms in households etc. mode is the only appropriate average.

Mode also suits to represent open end distributions. The mode is best suited where there is an outstandingly large frequency.

The Geometric Mean: Geometric mean is useful in averaging ratios and percentages and computing average rates of increase or decrease. It is the only average which may be employed in computing the percentage increase in sales, production, exports over a number of years. Geometric mean is particularly important in economic and business statistics in construction of index numbers.

The Harmonic Mean: The harmonic mean is suitable when small items are to be given more weight. Harmonic mean is useful in problems in which values of a variable are compared with a constant quantity of another variable, like rates, time distance covered within certain time and quantities purchased or sold per unit etc. It is always preferred unless peculiar circumstances like extreme values of open end class are present.

However the statistician who desires to present the accurate picture of the data is the final judge to decide which average is the most appropriate one.

It may also be pointed out that a complete description of a distribution occasionally calls for two or more of these averages, though two or more averages creates an added burden for the investigator as well as the user of the statistics. Still, if more than a single measure presents a more complete and clear description of the distribution, it is desirable to consider more than one average.

LIMITATIONS OF AVERAGE

1) An average should be properly interpreted since average is the value that represents a group of values, otherwise there is every possibility of a wrong conclusion. An average represents a group of values and lies somewhere in between the two extremes i.e. largest and the smallest item of the series. Average should not be interpreted as the uniform value throughout the series. The average height of women may be less than the average height of men but it does not mean that no woman can be taller than a man. A commonly cited example to this effect is, a person had to cross a river. He was not aware of the depth of the river. So he enquired from another man who informed him that average depth is 5'4". The man was 5'6" and he thought he can easily cross the river because at all times he would be above the level of water. However he drowned when he tried to cross the river and lost his life. The man was drowned because he had a misconception that average depth means uniform depth throughout.

2) An average may give us a value which does not exist in the data. For example arithmetic mean of 250, 300, 370, 390, 60, 100 is

$$= \frac{250 + 300 + 370 + 390 + 60 + 100}{6} = \frac{1470}{6} = 245$$

which is a value not found in the series yet it is said to be the representative of the series.

3) At times an average results in absurd figure. If the average number of children per couple is 2.3, average number of divorces is 1.5 per couple, average size of the family 4.8, average number of rooms in 100 households is 3.7, we get absurd results, as it is impossible to set children, divorces, number of persons in a family, rooms in a household in fractions.

4) Averages do not throw light on the formation of series. Two or more series may have the same central value but differ widely in composition. For example observe the following three series:

Series - A	Series - B	Series -- C
200	50	500
200	100	300
200	200	100
200	300	75
200	350	25
Total 1000	1000	1000
Average 200	200	200

Though the averages are same, the three series differ substantially and the averages fail to disclose this fact.

5) Average is a measure of central tendency, Hence, unless the data show a clear single concentration of observations, an average may not be meaningful at all.

LESSON - 2

TYPES OF AVERAGES

- ❑ INTRODUCTION
- ❑ COMPUTATION OF DIFFERENT TYPES OF AVERAGES
- ❑ SPECIFIC USES OF DIFFERENT TYPES OF AVERAGES
- ❑ ADVANTAGES AND DISADVANTAGES OF DIFFERENT TYPES OF AVERAGES
- ❑ RELATIONSHIP BETWEEN DIFFERENT AVERAGES

INTRODUCTION

As already discussed earlier, the averages condense the data into a single figure which adequately describes the data. The measures of central tendency help us to find out some central value around which the data tends to cluster. The significance of measure of central tendency in statistical analysis cannot be over-emphasized. The averages of the data are of utmost significance in the sense that

- 1) It represent the data in a single figure.
- 2) It helps in comparison of two or more sets of data.
- 3) It helps to trace the precise relationship between different groups of numerical items.

Various type of averages are employed in statistical studies for summarizing the huge mass of data. We have discussed the various types of averages in the previous lesson. Hence in the present lesson the computational aspects of those averages are going to be discussed.

COMPUTATION OF DIFFERENT TYPES OF AVERAGES**MEAN**

Arithmetic Mean of a series is the figure obtained by dividing the total values of various items by their number. Arithmetic mean may be a) Simple arithmetic mean or b) Weighted arithmetic mean.

Calculation of Simple Arithmetic Mean (Ungrouped data or individual observation): The process of computing the mean in case of individual observation (i.e. where frequencies are not given) is very simple. Add together the various values of the variable and divide the total by number of items. Symbolically if $X_1, X_2, X_3, X_4 \dots X_N$ are the values of variable, the mean is computed by the formula

$$\bar{X} = \frac{X_1 + X_2 + X_3 + X_4 + \dots + X_N}{N} \quad \text{OR} \quad \bar{X} = \frac{\Sigma X}{N}$$

Where \bar{X} = Arithmetic mean; ΣX = Sum of all the value of variable X i.e. $X_1 + X_2 + X_3 + X_4 \dots X_N$; N = Number of observations.

Steps

- i) Find out the value of ΣX by adding together the values of X
- ii) Ascertain the number of observations
- ii) Divide ΣX by N .

Example: Let us consider the marks obtained by 15 Students in an examination along with their roll number.

Roll No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Marks	20	28	34	50	53	72	74	39	54	75	42	78	79	59	68

The mean mark obtained is calculated as

Roll No.	Marks	Roll No.	Marks	Roll No.	Marks
	X		X		X
1	20	6	72	11	42
2	28	7	74	12	78
3	34	8	39	13	79
4	50	9	54	14	59
5	53	10	75	15	68
				$N = 15$	$\Sigma X = 825$

$$= \text{Mean } \bar{X} = \frac{\Sigma X}{N} = \frac{825}{15} = 55$$

Average Marks = 55

The above method can be used only when items are few and size of the figures are small. The arithmetic mean can be calculated by using an assumed mean which is known as an arbitrary origin. This method is a short-cut method.

Short-Cut Method: In this method we assume any arbitrary mean and find out the deviation of items from the arbitrary mean. The sum of the deviation of items from the actual mean is supposed to be zero. Thus the sum of deviation from assumed mean will not be zero. If the total deviation is divided by number of items and added to the assumed mean we shall get the actual Arithmetic mean. Symbolically.

$$\bar{X} = A + \frac{\Sigma dx}{N}$$

Where \bar{X} = Actual arithmetic mean

A = Assumed mean

Σdx = The sum of the deviations from assumed mean

N = Number of items

Steps:

- 1) Take an assumed mean
- 2) Take deviation of item from the assumed mean and denote them by dx
- 3) Obtain sum of deviation i.e. Σdx
- 4) Substitute the values of A Σdx and N into formula.

The same example if solved by this short cut method will be as follows:

Roll No.	Marks	Deviation from assumed mean 50 dx
1	20	-30
2	28	-22
3	34	-16
4	50	0
5	53	3
6	72	22
7	74	24
8	39	-11
9	54	4
10	75	25
11	42	-8
12	78	28
13	79	29
14	59	9
15	66	18
N = 15		$\Sigma dx = 75$

$$X = A + \frac{\Sigma dx}{N} \quad (A = 50 \quad \Sigma dx = 75 \quad N = 15)$$

$$= 50 + \frac{75}{15} = 50 + 5 = 55$$

Thus, average mark is 55.

Calculation of Arithmetic mean – (Discrete series)

Direct method – In a discrete series frequency distribution the values of the variables are multiplied by their respective frequencies and product so obtained are totalled. This total is divided by number of items, which in discrete series is equal to total of frequencies and the arithmetic mean is obtained by the following formula.

$$\bar{X} = \frac{\Sigma fx}{N}$$

Where f = Frequency

x = Variable in question

N = Total number of observation i.e. Σf

Step:

- i) Multiply all the individual values of the variable x by their respective frequencies and obtain Σfx
- ii) Divide Σfx by N i.e. Σf for the total number of observation.

Example: The following data gives the number of children in 100 families in a certain village.

No. of children Per family	1	2	3	4	5	6	7
No. of Families	7	9	25	22	18	11	6

The average number of children per family can be calculated as follows:

No. of children	No. of Family	fx
1	7	7
2	9	18
3	25	75
4	22	88
5	18	90
6	11	66
7	6	42

$$\text{Mean} = X = \frac{\Sigma fx}{N} \quad (\Sigma fx = 400 \quad N = 100)$$

$$= \frac{400}{100} = 4$$

- ii) Take deviation of the variable X from the assumed mean dx which is $x-4$

Thus average number of children per family is 4.

Short-cut Method: To avoid laborious calculation of direct method, shortcut method can be adopted. According to this method

$$\bar{X} = A + \frac{\sum f dx}{N}$$

where A = The assumed mean

$$dx = X - A \quad N = \sum f$$

Steps:

- 1) Take a particular size of the item from the variable X as assumed mean (A)
- 2) Take deviation of the variable x from the assumed mean dx which is $x - A$
- 3) Multiply all the individual values of dx by respective frequencies to have the value of $\sum f dx$.
- 4) Divide $\sum f dx$ by N.
- 5) Substitute the values obtained in the formula.

Example: Computation of mean by short cut method.

No. of children (X)	No. of Family f	Deviation from Assumed dx - x - 4	f dx
1	7	-3	-21
2	9	-2	-18
3	25	-1	-25
4	22	0	0
5	18	1	18
6	11	2	22
7	8	3	24
N = 100			$\sum f dx = 0$

$$\bar{X} = A + \frac{\sum f dx}{N} \text{ where } A = 4; \sum f dx = 0; N = 100$$

$$\text{Substituting the value: } \bar{X} = 4 + \frac{0}{100} = 4$$

Average number of children per family is 4.

Calculation of Mean (Continuous series or grouped data): In case of continuous frequency distribution the method of computing mean is same as that employed in discrete series but with minor difference. In continuous series arithmetic mean is calculated by applying, Step deviation method also, other than usual direct and shortcut method.

Direct method: When direct method is used:

$$\bar{X} = \frac{\sum fm}{N}$$

Where m = mid point of various class

calculated as $\frac{\text{Lower limit} + \text{Upper limit}}{2}$

f = The frequency of each class

N = The total frequency

Steps:

- 1) Obtain the mid points of each class and denote it by m .
- 2) Multiply the mid points by the respective frequency of each class and obtain the total $\sum fm$.
- 3) Divide $\sum fm$ by N or $\sum f$.

Example: Calculate the arithmetic mean of the following:

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Number of Students	6	10	25	30	20	5	4

Marks	Midpoint	No. of Students	fm
x	m	f	
0 - 10	5	6	30
10 - 20	15	10	150
20 - 30	25	25	625
30 - 40	35	30	1050
40 - 50	45	20	900
50 - 60	55	5	275
60 - 70	65	4	260
$N = 100$			3290

$$X = \frac{3290}{100} = 32.90$$

Short-cut method: When short-cut method is used arithmetic mean is computed by applying the following formula

$$\bar{X} = A + \frac{\sum fdx}{N}$$

Where A = Assumed mean

dx = deviation of midpoints from assumed mean i.e. m-A

N = total number of observation

Steps:

- i) Obtain mid points of the items of the variables
- ii) Take one of the mid points as assumed mean (A).
- iii) Take deviation of the midpoints from the assumed mean i.e. dx = m-A.
- iv) Multiply the values of dx by their respective frequencies and obtain fdx.
- v) Apply the formula.

Example: The same problem can be solved by shortcut method as follows:

Marks x	Midpoint m	No. of Students f	dx (m-25)	fdx
0 - 10	5	6	-30	-180
10 - 20	15	10	-20	-200
20 - 30	25	25	-10	-250
30 - 40	35	30	0	0
40 - 50	45	20	10	200
50 - 60	55	5	20	100
60 - 70	65	4	30	120

Assumed mean: 35

$$\bar{X} = A + \frac{\sum fdx}{N} = 35 + \frac{(-210)}{100} = 35 - 2.1 = 32.9$$

Step deviation method: The computation of mean may further be simplified by taking step deviation in place of simple deviation. The only additional point here is, a common factor from the deviation from mid points is taken and the result at last is multiplied with the common factor.

The formula is

$$\bar{X} = A + \frac{\Sigma f dx}{N} \times C$$

Where A = assumed mean

f = frequency

$$dx = \frac{m - A}{C}$$

N = total number of observation Σf

C = Common Factor.

Steps:

- i) Write down the midpoints of all the class (m)
- ii) Select any midpoint as assumed mean (A) and write down deviation of midpoints from assumed mean dx
- iii) Find step deviation dx dividing the absolute deviation by a common factor (c)
- iv) Multiply dx by their respective frequencies and obtain fdx
- v) Apply the formula.

Example: The same problem is solved by step deviation method.

Marks x	Midpoint m	No. of Students f	dx (m-35)	(m-35)/10 dx	fdx
0 - 10	5	6	-30	-3	-18
10 - 20	15	10	-20	-2	-20
20 - 30	25	25	-10	-1	-25
30 - 40	35	30	0	0	0
40 - 50	45	20	10	1	20
50 - 60	55	5	20	2	10
60 - 70	65	4	30	3	12
N = 100			$\Sigma f dx = -21$		

$$\begin{aligned}\bar{X} &= A + \frac{\sum fdx}{N} \times C \quad (\sum fdx = -21 \quad N = 100 \quad C = 10) \\ &= 35 + \frac{(-21)}{100} \times 10 = 35 - 2.1 = 32.9\end{aligned}$$

Correcting incorrect values: Sometimes due to oversight or by mistake some wrong items are taken while calculating mean. To find out the correct mean from the wrong mean the procedure to be followed is as follows: From the incorrect $\sum x$ deduct wrong item and add correct item and then divide correct $\sum x$ by number of items.

Example: The mean marks of 100 students were found to be 40; later it was found that a score of 43 was misread as 48. Find the correct mean corresponding to correct score.

We are given $N = 100$ $\bar{X} = 40$

$$\text{Since } \bar{X} = \frac{\sum X}{N} \quad \sum X = \bar{X} \times N = 100 \times 40 = 4000$$

But this is incorrect $\sum x$

Correct $\sum X =$ Incorrect $\sum X =$ Wrong item + Correct item

$$4000 - 43 + 48 = 4005$$

$$\therefore \text{Correct } \bar{X} = \frac{4005}{100} = 40.05$$

Mathematical Properties of Arithmetic Mean

1. The products of arithmetic mean \bar{X} and the total number of values N which the mean is based is equal to the sum of all given values.
2. The algebraic sum of deviation of the given values from the arithmetic mean is equal to zero (refer problem solved by short cut method for discrete series). Thus it is regarded as the point of balance.
3. The sum of square of deviations is minimum when taken from the arithmetic mean.

4. If we know the sizes and means of two series we can find the combined mean of these groups by the formula

$$\bar{X}_{12} = \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2}{N_1 + N_2}$$

Weighted Arithmetic Mean: In the calculation of simple arithmetic mean, all items in a series are given equal importance. But it is not so in the practical life. In case some items in the distribution carry more importance than the other, the simple arithmetic mean is not the true representative average. Weights are assigned to each item/value according to its importance and then average is calculated to make it more representative. The term weight stands for relative importance of the different items. The formula for computing weighted arithmetic mean is

$$\bar{X}_W = \frac{\sum WX}{\sum W}$$

In the expanded form

$$\bar{X}_W = \frac{X_1 W_1 + X_2 W_2 + X_3 W_3 + \dots + X_N W_N}{W_1 + W_2 + W_3 + \dots + W_N}$$

where \bar{X}_W = Weighted mean

$X_1, X_2, X_3 \dots X_N$ = Values of the variables

$W_1, W_2, W_3 \dots W_N$ = Weights assigned to variables.

Thus weighted arithmetic mean may be calculated by multiplying the various items in a series by certain values known as weight and product so obtained is divided by the total weights. It should be noted that

- 1) Simple arithmetic mean shall be equal to weighted arithmetic mean if equal weights are assigned to all items.
- 2) Simple arithmetic mean shall be less than weighted arithmetic mean if greater values are assigned greater weight and smaller weights to smaller value.

- 3) Simple arithmetic mean is greater than weighted arithmetic mean as higher values are attached to smaller weight and greater weights are attached to smaller value.

Steps

1. Multiply the various values of a variable by their respective weights.
2. Sum up the products obtained in step 1
3. Divide the total of the products by total weights and the resulting value is the desired weighted mean.

Example: An examination was held to decide the award of scholarship. The weights given to the various subjects were different. Marks of three applicants were as follows:

Subjects	Weights	Marks		
		A	B	C
Statistics	4	60	65	63
Accountancy	3	64	70	63
Economics	2	56	63	58
Principles of Management	1	80	52	70

If the candidate getting highest mark is to be awarded the scholarship, who should get it?

Calculation of weighted average mean marks of A B C:

Marks	Weights		Marks	Weights		Marks	Weights	
X_A	W_A	XW_A	X_B	W_B	XW_B	X_C	W_C	XW_C
60	4	240	65	4	260	63	4	252
40	3	192	70	3	210	65	3	195
56	2	112	63	2	126	58	2	116
80	1	80	52	1	52	70	1	70
$\Sigma W = 10$		$\Sigma XW_A = 624$	$\Sigma W = 10$		$\Sigma XW_B = 648$	$\Sigma W = 10$		$\Sigma XW_C = 633$

$$\bar{X} W_A = \frac{624}{10} = 62.4 \quad \bar{X} W_B = \frac{648}{10} = 64.8$$

$$\bar{X} W_C = \frac{633}{10} = 63.3$$

Scholarship should be awarded to candidate B as he is having the highest marks.

MEDIAN

The median by definition refers to the middle value in the distribution when the values are arranged in the ascending order or in the descending order. Thus median value divides the series into two equal parts, one part having values more than the median and second part having values less than median. Median is called a positional average in contrast to arithmetic mean which is called a calculated average. When the values are arranged in order of magnitude the median value is the value of the $(N+1/2)$ th item.

In case of odd number of items in the series, median is an actual value. For example if there are 5 items the median is the value of $5+1/2 = 3$ rd item. But if there are even number of items in the series there is no actual value exactly in the middle of the series there is no actual value exactly in the middle of the series. In such case the median is arbitrarily taken to be half way between two middle items. For example if there are 10 items in a series the median position is 5.5, that is, median value is half way between, the value of items that are 5th and 6th in order of magnitude. Thus when N is odd, the median is the actual value and if N is even, the median is a derived figure.

Calculation of Median - (Individual observation)

Steps:

- 1) Arrange the values of the observation in ascending or descending order.
- 2) Find out the middle item by applying the formula $(N+1/2)$
- 3) If ' N ' is odd number median is the size of the $(N+1/2)$ th item. If ' N ' is even median is calculated as arithmetic mean of middle observations after they are arranged in ascending or descending order.

Example: The following data give the marks of 9 students.

98 115 110 117 126 107 102 119 130

In this problem n is 9. Hence median value is $9+1/2 = 5\text{th}$ value in the series.

The marks are first arranged in ascending order of magnitude:

98 102 107 110 115 117 119 126 130

$1/2 = 5\text{th}$ value in the series.

Hence median is 115.

Example: The following data relates to marks obtained by 8 students in statistics.

56 49 53 51 58 65 62 63.

Calculate the marks when arranged in ascending order.

49 51 53 56 58 62 63 65

Median is the size of $(N+1/2)\text{th}$ item i.e. $M = \text{Size of } (8+1/2)\text{th item} = 4.5\text{th item.}$

$$\text{Size of } 4.5\text{th item} = \frac{4\text{th item} + 5\text{th item}}{2} = \frac{56 + 58}{2} = \frac{114}{2} = 57$$

Thus median marks are 57.

Calculation of Median (Discrete series)

Steps:

- 1) First write the value of a variable in order of magnitude along with their frequencies.
- 2) Find out the cumulative frequencies by an operation of addition.
- 3) Apply the formula, median = size of $n+1/2$
- 4) Look at the cumulative frequency column and find that total which is either equal to $N+1/2$ or next higher than that and determine the value of the variable corresponding this, that gives median.

Example: Five coins were tossed 100 times and at each toss the number of heads was observed. The number of tosses during which 0, 1, 2, 3, 4,

5, heads were observed is shown in the following table. Find the median size of the heads obtained.

Table

Find the median size of the heads obtained

No. of Heads (k)	No. of tosses (frequency)
0	5
1	15
2	35
3	29
4	6
5	10
$N \text{ or } \Sigma f = 100$	

No. of Heads (k)	No. of tosses	Cumulative frequency
x	f	(Σf)
0	5	5
1	15	20
2	35	55
3	29	84
4	6	90
5	10	100

$M = \text{Size of } (N+1/2)\text{th item i.e. size of } (100+1/2)\text{th item.}$

$= \text{Size of } 50.5\text{th item.}$

The value of item from 50.5th to 55th is 2

The value of 50.5 item is 2.

Calculation of Median (Continuous series)

Steps:

- 1) Convert individual frequencies into cumulative frequencies.
- 2) Locate the median class by moving along the column of cumulative frequencies until you reach at the cumulative frequency which is equal to or just greater than half of the total i.e. $(N/2)$
- 3) Compute the median from the median class by the formula.

$$m = L_1 + \frac{(N/2 - Cf)}{f} (L_2 - L_1)$$

Where L_1 = Lower limit of the class interval where median is likely to fall.

$N/2$ = Item whose value is to be found

CF = Cumulative frequency up to the median class.

f = Frequency of the median class.

Example: A marketing executive wants to know the average age group which prefers their new branded cosmetics. The details of a survey of 200 buyers is given below.

Age groups	No. of Customers	CF
00 - 25	25	25
25 - 35	119	144
35 - 45	39	183
45 - above	17	200
$N = 200$		

Let us find the positional average. As $200/2 = 100$ falls in 25-35 class interval it can be considered to be the median class. To know the exact age:

$$\text{Median} = L_1 + \frac{(N/2 - CF)}{F} \times (L_2 - L_1)$$

Where,

$$L_1 = 25; L_2 = 35; N = 200; Cf = 65; F = 79$$

$$= 25 + \frac{(200/2 - 65)}{79} \times (35 - 25)$$

$$= 25 + \frac{100 - 65}{79} \times 10 = 25 + 4.431 = 29.431$$

The data shows that customers in the age group of 25–35 are liking the brand introduced by the unit.

Mathematical Properties of Median: The sum of the deviations of the items from median ignoring signs is the least. For example the median of 4 6 8 9 12 is 8, (ignoring the signs) the deviation from 8 totals to $4 + 2 + 0 + 1 + 3 = 10$. Deviation taken from any other value will be more than 10; like, if we take 6 and take the deviation from the deviation totals to $2 + 0 + 2 + 3 + 6 = 13$.

Quartiles and Decies

Besides median there are other measures which divide a series into equal parts. Important amongst them are quartiles and deciles. Quartiles divide the total frequencies in to 4 equal parts and deciles divide the total frequencies into 10 equal parts. As one point divides a series into two equal parts to divide the series in to 4 equal parts there are only 3 quartiles and into 10 equal parts there are 9 deciles. Quartiles are denoted by Q and deciles by D. In economics and business statistics quartiles are used more than deciles. Among the 3 quartiles the lower quartile Q₁ divides the series into two parts where 1/4th of the values are less than it and 3/4th more than the same. Q₂ divides the series in to two equal parts. Thus it is nothing but the median the upper quartile Q₄ similarly divides the series in such a way 1/4th of the values remain above this and 3/4th rest are below.

Computation of Quartiles and deciles: The method of computing deciles and percentiles are same as median. While computing these values in

individual and discrete series we add 1 to N whereas in continuous series we do not add 1.

Thus Q_1 = Size of $(N+1)/4$ th item (individual observation and discrete series)

Q_1 = Size of $(N/4)$ th item (in continuous series)

Q_3 = Size of $(3*N/4)$ th item (individual observation and discrete series)

Q_3 = Size of $(3*N/4)$ th item (in continuous series)

Q_4 = Size of $(4*(N+1)/10)$ th item (in individual and discrete series).

D_4 = Size of $(4*N/10)$ th item (in continuous series)

Example: Calculate the lower and upper quartile and third decile from the following data:

Central Value	2.5	7.5	12.5	17.5	22.5
Frequency	9	18	23	27	23

Since we are given the mid points we will first find out the upper and lower limits of the various class. The method for finding these limits is to take the difference between the two central values and divide it by 2. Deduct the value so obtained from the lower limit and add to the upper limit. In the given case $(7.5 - 2.5/2) = (5/2) = 2.5$. Thus the first class would be 0-5 record 5-10 etc.

Calculation of Q_1 , Q_3 Dy

Class group	F	CF
00 - 05	9	9
05 - 10	18	27
10 - 15	23	50
15 - 20	27	77
20 - 25	23	100

N = 100

Lower quartile Q_1 = Size of $(N/4)$ th item $(100/4) = 25$ th item.

Q_1 lies in 5-10 class.

$$= Q_1 + \frac{(N/4 - CF)}{f} \times L_2 - L_1$$

Where

$$L_1 = 5; L_2 = 10; (N/4) = (100/4) = 25; CF = 27; f = 18$$

$$= Q_1 = 5 + \frac{(25 - 27)}{18} \times 10 - 5$$

$$= 5 + 18/8 \times 5 = 5 + 5 = 10$$

D_4 = Size of $(4 \cdot N/10)$ th item (in continuous series)

Example: Calculate the lower and upper quartile and third decile from the following data.

Central value	2.5	7.5	12.5	17.5	22.5
Frequency	9	18	23	27	23

Since we are given the mid points we will first find out the upper and lower limits of the various class. The method for finding these limits is to take the difference between the two central values and divide it by 2. Deduct the value so obtained from the lower limit and add to the upper limit. In the given case $(7.5 - 2.5/2) = (5/2) = 2.5$. Thus the first class would be 0-5 record 5-10 etc.

Calculation of Q_1 , WQ_3 Dy

Class group	f	CF
00-05	9	9
05-10	18	27
10-15	23	50
15-20	27	77
20-25	23	100
$N = 100$		

Lower quartile Q_1 = Size of $(N/4)$ th item $(100/4) = 25$ th item. Q_1 lies in 5-10 class.

$$Q_1 = L_1 + \frac{((N/4) - CF)}{f} \times L_2 - L_1$$

$$L_1 = 5 \quad L_2 = 10 \quad (N/4) = (100/4) = 25, \quad CF = 27 \quad f = 18$$

$$Q_1 = \frac{(5 + 25 - 7)}{18} \times 10 - 5$$

$$= 5 + 18/18 \times 5 = 5+5 = 10$$

Upper quartile Q_3 = Size of $(3 \times N)/4$ th item $((3 \times 100)/4) = 75$ th item
 Q_3 lies in 15-20 class

$$Q_3 = L_1 + \frac{(3 \times N/4) - CF)}{f} \times L_2 - L_1$$

$$L_1 = 15 \quad L_2 = 20 \quad N = 100, \quad CF = 50 \quad f = 27$$

$$Q_3 = 15 + \frac{(3 \times (100 - 50))}{27} \times 20 - 15$$

$$= 15 + (75-50)/27 \times 5 = 15 + .926 \times 5$$

$$= 15 + 4.63 = 19.63$$

Third decile D_3 = Size of $(3 \times N)/10$ th item $(3 \times 100)/10 = 30$ th item
 lies in the class 10-15.

$$D_3 = L_1 + \frac{((3 \times N)/10 - CF)}{f} \times (L_2 - L_1)$$

$$\text{Where } L_1 = 10 \quad L_2 = 15 \quad (3 \times N)/10 = 30 \quad CF = 27 \quad f = 23$$

$$D_3 = 10 + (30-27)/23 \times 15-10$$

$$= 10 + (3/23) \times 5 = 10 + .652 = 10.652$$

Determination of Median Quartiles etc. Graphically - Median can be determined graphically by applying any of the following two methods:

1. Draw two O give - one by less than and one by more than method (the curve depicting cumulative frequency group is popularly known as give) O. From the point where both these curves intersect each other, draw a perpendicular line on the X axis. The point where this perpendicular line touches X axis gives the values of median. To draw this graph, the class intervals are considered on X axis and cumulative frequency corresponding to different classes is noted on vertical axis.

2. Draw only one O give by less than method. Take the variable on the X-axis and frequency in the Y axis. Determine the median value by the formula $\text{median} = \text{size of } (N/2)\text{th item}$. Locate this value on the X-axis and from it draw a perpendicular line on X axis and the point where it meets the X axis is the median. The other partition values like quartiles, deciles etc. can also be determined graphically by following method No.2.

Example: Determine the median wage graphically from the data give below.

Wages (Rs.)	No. of Workers
20 - 40	4
40 - 60	6
60 - 80	10
80 - 100	16
100 - 120	14
120 - 140	7
140 - 160	3
	60

1st Method: Draw two cumulative frequency curves one by less than method and another by more than method. In the less than' ogive the upper limit of the first class interval (20-40) would be the starting point and in 'more than' ogive the lower limit of the class interval or 20 would be the first value of the variable. From the point of intersection of the two

ogives drawn, draw a perpendicular line on X-axis and the point where it touches X-axis would be the value of median.

Wages Less than	No. of Workers	Wages More than	No. of Workers
40	4	20	60
60	10	40	56
80	20	60	50
100	36	80	40
120	50	100	24
140	57	120	10
160	60	140	3
		160	3

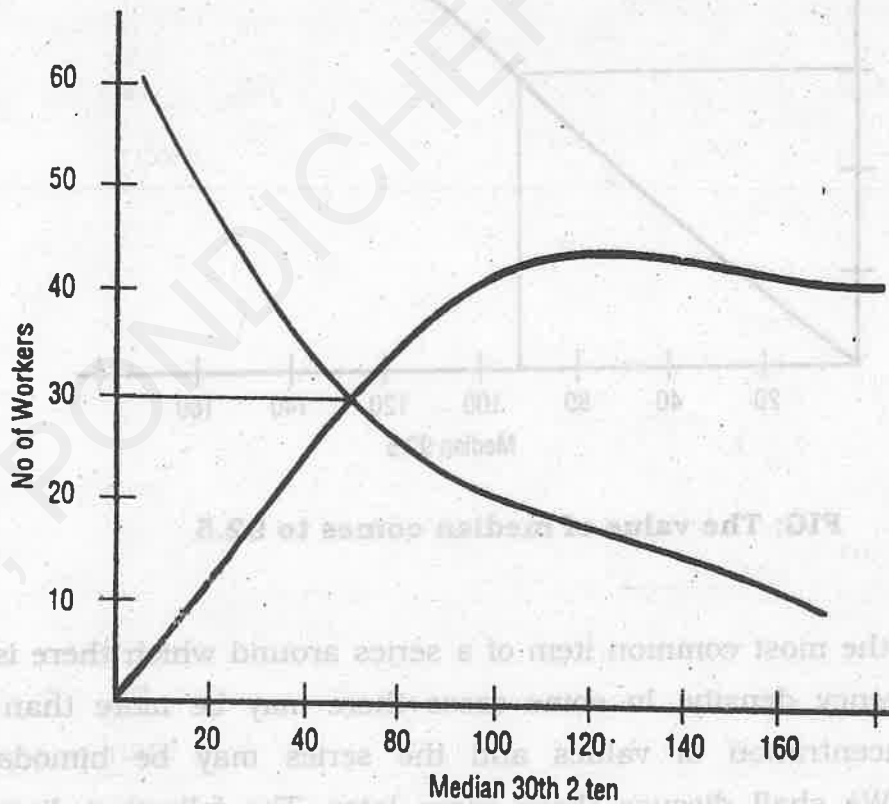


FIG: Locating median graphically

The value of median is 92.5

2nd method: If we draw only one ogive by the 'less than' method we can also determine the value of median from it.

This is shown in the following graph.

Median = Size of $(60/2) = 30$ th item.

Take 30 on the Y axis and draw a perpendicular line the ogive. From the point where it meets the ogive draw another perpendicular line on the x axis.

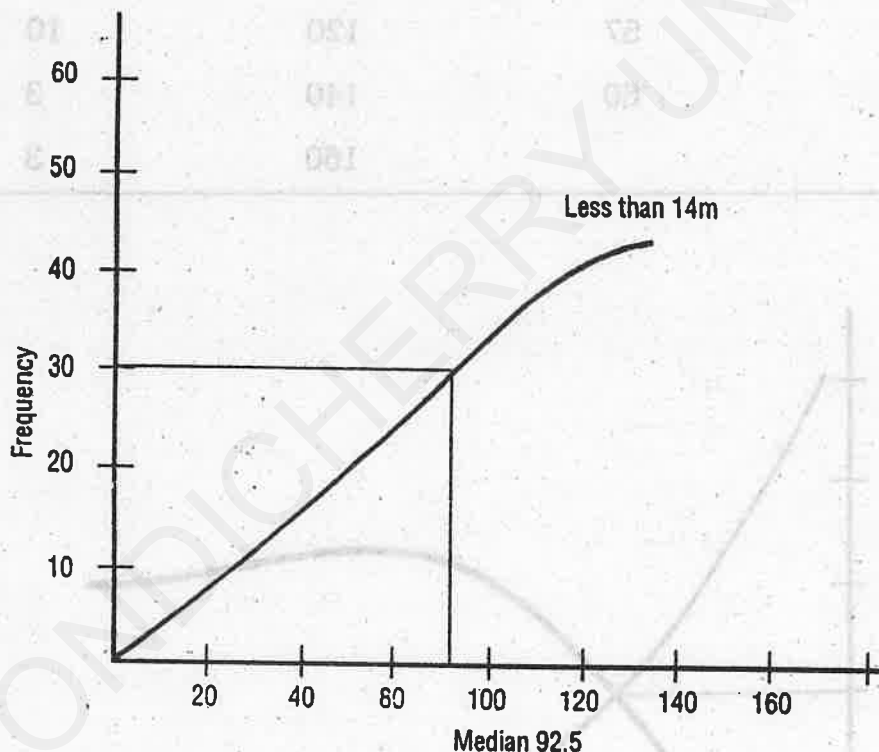


FIG: The value of median comes to 92.5

MODE

Mode is the most common item of a series around which there is the highest frequency density. In some cases there may be more than one point of concentration of values and the series may be bimodal or multimodal. We shall discuss these cases later. The following diagrams show the modal value.

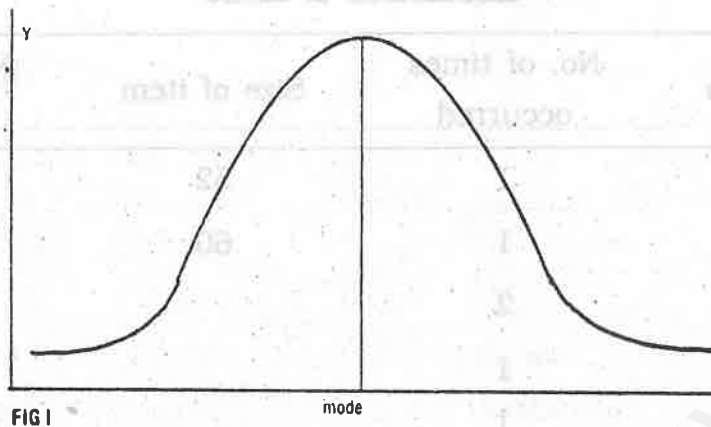


FIG I

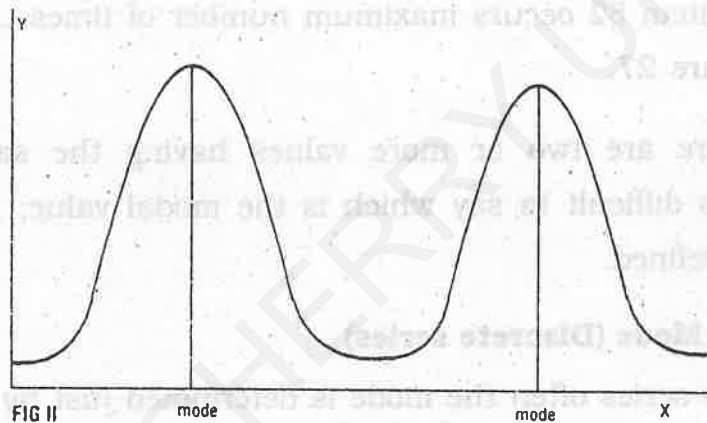


FIG II

Thus the value of the variable at which the curve reaches a maximum is called mode. When the series has two peaks there are two modes and such series are named bimodal series. The diagrams Fig I and II depict unimodal series.

Calculation of mode - (Individual observation) For determining mode, count the number of times the various values repeat themselves and the value occurring maximum number of times in the modal value. The more often the modal value appears relatively, the more valuable the measure is an average to represent data.

Example: The following data gives the marks of 10 students.

25, 42, 39, 45, 50, 52, 60, 52, 42, 52.

Calculation of Mode

Size of item	No. of times occurred	Size of item	No. of times occurred
25	1	52	3
39	1	60	1
42	2		
45	1		
50	1		
			Total 10

Since the item 52 occurs maximum number of times i.e. 3 hence the modal marks are 27.

When there are two or more values having the same maximum frequency, it is difficult to say which is the modal value; hence mode is said to be ill defined.

Calculation of Mode (Discrete series)

In discrete series often the mode is determined just by inspection i.e. by looking to that value of the variable round which maximum of the items are concentrated. For example observe the following data.

Size of the shoe	5.5	6	6.5	7	7.5	8	8.5	9
Frequency of sale	12	17	25	60	90	45	20	12
					↑	mode size		

From the above data we can clearly say that the modal size is 7.5.

In many cases it is not possible to locate the modal class by mere inspection and also where the mode is determined most by inspection, an error of judgment is possible where the difference between the maximum frequency and the frequency preceding it or succeeding it is very small and the items are concentrated on either side. A grouping table and an analysis of the grouping table is to be done in such case.

A grouping table has six columns; in first column the highest frequency is marked. In second column frequencies are grouped into two's, in third column leave first frequency then group the remaining in two's, in fourth column group frequencies in three's. In fifth column leave first frequency and group remaining in three's and in sixth column leave first two frequencies and then group the remaining in three. In each of these cases take the maximum total and mark it in a circle or mark the maximum total by bold type.

After preparing the grouping table, the analysis table is prepared. While preparing the analysis table the column number is put on the left hand side and various probable values of mode on the right hand side. The highest frequencies which are marked by bold type in grouping table are entered by bars in the analysis table corresponding to the value they represent.

The procedure of preparing grouping table and analysis table shall be clear from the following example. Take a simple case.

Size of shoe	1	2	3	4	5	6	7	8	9	10
Frequency	10	5	13	6	23	32	14	35	9	8

By inspection it can be said that the modal size is 8 as it has got the maximum frequency. But this test is not fool proof. It is not only the frequency of a class but also the frequency of the neighbouring class that decides the modal class. Thus by preparing grouping table we would be able to find out the actual modal size.

Grouping Table

Size of the shoe	frequency					
	1	2	3	4	5	6
1	10					
2	5					
		15				
3	13	19	18			
4	6			28	24	
			29			42
5	23			61	69	
6	32					
		55				
7	14		46			81
8	35	49				
			44	58	52	
9	9	17				
10	8					

Analysis Table

Col. No.	Size of the shoe									
	1	2	3	4	5	6	7	8	9	10
1								1		
2					1	1				
3						1	1			
4				1	1	1				
5					1	1	1			
6						1	1	1		
				1	3	5	3	2		

Here 4 occurs once, 5 occurs three times, 6 occurs 5 times, 7 occurs three times and 8 occurs twice only. Thus 6 occurs maximum number of times. Thus 6 is the modal size. It is to be noted here how there was an

error of judgment by determining mode by inspection. By inspection we had stated mode is 8 whereas the size 6 occurs maximum number of times.

Calculation of Mode - (Continuous series)

In continuous series the determination of mode involves two steps. First by process of grouping the modal class will be located. After this the mode is interpolated by use of a formula.

Steps:

- 1) By preparing grouping table and analysis table or by inspection determine the modal class.
- 2) Determine the value of mode by applying the following formula

$$Mo = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

Where L = Lower limit of the modal class

Δ_1 = The difference between the frequency of modal and pre modal i.e. preceding class ignoring sign.

Δ_2 = The difference between frequency of modal and post modal i.e. succeeding class ignoring sign.

i = Class interval of the modal class.

While applying the above formula for calculating mode it is necessary to see that the class intervals are uniform throughout. If they are unequal they should be made equal on the assumption that the frequencies are equally distributed throughout the class, otherwise it may lead to misleading results. In a bimodal distribution also mode cannot be determined by the procedure discussed above. When the series is ill defined the mode can be ascertained by the following formula based upon the relationship between mean, median, mode.

$$\text{mode} = 3 \text{ median} - 2 \text{ mean}$$

Example: The following data refer to the weekly wages of 460 workers of a factory.

Wages (Rs.)	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of Workers	85	120	110	67	50	20	6

Calculation of Mode

Grouping table							
Wages	Frequency						
	1	2	3	4	5	6	
20-30	85	205	230	315	297	227	
30-40	120						
40-50	110	177	177	137	79		
50-60	67						
60-70	50	70	26				
70-80	20						
80-90	14						

Analysis Table

Col. No.	Wages with maximum frequency						
	20-30	30-40	40-50	50-60	60-70	70-80	80-90
1		1					
2	1	1					
3		1	1				
4	1	1	1				
5		1	1	1			
6			1	1	1		
No. of times the class has occurred	2	5	4	2	1		

From the above it is evident that class 30-40 is repeated **maximum** times. Thus mode lies in the class 30-40. Hence

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

$$L = 30 \quad \Delta_1 = 120 - 85 = 35 \quad \Delta_2 = 120 - 110 = 10 \quad i = 10$$

$$\text{Mode} = 30 + \frac{35}{35 + 10} \times 10 = 30 + .78 \times 10 = 37.8$$

Example: (for Calculation of mode when mode is ill defined) From the following data find the modal value.

Marks	No. of Students
40 - 50	7
50 - 60	8
60 - 70	10
70 - 80	6
80 - 90	13
90 - 100	10
100 - 110	12
110 - 120	7

In this case it is difficult to ascertain the mode by inspection. Thus we find out the modal class by the help of grouping table and analysis table.

Grouping table						
Marks	Frequency					
	1	2	3	4	5	6
40-50	7					
50-60	9	16	19			
60-70	10			26	25	29
70-80	6	16	19			
80-90	13			29	35	29
90-100	10	23	22			
100-110	12					
110-120	17	19				

Analysis Table

Col. No.	Classes in which mode is expected to lie						
	50-60	60-70	70-80	80-90	90-100	100-110	110-120
1				1			
2				1	1		
3					1	1	
4			1	1	1		
5				1	1	1	
6		1	1	1	1	1	1
Total		1	2	5	5	3	1

It is clear from the analysis table that this is a bimodal frequency distribution. Hence mode is determined by the formula

$$\text{Mode} = 3 \text{ median} - 2$$

For this purpose we have to find out median and mean, median size of $(N/2)$ th item = Size of $(74/2)$ i.e. 37th item. Hence median lies in the class 80-90 applying the formula

$$\text{Median} = 80 + \frac{37 - 32}{13} \times 10 = 83.84$$

Calculation of Mean

Marks	Frequency f	mid point x	dx ((x-85)/10)	fdx
40-50	7	45	-4	-28
50-60	9	55	-3	-27
60-70	10	65	-2	-20
70-80	6	75	-1	-6
80-90	13	85	0	0
90-100	10	95	1	10
100-110	12	105	2	24
110-120	7	115	3	21
$\Sigma f = 74$				$\Sigma f dx = -36$

$$\text{Mean } (\bar{x}) = (85 + (-36/74) \times 10) = 8.5 - 4.85 = 80.14$$

$$\text{Mode } (s) = (83.84) - 2(80.14) = 251.52 - 160.28 = 91.24$$

Determination of Mode (Graphic Method)

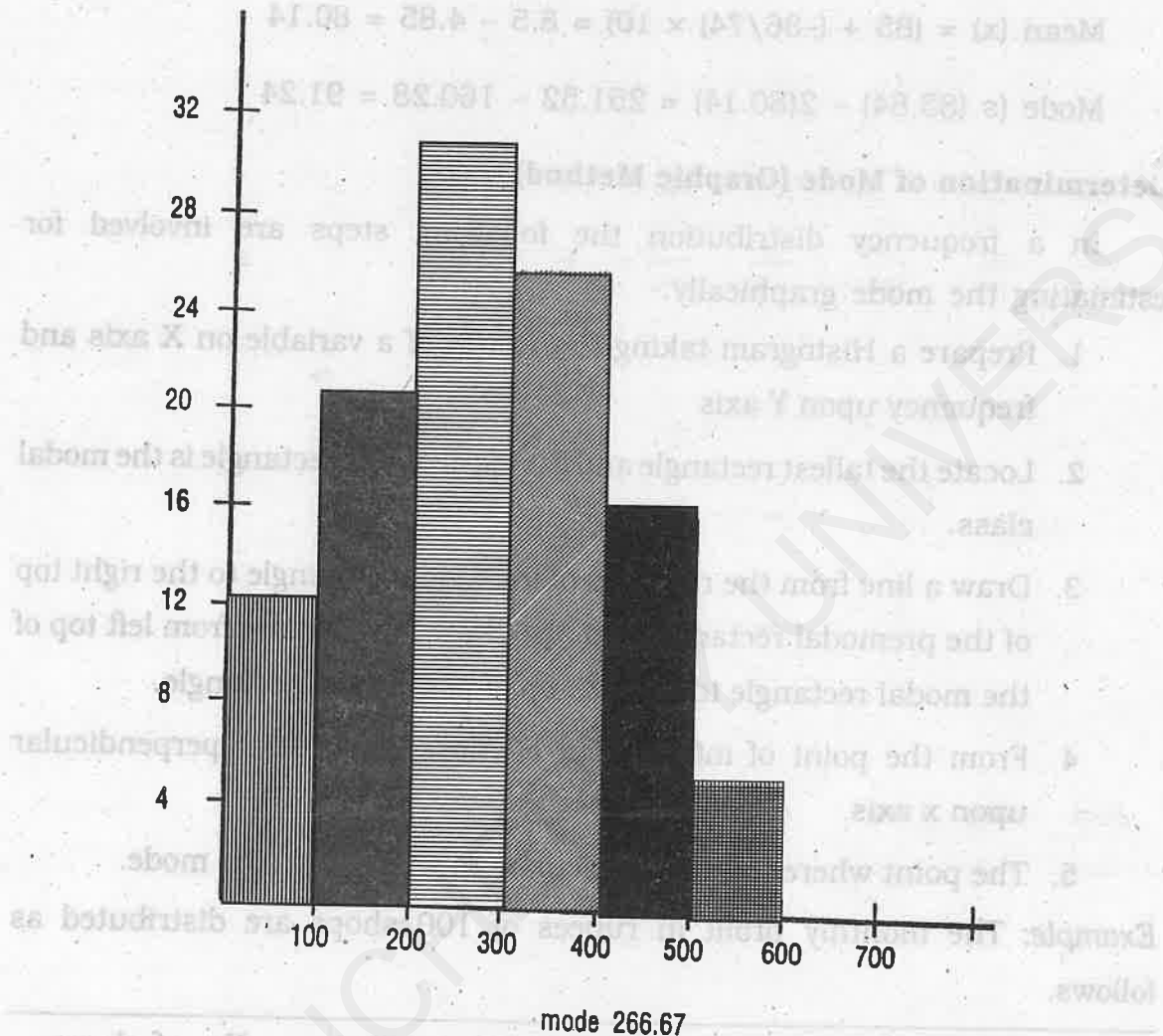
In a frequency distribution the following steps are involved for estimating the mode graphically.

1. Prepare a Histogram taking the values of a variable on X axis and frequency upon Y axis
2. Locate the tallest rectangle and the class of this rectangle is the modal class.
3. Draw a line from the right top of the modal rectangle to the right top of the premodal rectangle and also draw another line from left top of the modal rectangle to the left top of post modal rectangle.
4. From the point of intersection of these lines, draw perpendicular upon x axis.
5. The point where the perpendiculars meet x axis is the mode.

Example: The monthly profit in rupees of 100 shops are distributed as follows.

Profit per shop	No. of shops	Profit per shop	No. of shops
0-100	12	300-400	25
100-200	20	400-500	16
200-300	30	500-600	5

Calculate the average profit earned by maximum shops.



Direct calculation

Mode lies in the class 200.- 300

$$\begin{aligned}
 \text{Mode} &= L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i = 200 + \frac{30 - 20}{(30 - 20) + (30 - 25)} \times 100 \\
 &= 200 + \frac{10}{10 + 5} \times 100 = 200 + 66.67 = 266.67
 \end{aligned}$$

Properties of mode

1. The mode is the typical value; therefore it is often more representative.

2. The value of mode is independent of extreme value.
3. The true mode is difficult to compute but approximate mode is easily found.
4. It is positional average.

GEOMETRIC MEAN

Geometric mean is defined as the n th root of the product of n items. symbolically geometric mean of values $X_1, X_2, X_3 \dots X_n$ is given by the expression

$$G = \sqrt[N]{X_1 \times X_2 \times X_3 \dots X_N}$$

Calculation of Geometric mean

$$\text{Log } G_m = \frac{\text{Log } X_1 + \text{Log } X_2 + \text{Log } X_3 + \dots + \text{Log } X_N}{N}$$

$$\text{or } \text{Log } G_m = \frac{\sum \text{Log } X}{N} \quad \text{or } G_m = \frac{\text{Antilog } \sum \text{log } X}{N}$$

$$\text{In discrete series Geometric mean} = \frac{\text{Antilog } \sum f \text{log } X}{N}$$

$$\text{In continuous series } G_m = \frac{\text{Antilog } \sum f \text{log } X}{N}$$

Calculation of geometric mean (Individual values)

Steps:

- 1) Take log of various values of a variable
- 2) Sum up the log of various values of a variable $\sum \log X$.
- 3) Divide $\sum \log X$ by N .
- 4) Read and antilog of the value obtained in step 3 and in this is the geometric mean

$$G_m = \frac{\text{Antilog } \sum \text{log } X}{N}$$

Example: Find Geometric Mean of the following series:

135 79 286 176 7 59 38

Calculation of Geometric mean

Values (x)	Log X	Values (X)	Log X
135	2.1303	7	0.8451
79	1.8976	59	1.7709
286	2.4564	38	1.5798
		N = 7	$\Sigma \log X$ 12.9256

Geometric mean

$$G = \text{Antilog } \frac{\Sigma \log X}{N} \text{ where } \log X = 12.9257$$

$$N = 7$$

$$G = \text{Antilog } \frac{12.9256}{7} = \text{Antilog } 1.8465 = 70.23$$

Calculation of Geometric Mean (Discrete Series)

Steps:

- 1) Find the logarithms of the variable X
- 2) Multiply these logarithms with the respective frequencies and obtain the total $\Sigma f \log X$
- 3) Divide $\Sigma f \log X$ by total frequency
- 4) Take antilog of the value obtained in steps (3)

$$GM = \text{Antilog } \frac{\Sigma f \log x}{N}$$

Example: Find the geometric mean of the following distribution:

Value	8	25	17	30
Frequency	5	3	4	4

Calculation of geometric mean

X	f	Log X	flog
8	25	0.9031	4.5155
25	3	1.3979	4.1937
17	4	1.2304	4.9216
13	4	1.4771	1.9084
	16		15.5392

$$GM = \text{Antilog } \frac{\sum (f \log x)}{N} \quad \text{Where } \sum f \log X = 15.5392$$

$$N = 16$$

$$GM = \text{Antilog } \frac{15.5392}{16} = \text{Antilog } 9712 = 9.358$$

Calculation of Geometric Mean (Continuous series)

Steps:

- 1) Write the mid point of the Class X
- 2) Find the log value of mid point log X
- 3) Multiply the log values with their frequency and obtain the total $\sum f \log x$
- 4) Divide the total in step 3 by the total frequencies
- 5) Take antilog of value obtained in step 4

Example: Give the geometric mean of the data below

Marks	30-40	40-50	50-60	60-70
No. of Students	15	8	4	3

Marks	Frequency f	mid point x	dx ((x-85)/10)	fdx
30-40	15	35	1.5440	7.7200
40-50	8	45	1.6535	13.2256
50-60	4	55	1.7403	6.9612
60-70	3	65	1.8129	5.4387
	30			$\Sigma f \log m =$ 33.3455

$$GM = \text{Antilog} - \frac{\Sigma f \log m}{N} \text{ Where } \Sigma f \log m = 33.3455$$

$$N = 30$$

$$GM = \text{Antilog} = \frac{33.3455}{30} = \text{Antilog } 1.115$$

Hence geometric mean = 12.93

Properties of Geometric Mean

The following are two mathematical properties of geometric mean.

1. The product of the value of series will remain unchanged when the value of geometric mean is substituted for each individual value. For example the geometric mean for series 2,4,8 is 4 therefore we have $2 \times 4 \times 8 = 64 = 4 \times 4 \times 4$

2. The sum of the deviations of the logarithms of the original observations above or below the logarithms the geometric mean is equal. This also means the value of geometric mean is such as to balance the ratio deviations of the observations from it, i.e. the product of the ratios of geometric mean to the items below or equal to it is equal to the product of ratios above the geometric mean. For example geometric mean of number 2, 4, 8 as 4 than $(G/2) \times (G/4) = (8/G)$ where $G = 4$

$$\text{i.e. } (4/2) \times (4/4) = (8/4)$$

3. It is amenable to algebraic treatment i.e. we can find the geometric mean of two or more series. For example geometric mean of two series can be calculated as

$$G_{12} = \text{Antilog } \frac{N_1 \log G_1 + N_2 \log G_2}{N_1 + N_2}$$

4. Geometric mean is always smaller than arithmetic mean and greater than harmonic mean.

HARMONIC MEAN

Harmonic mean is based on the reciprocals of the numbers averaged. It is defined as the reciprocal of the arithmetic mean of the reciprocal of individual observation.

Computation of harmonic Mean - The harmonic mean can be expressed by the general formula

$$HM = \frac{N}{\frac{1}{X_1} + \frac{1}{X_2} + \frac{1}{X_3} + \dots + \frac{1}{X_N}}$$

However for specific type of series the formulas are

For individual observation $HM = \frac{N}{\sum (1/X)}$

For discrete series $HM = \frac{N}{\sum fX \frac{1}{x}} = \frac{N}{\sum f/x}$

For continuous series $HM = \frac{N}{\sum (fX1/m)} = \frac{N}{\sum (f/m)}$

Calculation of Harmonic Mean - (Individual observation)

Steps:

- 1) Find the reciprocals of the various values X
- 2) Sum the reciprocals of the values

3 Divide number of values by the sum of reciprocal values

$$\text{Formula, } HM = \frac{N}{\sum (1/x)}$$

Example: Find the harmonic mean of the following values.

35 45 75 89 68 110 135 87 93 120

Calculation of Harmonic Mean

Value X	Reciprocal 1/x
35	.0286
45	.0222
75	.0133
89	.0112
68	.0147
110	.0091
135	.0075
87	.0115
93	.0108
120	.0083
N = 10	$\sum(1/x) = .1371$

$$\text{Harmonic Mean} = \frac{N}{\sum (1/x)} = \frac{10}{.1371} = 72.94$$

Calculation of Harmonic Mean (Discrete series)

Steps:

- 1) Find the reciprocals of the various values of X, (1/X)
- 2) Multiply the reciprocals of the values with their respective frequency f(1/x)
- 3) Total the product obtained in step two
- 4) Put the values in the formula given below

$$HM = \frac{N}{\sum f(1/x)}$$

Example: From the following data compute the Harmonic mean.

Marks	10	20	25	40	50	60
No. of Students	30	40	45	31	10	5

Calculation of Harmonic Mean

Marks	No. of Students	F/x
10	30	3.000
20	40	2.000
25	45	1.800
40	31	.775
50	10	.200
60	5	.067
N = 160		$\Sigma(f/x) = 7.842$

$$HM = \frac{N}{\Sigma(f/m)} = \frac{160}{7.842} = 20.402$$

Calculation of Harmonic Mean - (Continuous Series)

Steps:

- 1) Write the mid points of all classes (m)
- 2) Find the reciprocals of mid points
- 3) Multiply the reciprocals of each mid point by respective class frequency $f(1/m)$
- 4) Sum the products obtained in step 3
- 5) Put the values in the formula $HM = \frac{N}{\Sigma(f/m)}$

Example: From the following data compute the value of harmonic mean

Class interval	10-20	20-30	30-40	40-50	50-60
Frequencies	7	9	15	11	4

Calculation of Harmonic Mean

Class interval	Mid point (m)	Frequency (f)	f/m
10-20	15	7	.467
20-30	25	9	.360
30-40	35	15	.429
40-50	45	11	.244
50-60	55	4	.073
		N = 46	$\Sigma(f/m) = 1.573$

$$= HM = \frac{N}{\Sigma f/m} = \frac{46}{1.573} = 29.24$$

Specific use of types of averages

1. *Weighted arithmetic mean* - the weighted arithmetic mean is used specifically in the problems relating to

a). Construction of index number

b) Standardised birth and death rates.

2. *Median* - The median is useful for distributions containing open end intervals and also since the median is affected by the number rather than the size of items it is frequently used as a measure of central tendency instead of mean in cases where such values are likely to distort the mean.

3. *Mode* is employed when the most typical value of the distribution is desired. It is most meaningful measure when the series is skewed or not normally distributed.

4. *Quartiles and Deciles* - quartiles are more widely used in economics and business problems to find out the positional averages. Deciles however are important in psychological and educational statistics concerning grade ratings and ranks.

5. *Geometric mean* - Geometric mean is specially useful in the following cases:

a) It is used to find the average percent increase in sales, production population or other business or economic series. For example for 1991 to 1993 the prices increased by 5% 10% and 18% respectively. The average annual increase is not 11% as calculated by arithmetic average $((5+10+18)/3)=11$ but 10.9% as obtained by the geometric mean.

b) Geometric mean is considered to be best average in construction of index numbers.

c) It is specifically more suitable when large weights have to be given to small items and small weights to large items.

6. *Harmonic mean* - The field of usefulness of harmonic mean is very much restricted. It is useful in cases like (a) Computing average rate of increase in profit of a concern or average speed at which the journey has been performed or average price at which the article has been sold. The rate usually indicates the relation between two different types of measuring units that can be expressed reciprocally, like, a man walked 10 km in 2 hrs the rate of his walking speed can be expressed $(10 \text{ km}/2 \text{ hrs}) = 5 \text{ km per hour}$.

Advantages and Disadvantages of Different Type of Averages -

MEAN

Advantages: Arithmetic mean is most widely used in business and economics because of the following reasons:

1. It is the simplest average to understand and easiest to calculate.
2. It is affected by the value of every item in the series. As it is based on all the observations in a series and cannot be computed even if a single value is missing, it is the most representative measure of central tendency.
3. It is defined rigidly by a mathematical formula, therefore anybody calculating the mean gets the same answer from the same set of data. Here the bias of an investigator cannot do any harm.

4. As it is a computed average it can be used for further statistical computation so that its utility is enhanced. In other words it lends itself to subsequent algebraic treatment.
5. It is relatively simple to calculate. It can be calculated even if individual items are not known but the sum of their number is known.
6. It is relatively reliable in the sense that it is affected least by fluctuations of sampling.
7. It gives weight to all items in direct prorate to their size which is not done by other averages.
8. Mean is in typical sense, the centre of gravity balancing the values on either side of it, thus the total of deviations from arithmetic mean is zero.
9. It is a calculated value but not based on position or direction of items in the series. Thus we might almost say that in case of doubt, the arithmetic mean should be used and other average should be used only when there is some clear reason.

Disadvantage: The main weaknesses of mean are

1. The extreme items in the series unduly affect the average as this average depends on each and every item in the series. For example the mean income of five persons is Rs. 1700, Rs. 150, Rs. 80, Rs. 70 is 500. Thus the mean is affected by the high figure Rs. 1700. The smaller the number of observations the greater is the impact of extreme value in this average.
2. The mean cannot be computed when the frequency distribution has open end class at both ends.
3. Though the calculation of mean is very simple, it can hardly be located by inspection but median and mode can be. It can ignore any single item only at the cost of losing its accuracy.

4. It may not correspond to any observed value in a series. For example the mean of 4, 8, 9, 10, is 7.78 but no observation is 7.78 in the series.

5. In many cases arithmetic mean gives meaningless results. For example the average number of children born per family may work out to be 2.7. This figure makes no sense as children cannot be in fractions.

6. In many situations mean gives entirely misleading results. It is possible that mean of different series may have wide difference in the individual values of the series. For example mean of the following 3 series are same but the series are quite different:

	A	B	C
	10	30	12
	10	6	10
	10	3	9
	10	1	9
Total	40	40	40
Mean	10	10	10

Mean is representative of series A and C but totally unrepresentative of B series.

7. Arithmetic mean provides a characteristic value in the series indicating where most of values lie when the distribution of variables are normal. In case of U shaped distribution the mean is not likely to serve any useful purpose.

MEDIAN

Advantages: The Median possesses the following merits:

1. It is easily understood and easily calculated especially in the case of individual values on discrete series.
2. Median is especially useful in case of open end classes since it is enough to know the position and not the value of item to find out this average.

3. Median is recommended for distributions having unequal classes since it is easier to compute than mean.
4. The value of median cannot be distorted by a few items exceptionally large or small as they are not taken into consideration, except for arranging the data in order of magnitude.
5. It is the most appropriate average in dealing with qualitative data i.e. when ranks are given or other type of items that are not counted or measured but are scored.
6. The median as a measure of central tendency is mostly used in markedly skewed distributions such as income distribution.
7. The value of median can be determined graphically but not the value of mean.
8. Median possesses the simple property of being the central or the middle most value. This clear cut meaning makes median easily understood. As it is the middle most value, it is easy to determine and sometimes it can be done by mere inspection.

Disadvantages: Despite the above merits median suffers from the following weakness:

1. Since median is a positional average, its value is not determined by each and every observation.
2. The median is not amenable to algebraic treatment like mean.
3. For calculation of median, it is necessary to arrange the data. In case of large data the arrangement of items in order of magnitude is a tedious job. This is a time consuming process also.
4. It is erratic in case of small number of items.
5. The value of median is affected more by sampling fluctuation than the value of arithmetic mean.
6. The assumption that all values of a variable are uniformly distributed in median class in a continuous series may not hold good.
7. If big or small items in a series are to receive greater importance median would be an unsuitable average.

MODE

Advantage: The following are the main advantages of mode:

1. By definition mode is the most typical or representative value of a distribution. Thus to the ordinary mind it seems to be best representative of the group.
2. It is simple to understand, easy to calculate and in most cases it can be located by mere inspection.
3. It is not affected by extremely large or small items. For example the mode of values 1, 5, 5, 1000 is 5 and mode of values 1, 5, 5, 10,000 also is 5.
4. The value of mode can be determined in open end distribution without knowing the class limits.
5. It can be very easily determined from graph.
6. It is not affected by sampling fluctuations.
7. Since mode is the most common item of a series it is not an isolated example like mean. Unlike arithmetic average it cannot be a value which is not found in the series.
8. It can be used to describe qualitative phenomenon For example if we want to study the consumer preference for products, mode would be the most suitable tool. Thus mode can be profitably employed in market research.
9. It is not necessary to know all the items in a series to compute mode. The matter to be known is the point of maximum concentration. Thus extreme items in the series gets ignored as these items have less concentration.

Disadvantages: Mode suffers mainly from the following weaknesses:

1. The value of mode cannot always be determined. In some case we may have bimodal or multimodal series. Thus mode has little meaning unless there is large number of items and majority of them show a marked concentration towards a particular value.

2. It is not capable of algebraic manipulation or treatment. For example from the modes of two set of data one cannot calculate the overall mode of combined data.
3. It is not based on all observations of a series of data.
4. In the computations of mode, groups of data in a frequency distribution is a prerequisite.
5. In some series there may not be a modal value, i.e. if the data contains values that does not occur more than once.
6. It is not rigidly defined. There are several methods of calculating mode all of which usually give somewhat different answers. Thus mode is regarded as the most unstable average and its value is difficult to determine.
7. Calculating mode is a laborious and time consuming process as frequencies are to be grouped to ascertain the concentration of frequencies.
8. While dealing with quantitative data the disadvantages of mode outweigh its advantages and hence is seldom used.

GEOMETRIC MEAN

Advantages: The advantages of geometric mean are as follows:

1. Geometric mean is a calculated value and it depends upon the size of all the items.
2. It is based on all the observations and also gives less importance to extreme items or values than does the arithmetic mean. Thus it is more typical than arithmetic mean.
3. It is amenable to algebraic treatment in the sense that we can find out the combined geometric mean of two or more series.
4. It is useful in average ratios and percentages and in determining the rate of increase or decrease it is particularly adopted.
5. Geometric mean is rigidly defined and its value is a precise figure.

Disadvantages:

1. It is not easy to understand and comparatively difficult to calculate as it involves complex mathematical calculations like roots, logs, antilogs etc.
2. If any value in a series is zero the geometric mean also becomes zero. In such cases geometric mean cannot be calculated similarly if a value is negative the geometric mean becomes an imaginary figure.
3. Like mean it cannot be computed in case the distribution has open end class.
4. The value of geometric mean may not correspond to any value in the series as the case with arithmetic mean.
5. The property of giving more weight to smaller items may in some case not be desirable which proves to be a drawback of geometric mean. In cases where smaller items have to be given smaller weight and bigger items bigger weights, geometric mean is not an ideal average.
6. It is not a widely used average.
7. Because of abstract mathematical character geometric mean is not easy to calculate for a non mathematical person and also to understand and interpret.

HARMONIC MEAN**Advantages:**

1. Harmonic mean is rigidly defined and its value is definite.
2. It tends itself to algebraic manipulation.
3. It is based on all the observation of series.
4. It is not affected by sampling fluctuation.
5. It measures relative changes and is primarily used in averaging rates.
6. Like geometric mean this average is also not affected much by sampling fluctuation.
7. It gives greater importance to small items and as such a single big item cannot push up its value.

Disadvantages: Harmonic mean suffers from the defects like

1. It is difficult to compute and is not understandable to common man.
2. It gives more weight to smaller items which may not be desirable in many cases. As such this average is not very useful for analysis of economic data.
3. Harmonic mean cannot be obtained if any one of the observation is zero.
4. Generally it is not a good representative of a statistical series unless the phenomenon is such that it needs smaller items to be magnified.

RELATIONSHIP AMONG AVERAGES

There exists a definite relationship among three principal measures of central tendency namely mean, median and mode. Depending on the degree of variability in the value of data the size of difference among averages measure. In case the frequency distribution is perfectly symmetrical, the values of mean, median and mode are identical i.e. $\text{mean} = \text{median} = \text{mode}$.

As the frequency distribution departs from symmetry these values differ. However they still maintain a definite relation to each other. A distribution may be asymmetrical either to the left or to the right.

- a) When the distribution is skewed to the left median is less than mode and mean is less than median.

i.e. $\text{Mean} < \text{Median} < \text{Mode}$.

- b) When the distribution is skewed to the right the mean is greater than median and mode. Their relationship may be put as

$\text{Mean} > \text{Median} > \text{Mode}$.

It is evident from the above that median always lies between mean and mode. When mode is minimum the distribution is skewed to the right and when it is maximum the distribution is skewed towards left. For a moderately skewed distribution Karl Pearson's has given the relationship among 3 principal measures of central tendency as given below.

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

Mode = 3 median - 2 mean and median = mode + $\frac{2}{3}$ (mean-mode)

The relationship is based on the fact that distance between mean and median is half of the distance between mode and median. The figures below show the relative position of mean, median and mode for frequency distributions which are moderately asymmetrical.

LINE GRAPHS

Further in any distribution there exist a particular relationship between Arithmetic mean, geometric mean and harmonic mean when original items differ in size in any distribution: AM GM and HM will differ and will be in the following order. M, GM, HM i.e. arithmetic mean is greater than geometric mean and geometric mean is greater than harmonic mean. When all the numbers are equal, the equality sign holds good.

UNIT - IV**LESSON - 1**

MEASURES OF DISPERSION

- ❑ INTRODUCTION
- ❑ DEFINITION
- ❑ OBJECTIVE AND SIGNIFICANCE
- ❑ TYPES OF DISPERSION
- ❑ MEASURES OF DISPERSION
- ❑ PROPERTIES OF GOOD MEASURE OF DISPERSION
- ❑ METHODS OF STUDYING DISPERSION, THEIR MERITS
DEMERITS AND USES
- ❑ RANGE
- ❑ INTERQUARTILE RANGE AND QUARTILE DEVIATION
- ❑ MEAN DEVIATION
- ❑ STANDARD DEVIATION

INTRODUCTION

Averages or measures of central tendency viz. mean, median, mode identify a single figure from an array of data which gives us an idea of the concentration of observations around central value of the distribution. However these measures do not tell us how the individual items have been distributed around the value and whether the deviation of individual items from the average are large or small. In addition, there may be several series whose averages may be identical but may be differing from each other in many ways. In such cases further statistical analysis of data is required so that the difference between the series may also be studied and accounted for. If this is done statistical analysis would be more meaningful and accurate. An average can be more meaningful only when it is examined in the light of deviation of individual item from the same. Measurement of dispersion helps us in a studying this important characteristic of the

frequency distribution. To illustrate the point let us consider the following example. Details of sales of three retail shops during a week is given below:

Daily Sales (Rs.) (A)	Daily Sales (Rs.) (B)	Daily Sales (Rs.) (C)
4000	7000	4000
4000	3000	4500
4000	2000	4000
4000	2500	5000
4000	3750	3000
4000	5000	4000
4000	3750	3750
$\bar{X} = 4000$	$\bar{X} = 4000$	$\bar{X} = 4000$

In this example all the three shops show an average daily sale of 4000. But if we carefully observe the daily sales figures we observe shop A has uniform sales everyday whereas in B and C there are fluctuations. The fluctuation in B, in addition, is more than C. Average as a measure to present the real characteristic of sales performance seems insufficient, as average for all the shops are same. Thus average to be meaningful is to be supplemented by measure of dispersion, which indicates the extent to which the individual of a data fall away from the average or central value. The measure of dispersion helps us to identify the homogeneity (Compactness) or heterogeneity of the data in a distribution. In this section we shall be discussing the measures of variability or spread or dispersion with regards to its degree only. We shall discuss the direction of the spread or variability in the next lesson of this unit.

DEFINITION

Some important definitions of dispersion are given below:

1. "Dispersion or spread is the degree of the scatter or variation of the variables about a central value" – Books and Disk.
2. "Dispersion is a measure of variation of items" - A.L. Bowley.

3. "Measures of variability are usually used to indicate how tightly bunched the sample values are around the mean" - Dyckman and Thomas.
4. "The degree to which numerical data tend to spread about an average value is called the variation or dispersion of the data" - Spiegel.

From the above definitions it is clear that in general sense the term dispersion refers to the variability in the size of items from the average size. The term dispersion not only gives a general impression about the variability of a series, but also a precise measure of such variation. In a precise study of dispersion, the deviation of items from the measure of central tendency are found out and then deviations are averaged to give a single figures representing the dispersion of the series. This figure can be compared with similar figures representing other series. Obviously such comparisons would give a better idea about the series than a simple comparison of their averages. Since measures of dispersion give an average of the difference of various items from an average, they are called averages of second order. However mean, median, mode, harmonic mean, geometric mean are called average of first order. Since in the calculation of measures of dispersion we average values derived by the use of the averages of first order, these measures are called averages of second order.

OBJECTIVE AND SIGNIFICANCE OF DISPERSION

Measures of dispersion or variation are calculated to serve the following four basic purpose:

1. To judge the reliability of an average.
2. To make a comparison of two or more series with regard to their variability.
3. To identify the cause of variability which will serve as a basis for control of variability.
4. To facilitate further statistical analysis.

A brief explanation of these are given below:

1. The measures of dispersion are employed to describe the structure of a frequency distribution and to indicate the reliability of an average as a representative of a distribution. When dispersion is small, the average would be representative value, on the other hand, when dispersion is large the average is not so typical and unless the sample is very large, the average may be quite unreliable.
2. The measures of dispersion are extremely useful in comparing two or more series with regard to their variability. The higher the degree of variability the lesser is the consistency or uniformity in the values of the variables. Whereas a low degree of variation would mean great uniformity or consistency. For example two cricket players A and B have an average batting score of 40 in five innings.

A's Score	12	130	50	4	4
B's Score	36	42	36	40	44

When we compare these scores in five innings, we find that player B is more consistent than player A. Thus measure of dispersion enables us to compare variability between two series.

3. Another objective of measuring dispersion is to determine nature and cause of variation in order to control the variation. In industrial production to control the quality variation dispersion is measured. In matters of health, variation in body temperature, pulse rate, blood pressure are the basic guides for diagnosis of the disease. Thus measurement of dispersion is basic to the control of cause of variation.
4. Measures of dispersion are used for other statistical analysis like correlation, regression, testing of hypothesis, analysis of variance, statistical quality control, etc.

TYPES OF DISPERSION

Dispersions are of two types - Absolute and Relative.

Absolute dispersion are expressed in the same statistical units in which the original data is expressed such as rupees, kilogram, tonnes, meters etc. Thus its applicability is restricted to the extent that the series to be compared should be expressed in same statistical units. In case two sets of data are expressed in different units the absolute measure of dispersions are not comparable.

Relative Dispersion: It is the ratio of a measure of absolute dispersion to an appropriate average. It is often called as coefficient of dispersion. While computing relative dispersion, it should be taken note of, that the average used as base should be the same as from which the absolute deviations were measured. The relative dispersion is expressed as an abstract number such as a ratio, percentage etc. It is always independent of units in which problem is expressed. Suppose there are 3% variation in weight, 5% variation in height of a group of students it implies that there are more variation in height compared to weight of this particular group of students. Relative measures of dispersion can be found out by calculating coefficients of absolute measures of dispersion.

MEASURES OF DISPERSION

Various measures of dispersion may be divided into two categories:

1. Absolute measures of dispersion

- a) Range
- b) Interquartile range
- c) Quartile deviation
- d) Mean deviation
- e) Standard deviation

2. Relative Measures of Dispersion

- a) Coefficient of range
- b) Coefficient Quartile deviation
- c) Coefficient of mean deviation
- d) Coefficient of variation.

Range, interquartile range and quartile deviation are positional measures as they depend on the value at a particular position in the distribution. They are counterparts of positional averages like median, quartiles, decile etc. The others are algebraic measures.

PROPERTIES OF A GOOD MEASURE OF DISPERSION

A satisfactory measure of dispersion should possess the same characteristics as prescribed for good measures of central tendency. However a good measures of dispersion should possess as far as possible the following properties.

1. It should be simple to understand,
2. It should be easy to calculate,
3. It should be rigidly defined,
4. It should be based on all the observations,
5. It should not be affected by sampling fluctuation, i.e. should have sampling stability.
6. It should be amenable to further algebraic treatment.
7. It should not be affected by extreme items.

METHODS OF STUDYING DISPERSION

Range: Range is the simplest of all the measures of dispersion. It is defined as the difference between the value of smallest and largest item in the series. Symbolically,

$$\text{Range} = L - S \quad \text{Where } L \text{ is the largest value} \\ S \text{ the smallest value}$$

The relative measure corresponding to range, called the coefficient of range, is obtained by applying the following formula:

$$\text{Coefficient of Range} = \frac{L - S}{L + S}$$

If the average of two distributions are about the same, a comparison of range indicates that the distribution with smaller range has less dispersion, hence its average is more typical or representative of the group.

Example: Calculate the range and the coefficient of range of the data given below:

Weekly earnings (in Rs) 147 150 149 170 163 155 181

Solution: Largest earning (L) = 181

Smallest earnings (S) = 147

Range (181 - 147) = 34

$$\text{Co-efficient of range} = \frac{181 - 147}{181 + 147} = \frac{34}{328} = 0.1036$$

When the data is available in the form of a continuous frequency distribution then range is calculated by any one of the following two methods:

1. By subtracting the lower limit of the lowest class from the highest limit of highest class. This method is known as method of limits.
2. By subtracting the mid value of lowest class from mid value of highest class (method of mid point). The following example will illustrate the above point.

Computation of Range

Example: Compute the range and co-efficient of range from the following frequency distribution.

Marks	10-20	20-30	30-40	40-50	50-60
No. of Students	7	10	13	8	4

Solution:

Marks Class	Mid point
10-20	15
20-30	25
30-40	35
40-50	45
50-60	55

1st method - Range = $60 - 10 = 50$ ($R = L - S$)

$$\text{Coefficient of range} = \frac{60 - 10}{60 + 10} \quad (\text{Coefficient of range} = \frac{L - S}{L + S})$$

$$= \frac{50}{70} = .7143$$

2nd method - Range = $55 - 15 = 40$

$$\text{Coefficient of range} = \frac{55 - 15}{55 + 15} = \frac{40}{70} = .5714$$

It should be noted that in calculation of range only the value of variables are taken in to account and not frequencies.

Merits:

1. It is easy to calculate and simple to understand as it is just the difference between maximum and minimum value.
2. It is mostly used in the construction of quality control charts.
3. It results in saving of time and labour.

Demerits:

Range though is simplest it is the crudest measure of dispersion due to the following weaknesses:

1. It is not based on all observations of the series.

2. It is affected by sampling distribution i.e. a simple variation in the value of extreme items affect the value of the range because it takes only the extreme items into consideration.
3. It can not be used in case of open end distribution.
4. Range does not speak of the character of distribution within the two extreme observation.
5. In case of continuous data range is likely to be incorrect as the exact values of the smallest and the largest item would not be known.

Uses of Range

With all its limitations range is commonly used in fields like quality control, weather forecasting, comparing two or more series with regard to variability etc.

INTERQUARTILE RANGE OR QUARTILE DEVIATION

Just as in case of range the difference of extreme items is found, similarly, it would give us inter quartile range if the difference in the values of two quartiles is calculated. Quartiles are the points which divide the array of data into four equal parts and they are positional measures. The difference between 1st and 3rd quartile is the inter quartile range which contains the middle half of the items. The dependence of range on extreme items can be avoided by adopting this measure. It is obvious that only the middle half of the data ($Q_3 - Q_1$) is required for computing the value which is supposed to be more representative of the group. Symbolically interquartile range is ($Q_3 - Q_1$). Very often the interquartile range is reduced to the form of semi-interquartile range or quartile deviation by dividing it by two. Thus

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

The more the data are spread out the wider the interval between two quartiles within which the middle half of the case lie. Quartile deviation gives the average amount by which the two quartiles differ from median

In a symmetrical distribution the two quartiles ($Q_3 - Q_1$) are equidistant from median and as such the difference can be taken as a measure of dispersion. When quartile deviation is a very small it describes high uniformity of central 50% of items. Quartile deviation is an absolute measure of dispersion. The relative measure called coefficient of quartile deviation is calculated as follows:

$$\text{coefficient of quartile deviation} = \frac{Q_3 - Q_1/2}{Q_3 + Q_1/2} + \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

The following examples would clarify computation of quartile deviation for individual observation, discrete series and continuous series.

Computation of Quartile Deviation

(a) Individual Observation

Example: In the following table marks obtained by 12 students are given.

96, 88, 24, 28, 31, 40, 49, 82, 58, 56, 65, 82

Find the quartile deviation and its coefficients.

Solution:

Arranging the data in ascending order we get

24, 28, 31, 40, 49, 56, 58, 65, 82, 88, 96

$$Q_1 = \text{Size of } \left[\frac{N+1}{4} \right]^{\text{th}} \text{ item} = \text{Size of } \left[\frac{11+1}{4} \right]$$

item = 3rd item

size of 3rd item is 31 thus $Q_1 = 31$

$$Q_3 = \text{Size of } 3 \left[\frac{N+1}{4} \right]^{\text{th}} \text{ item} = \text{Size of } 3 \left[\frac{11+1}{4} \right]$$

item = 3rd item

$$Q_3 = 3 \times \frac{12}{4} = 9\text{th item i.e. } 82$$

$$QD = \frac{Q_3 - Q_1}{2} = \frac{82 - 31}{2} = \frac{51}{2} = 25.5$$

$$\text{Coefficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{82 - 31}{82 + 31} = \frac{51}{113} = 0.451$$

b) Discrete Series

Example: Find out quartile deviation and its coefficient from the following observation:

Marks	10	15	20	25	30	35	40	90
No. of Students	6	17	30	38	20	14	19	1

Solution:

Calculation of Quartile Deviation

Mark (x)	Frequency(f)	Cumulative frequency(f)
10	6	6
15	17	23
20	30	53
25	38	91
30	20	111
35	14	125
40	9	134
90	1	135
N = 135		

$$Q_1 = \text{Size of } \left[\frac{N+1}{4} \right]^{\text{th}} \text{ item} = \left[\frac{135+1}{4} \right] = 34\text{th item}$$

which corresponds to 20; thus $Q_1 = 20$

$$Q_3 = \text{Size of } 3 \left[\frac{N+1}{4} \right]^{\text{th}} \text{ item} = \text{Size of } 3 \left[\frac{135+1}{4} \right]$$

$$= 102^{\text{th}} \text{ item}$$

Size of 102th item is 30; thus $Q_3 = 30$

$$\text{Quartile Deviation} = \frac{30 - 20}{2} = 5$$

$$\text{Coefficient of Quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{30 - 20}{30 + 31} = \frac{10}{50} = 0.2$$

C) Continuous Series

Example: Compute coefficient of quartile deviation from the following data

Age group	20-25	25-30	30-35	35-40	40-45	45-50
No. of persons	50	70	100	180	150	70

Solution:

Calculation of Inter Quartile Range

Age group	No. of persons	Cumulative frequency
20-25	50	50
25-30	70	120
30-35	100	220
35-40	180	400
40-45	150	550
45-50	70	620
N = 620		

$$\text{Lower quartile } Q_1 = \text{Size of } \frac{N}{4} = \frac{620}{4} = 155^{\text{th}} \text{ item}$$

Q_1 lies in the class 30-35

$$Q_1 = L_1 + \frac{\frac{N}{4} - C}{F} (L_2 - L_1)$$

Where $L_1 = 30$; $L_2 = 35$; $\frac{N}{4} = 155$; $C = 120$; $f = 100$

Substituting the values we get

$$30 + \frac{\frac{620}{7} - 120}{100} \times 5 = 30 + \frac{150 - 120}{100} \times 5$$

$$= 30 + \frac{35}{20} = 30 + 1.75 = 31.75$$

Upper quartile = Q_3 = Size of $\frac{3N^{\text{th}}}{4}$ item = $3 \left[\frac{620}{4} \right]$
 $= 465^{\text{th}}$ item

465^{th} item lies in the 40-45 class

$$\text{The value of } Q_3 = L_1 + \frac{3 \left[\frac{N}{4} \right] - C}{f} (L_2 - L_1)$$

Where $L_1 = 40$; $L_2 = 45$; $\frac{N}{4} = 155$; $C = 400$; $f = 150$

Substituting the values

$$Q_3 = 40 + \frac{465 - 400}{150} (45 - 40)$$

$$= 40 + \frac{65}{150} \times 5 = 40 + 2.17 = 42.17$$

Inter quartile range = $Q_3 - Q_1 = 42.17 - 31.75 = 10.42$

Quartile deviation = $\frac{Q_3 - Q_1}{2} = \frac{10.42}{2} = 5.21$

Coefficient of quartile deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{42.17 - 31.75}{42.17 + 31.75}$
 $= \frac{10.42}{72.92} = 0.143$

Merits and Demerits of Quartile Deviation

Merits:

The quartile deviation possesses the following merits:

1. It is simple to calculate and easy to understand.
2. It is not affected by extreme values like range.
3. It can be computed in case of open end distribution.
4. It is used in cases wherein the data may be ranked but not measured quantitatively.
5. It can be computed even if the distribution has unequal class intervals.

Demerits:

The following are the main limitations of quartile deviation:

1. Quartile deviation ignores 50% of the items i.e. the first and last 25%. As the value of quartile deviation does not depend on all the observation on the series it can not be regarded as a good method of measuring dispersion.
2. It is not amenable to further algebraic treatment.
3. As it is mainly dependent on the values of central items if these values are irregular and abnormal, the result is bound to be affected i.e. it is affected much by sampling distribution.
4. This measure is not affected by distribution of the items between the quartiles or by the distribution outside the quartiles. In other words the value of quartiles may be same for two quite dissimilar distribution.
5. As quartile deviation does not show the scatter of values in the series around the average it is in the true sense of the term, not regarded as a measure of dispersions. It simply shows a distance on a scale (i.e. distance between Q_3 and Q_1); consequently it is regarded more as a measure of partition rather than dispersion.

Uses of Quartile Deviation

Because of the above limitations quartile deviation is not often useful for statistical inference. Uses of quartile deviation as a measure of dispersion is done mainly for open end distribution.

MEAN DEVIATION

The range, the inter quartile range and the quartile deviation suffer from a common defect i.e. they are calculated on the basis of only two values in a series. Such measures of dispersion suffer from a number of demerits. In the strict sense of the term 'dispersion' the above measures are not regarded as the measures of dispersion as they do not show the scatteredness around the average. However to study the formation of a distribution one should take the deviation from the average. Mean deviation helps us in achieving this goal. The mean deviation alternatively called average deviation, is the arithmetic average of deviations of the values from a measure of central tendency without taking their signs (+ or -) into consideration. Theoretically deviation can be taken from any of the three averages i.e. mean, median, mode, but in actual practice mean deviation is calculated either from mean or from median. The use of mode is not recommended because in most of the series mode is ill defined. As regards the choice between mean and median, median is supposed to be better than mean. More reliance is placed upon median because the sum of deviation from the median is minimum compared to the sum of deviations of the values from mean. It should be noted here the sign of the deviation should be ignored while aggregating the deviations as, if it is taken into account the sum of the deviation from mean would be zero and median would be nearly zero. Since the purpose of a measure of dispersion is to study the variation of items from the central value it matters least if the signs are ignored. Mean deviation is also known as first moment of dispersion. Symbolically,

$$(i) \quad \alpha \bar{x} = \frac{\sum (dx)}{N}$$

Where αx stands for mean deviations from mean \bar{x} for the deviations from the mean and N for the number of items.

$$(ii) \sigma_m = \frac{\Sigma (dm)}{N}$$

Where σ_m stands for mean deviation from median \bar{d}_x for deviations from median and N for number of items. Mean deviation or first moment of dispersion as calculated above would be an absolute measure of dispersion. In order to transform it into relative measure it is divided by the average from which the deviation is taken. It is then known as mean coefficient of dispersion.

Thus the coefficients of dispersion from mean, median would be respectively.

$$\frac{\sigma_x}{\bar{x}}, \quad \frac{\sigma_m}{\bar{m}}$$

Computation of Mean Deviation

If $X_1, X_2, X_3, \dots, X_n$ are N given observations then the deviation about an average A is given by

$$MD = \frac{1}{N} \Sigma (X - A) = \frac{1}{N} \Sigma (D) \text{ or } \frac{\Sigma (d)}{N}$$

Steps:

1. Compute the average (mean or median) of the series.
2. Add the deviations of items from the average ignoring + sign and denote these deviation by (D)
3. Obtain the total of these deviations i.e. $\Sigma(D)$
4. Divide the total obtained in step(3) by the number of observation N .

In practice, Arithmetic mean is more frequently used in calculating the value of average deviation.

Example: Calculate mean deviation and its coefficient of the mean deviation of the two income group of 5 and 7 members given below.

I Rs. 3000 3050 4000 4400 4700 4800 5800

II Rs. 4000 4050 4400 4700 4900

Solution:

Calculation of Mean Deviation

GROUP I	Deviation from median 4400 (D)	GROUP II	Deviation from median 4400 (D)
3000	1400	4000	400
3050	1350	4050	350
4000	400	4400	0
4400	0	4700	300
4700	300	4900	500
4800	400		
5800	1400		
	5250		1550

Mean deviation Group I = $Md = \frac{\sum IDI}{N}$ where IDI is

Deviation from median ignoring signs

Median size of $\frac{N + 1th}{2}$ item = size of $\frac{7 + 1th}{2}$ item

i.e 4th item

Size of 4th item is 4400

Mean Deviation is $\frac{5250}{7} = 750$

Mean Deviation Group B

Median size of $\frac{N + 1th}{2}$ item = $\frac{(5 + 1)th}{2}$ item = 3rd item

Size of 3rd item = 4400 = ΣIDI = 1550 $N = 5$

$$MD = \frac{1550}{5} = 310$$

Coefficients of mean deviation group I = $\frac{MD}{Median}$

$$\frac{750}{4400} = 0.1682$$

$$\text{group II} = \frac{310}{4400} = 0.0705$$

Calculation of Mean Deviation (Discrete Series)

In discrete series the formula for calculating mean deviation is

$$MD = \frac{\Sigma f|DI}{N}$$

IDI denotes deviation from median ignoring signs.

Steps:

1. Calculate the median of the series.
2. Take the deviation of items from median ignoring signs and denote them by IDI
3. Multiply these deviations by respective frequencies and obtain the total of $\Sigma f|DI$
4. Divide the total obtained in step 3 by the number of observations.
This gives the value of mean deviation.

Examples: Find the mean deviation of the distribution:

No. of Printing mistakes per page	0	1	2	3	4	5	6	7
No. of Pages	21	16	17	13	11	10	7	5

Solution:**Calculation of Mean Deviation**

Number of printing mistakes per page x	No. of pages f	CF	$X - 2$ IDI	$f D $
0	21	21	2	42
1	16	37	1	16
2	17	54	0	0
3	13	67	1	13
4	11	78	2	22
5	10	88	3	30
6	7	95	4	28
7	5	100	5	25
$N = 100$				$\Sigma f D = 176$

$$\text{Median} = \text{Size of } \frac{N^{\text{th}}}{2} \text{ item}$$

$$= \text{Size of } \frac{100^{\text{th}}}{2} \text{ item} = 50^{\text{th}} \text{ item}$$

The size of 50th item is 2

Thus mean deviation is

$$MD = \frac{\Sigma f|D|}{N}$$

$$= \frac{176}{100}$$

$$= 1.76$$

Thus mean deviation is 1.76.

Example: Calculate mean deviation from the following data:

X	2	4	6	8	10	12	14	16
f	2	3	4	5	3	2	1	2

Calculation of Mean Deviation from Mean

X	f	fx	IDI x-Mean	fIDI
2	2	4	6	12
4	2	8	4	8
6	4	24	2	8
8	5	40	0	0
10	3	30	2	6
12	2	24	4	8
14	1	14	6	6
16	1	16	8	8
N = 20		160		$\Sigma fIDI = 56$
$X = \frac{\Sigma fx}{N} = \frac{160}{20} = 8$				

$$\text{Mean deviation} = \frac{\Sigma fIDI}{N} = \frac{56}{20} = 2.8$$

Calculation of Mean Deviation - Continuous series

For calculation of mean deviation in a continuous series the procedure to be adopted is the same as in case of discrete series. The difference between discrete series and continuous series calculation is that we have to obtain mid points of the class and the classes are to be replaced by mid points.

Steps:

Computation of mean deviation from continuous series involves the following steps:

1. Write down the mid points of all classes.
2. Compute the mean or median.

3. Write down the deviations of the mid points either from mean or from median IDI, disregarding the signs.
4. Multiply the deviations of each class by their respective class frequencies $f \text{IDI}$.
5. Sum the product obtained in Step 4. $\Sigma f \text{IDI}$.
6. Divide the total products by total frequency; the resulting value is the mean deviation.

It may be expressed in the formula

$$\text{MD} = \frac{\Sigma f \text{IDI}}{N}$$

Example: Calculate the mean deviations for the following series and find out its coefficient.

No. of

Marks 10-20 20-30 30-40 40-50 50-60 60-70 70-80 80-90

No. of Students 2 6 12 18 25 20 10 7

Marks	Mid Value	f	cf	d = X-54.80	$\Sigma f \text{IDI}$
10-20	15	2	2	39.8	79.60
20-30	25	6	8	29.8	178.80
30-40	35	12	22	19.8	237.60
40-50	45	18	38	9.8	176.40
50-60	55	25	63	0.2	5.00
60-70	65	20	83	10.2	204.00
70-80	75	10	93	20.2	202.00
80-90	85	7	100	30.2	
					1284.00

$$\text{Median Class} = \frac{N}{2} = \frac{100}{2} = 50$$

Median lies in 50-60 items

$$\text{Median} = L_2 + \frac{\frac{N}{2} - C}{f} (L_2 - L_1)$$

$$= 50 + \frac{50 - 30}{25} \cdot 10$$

$$= 50 + 4.80$$

$$= 54.80$$

$$\text{Mean Deviation} = \frac{\sum f|DI|}{N}$$

$$\frac{1294.8}{100}$$

$$= 12.948$$

$$\text{Coefficient of Mean deviation} = \frac{\text{MD}}{\text{Median}} = \frac{12.948}{54.8} = .2363$$

Calculation of mean deviation from mean

Example: Find out the mean deviation from the mean from the following data:

x	7.5-12.5	12-17.5	17.5-22.5	22.5-27.5	27.5-32.5	32.5-37.5	37.5-42.5
f	5	17	28	38	25	14	8
X	Mid Value	f	$\frac{d'm - 25}{5}$	fd	IDI	fIDI	
7.5-12.5	10	5	-3	-15	3	15	
12.5-17.5	15	17	-2	-34	2	34	
17.5-22.5	20	28	-1	-28	1	28	
22.5-27.5	25	38	0	0	0	0	
27.5-32.5	30	25	1	+24	1	25	
32.5-37.5	35	14	2	+28	2	28	
37.5-42.5	40	8	3	-24	3	24	
N = 133		fd' = 0		$\Sigma fIDI = 154$			

$$\bar{X} = A + \frac{fd'}{N} \times C \quad (4 \text{ Assumed Mean } 25)$$

$$fd' = 0; N = 135; c = 5$$

$$= 25 + \frac{0}{135} \times 5 = 25$$

$$\text{Mean Deviation} = \frac{\sum f|D|}{N} \quad C = \frac{154}{135} \times 5 = \frac{154}{27}$$

$$= 5.074$$

Calculation of Mean Deviation-Shortcut Method- 1

When mean or median is in fraction, calculation of mean deviation becomes difficult. The computation can be done in a simplified way by following the short cut method, the formula for which is as follows:

$$\text{MD (from mean)} = \frac{\sum mfA - \sum mFB - (\sum fA - \sum fB) \bar{x}}{N}$$

$$\text{MD (from median)} = \frac{\sum mfA - \sum mFB - (\sum fA - \sum fB) \text{median}}{N}$$

$\sum mfA$ and $\sum mfB$ stand for total of products of mid points and frequencies corresponding to mid points above and below the average value respectively.

$\sum fA$ and $\sum fB$ represent the total frequencies pertaining to mid point above and below the average value.

Example: From the following, calculate mean deviation from mean.

Class	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency	14	18	24	30	45	35	22	12

X	Mid Points	f	cf	mf	
0-5	2.5	14	14	35	
5-10	7.5	18	32	135	$\Sigma mfB = 995$
10-15	12.5	24	56	300	$\Sigma fB = 86$
15-20	17.5	30	86	525	
20-25	22.5	45	131	1012.5	
25-30	27.5	35	166	962.5	$\Sigma mfA = 3140$
30-35	32.5	22	188	715	$\Sigma fA = 114$
35-40	37.5	12	200	450	
		N = 200	$\Sigma mA = 4135$		

$$\bar{X} = \frac{\Sigma mf}{N} = \frac{4135}{200} = 20.675$$

$$MD \text{ (from mean)} = \frac{\Sigma mfA - \Sigma mFB - (\Sigma fA - \Sigma fB)}{N}$$

$$= \frac{3140 - 995 - (114 - 86)}{200} \times 20.675$$

$$= \frac{3140 - 995 - 578.9}{200} = 1516.1 = 7.8305$$

Calculating of Mean Deviation from Median

$$\text{Median} = \text{Size of } \frac{N}{2} \text{ item } = \frac{210}{2} = 100\text{th item}$$

Thus Median lies in 20-25 class

$$\text{Median} = L1 + \frac{\frac{N}{2} - Cf}{f} \times 2$$

$$= \frac{200}{2} - 86$$

$$= 20 + \frac{2}{48} \times 5$$

$$= 20 + \frac{14}{45} \times 5 = 21.56$$

$$\text{Mean deviation} = \frac{\Sigma mfA - \Sigma mfB - (114 - 86) \text{ median}}{N}$$

$$= \frac{3140 - 995 - (114 - 86) 21.56}{200}$$

$$= \frac{3140 - 995 - 603.68}{200} = 7.71$$

Short cut method 2: Mean deviation can be computed by another short cut method where the deviations are taken from assumed mean or median. The formulas are

$$\text{MD (from mean)} = \frac{\Sigma f|dx| + (Fa - Fb) \bar{X} - A}{N}$$

When $\Sigma f|dx|$ - sum of products of the deviations times frequencies when deviations are taken from the assumed mean.

fa = sum of frequencies above the mean

fb = sum of the frequencies below the mean

XA = Difference between real mean and assumed mean

The same problem is worked out by the help of short cut method 2 as under:

Calculation of Mean Deviation

Class	Mid Points m	frequency f	Assumed mean 22.5 $dx = \frac{X - 22.5}{5}$	fdx
0-5	2.5	14	-4	-56
5-10	7.5	18	-3	-57
10-15	12.5	24	-2	-48
15-20	17.5	30	-1	-30

20-25	22.5	45	0	0
25-30	27.5	35	1	35
30-35	32.5	22	2	44
35-40	37.5	12	3	36
<hr/> N = 200			<hr/> $\Sigma fdx = -73$ <hr/>	

$$\text{Mean } (X) = A + \frac{\Sigma fdx}{N} \times C$$

$$= 22.5 + \frac{73}{200} \times 5 = 22.5 + \frac{-365}{200} = 20.7$$

$$MD = \frac{\Sigma f|dx| + (f_a - f_b) \times (X - A)}{N}$$

Where $\Sigma f|dx|$ = Total deviation from assumed mean ignoring signs

$$303 \times 5 = 1515$$

f_a = Sum of frequency above mean (20.7)

$$= 14 + 18 + 24 + 30 + 86$$

f_b = sum of frequencies below the mean (20.7)

$$= 45 + 35 + 22 + 12 = 114$$

$$= \bar{X} - A = \text{Difference between real mean and assumed man.}$$

Substituting the values in the formula

$$MD = \frac{1515 + (86 - 116) (-18)}{200} = \frac{1565.4}{200}$$

Thus mean deviation is 7.82.

Merits:

Mean deviation possesses the following merits:

1. The method of calculation is very simple as it can be done by just averaging the deviation from a measure of central tendency.
2. It is based on each and every individual item in the series.

3. Mean deviation is rigidly defined and its value is precise and definite.
4. Mean deviation is less affected by the value of extreme items like range and quartile deviation.
5. In a symmetrical distribution 57.5% of the items in a series fall in a range of $\text{mean} \pm \text{mean deviation}$ or $\text{median} \pm \text{mean deviation}$.

Demerits:

1. It is not amenable to algebraic treatment as it ignores the algebraic signs which is mathematically illogical.
2. Its computation is laborious and time consuming in case of large samples and frequency distributions.
3. This method may not give us very accurate results. The reason is that mean deviation gives best result when deviations are taken from median. But median is not a satisfactory measure when degree of variability in the series is very high.
4. From the mean deviation of several groups of observations it is not possible to find the mean deviation of combined group.

Uses

Despite the demerits mean deviation still has its practical utility. Because of its simplicity in meaning and computation it is mostly used by economists and businessmen. It is specially effective in reports presented to general public or to groups not familiar with statistical methods. This measure is useful for small samples with no elaborate analysis required.

STANDARD DEVIATION

The standard deviation (SD) is another method of summing up of deviations from measure of central tendency and finding a measure of dispersions of the data. SD is by far the most important and widely used measure of studying dispersion. Its significance lies in the fact that it is free from the defects which the earlier methods suffer and satisfy most of the properties of a good measure of dispersion. It mathematically represents the variability which is very important for interpreting statistical data.

The standard deviation measures the absolute dispersion or variability of a distribution; the greater the amount of dispersion the greater is the standard deviation. Thus standard deviation is extremely useful in judging the representatives of the mean. Standard deviation is the square root of the arithmetic average of the squares of the deviations measured from the mean. That is why it is otherwise known as root mean square deviation. Standard deviation is denoted by the small greek letter σ (read as sigma).

Standard deviation differs from mean deviation mainly on two respects:

1. Algebraic signs are not ignored while computing standard deviation unlike mean deviation.
2. Standard deviation is calculated from arithmetic mean only whereas mean deviation is computed either from mean or median.

Calculation of Standard Deviation (Individual observation)

The standard deviation in individual series can be calculated in any of the following methods:

A. Direct Method

Steps:

1. Find the actual mean of the series
2. Find the deviation of each item from the mean i.e. find $(x - \bar{x})$ denote it by (d)
3. Square these deviations and total them (Σd^2)
4. Divide Σd^2 by number of items
5. Now find out the square root of the value obtained in step 4.

$$\text{Formula applied in this method is } = \sqrt{\frac{\Sigma d^2}{N}}$$

B. Short Cut method**Steps:**

1. Assume a mean and take deviations (dx) of each item from assumed mean (i.e. calculate $X-A$) and square them up. Total all the deviations Σdx as well as such square deviation or find out Σdx^2
2. Substitute the value of Σdx , Σdx^2 and N in the formula

$$\sqrt{\frac{\Sigma dx^2}{N} - \left| \frac{\Sigma dx}{N} \right|^2}$$

The following example would illustrate the use of these formula

Example: Calculate the Standard Deviation of the Production of rice of 10 firms of equal sizes, production in quintals 49, 52, 59, 66, 55, 40, 60, 57, 52, 61.

Solution:**Calculation of Standard Deviation**

Production quintals x	Deviation from Mean 55 d	Deviation squared d^2
49	-7	49
52	-3	9
59	4	16
66	11	121
55	-15	225
40	5	25
60	2	4
57	-3	9
52	6	36
61		
$\Sigma x = 550$	$\Sigma dx = 0$	$\Sigma d^2 = 494$

$$\text{Arithmetic mean} = \frac{\Sigma x}{N} = \frac{550}{10} = 55$$

$$\sigma = \sqrt{\frac{d^2}{N}} = \frac{494}{N} = \sqrt{49.4} = 7.03$$

The same question is solved by other methods i.e. short-cut method.

Short-cut Method

Production quintals x	Deviation from assumed Mean 55 dx (57)	dx ²
48	-9	81
52	-5	25
59	2	4
66	9	81
55	-2	4
40	-17	289
60	3	9
57	0	0
52	-5	25
61	4	16
$\Sigma dx = -20$		$\Sigma dx^2 = 534$

$$\sigma = \sqrt{\frac{\Sigma dx^2}{N} - \left| \frac{\Sigma dx}{N} \right|^2} = \frac{534}{10} - \left[\frac{-20}{10} \right]^2$$

$$\sqrt{53.4 - 4} = \sqrt{49.4}$$

$$= 7.03$$

Calculation of Standard Deviation - Discrete Series

For calculating standard deviation in discrete series any of the following methods can be applied:

1. Actual mean method
2. Assumed mean method
3. Step deviation method

Actual Mean Method

Steps:

1. Calculate arithmetic mean.
2. Find out the deviations of various values from the mean. Square these deviation (d^2)
3. Multiply d^2 with respective frequencies of against various values and add all the value Σfd^2
4. Divide Σfd^2 by number of items (N) and find out the square root of the figure

$$\text{Formula } \sigma = \sqrt{\frac{\Sigma fd^2}{N}}$$

Assumed Mean Method

1. Assume mean x and take deviations (dx) and square them up (dx^2)
2. Multiply dx with respective frequency f to get fdx add them up to get Σfdx .
3. Multiply dx^2 with respective frequencies (f) to get fdx^2 . Total them Σfdx^2
4. Put them in formula

$$\sigma = \frac{(\Sigma fdx^2)}{N} - \frac{(\Sigma fdx)^2}{N}$$

5. Step Deviation method - In this method we take a common factor from the given data. The formula for computing standard deviation is

$$\sigma = \sqrt{\frac{\Sigma fdx^2}{N} - \left(\frac{\Sigma fd1^x}{N}\right)^2} \times i$$

$$d^1 x = \frac{X - A}{i} \text{ and } i \text{ is the common factor:}$$

Example: Find out standard deviation from the following data

x	6	7	8	9	10	11	12
f	3	6	9	13	8	5	4

Calculation of Standard Deviation (Actual Mean Method)

x	f	fx	($\bar{x} - x$)	d ²	fd ²
6	3	18	-3	9	27
7	6	42	-2	4	24
8	9	72	-1	1	9
10	8	80	1	1	8
11	5	55	2	4	20
12	4	48	3	9	36
$\Sigma f = 48$		$\Sigma fx = 432$			$\Sigma fd^2 = 124$

$$\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{432}{48} = 9$$

$$\sigma = \sqrt{\frac{\Sigma fdx^2}{N}}$$

$$= \sqrt{\frac{124}{48}} = \sqrt{2.58} = 1.6$$

Calculation of Standard Deviation (Assumed Mean Method)

x	f	(X-A)	d ²	fdx	fdx ²
6	3	-4	16	-12	48
7	6	-3	9	-18	54
9	13	-1	4	-18	36
10	8	0	0	0	0
11	5	1	1	5	5
12	4	2	4	8	16
$\Sigma f = 48$				$\Sigma fdx = -48$	$\Sigma fdx^2 = 172$

$$\sigma = \sqrt{\frac{\Sigma fdx^2}{N} - \left| \frac{\Sigma fdx}{N} \right|^2} = \sqrt{\frac{172}{48} - \left| \frac{-48}{48} \right|^2}$$

$$\sqrt{3.58 - 1} = 1.6$$

Calculation of Standard Deviation (Step Deviation Method)

Calculate Standard deviation for the following distribution:

x	10	20	30	40	50	60	70
f	1	5	16	24	16	10	5

Solution:

x	f	Deviation from assumed mean 24 dx/10	fd ¹ x	dx ²	fdx ²
10	1	-3	-3	9	9
20	5	-2	-10	4	20
30	15	-1	-15	1	15
40	24	0	0	0	0
50	16	1	16	1	16
60	10	2	20	4	40
70	5	3	15	9	45
$\Sigma f = 76$			$\Sigma fdx = 23$		$\Sigma fdx^2 = 145$

$$\sigma = \sqrt{\frac{\Sigma fd^2 x^2}{N} - \left| \frac{\Sigma fd^1 x^2}{N} \right|^2 \times 2} = \sqrt{\frac{145}{76} - \left| \frac{23}{76} \right|^2 \times 10}$$

$$\sqrt{1.91 - 0.09} \times 10$$

$$\sqrt{1.82 \times 10} = 1.35 \times 10 = 13.50$$

Calculation of Standard Deviation-Continuous Series

In continuous series any of the methods discussed above for discrete frequency distribution can be used. However in practice step deviation method is generally used.

The formula is =

$$\sigma = \sqrt{\frac{\sum fd^1 x^2}{N} - \left| \frac{\sum fd^1 x}{N} \right|^2 \times 2 =}$$

Where $d^1 x = \frac{m - A}{c}$

c = common factor.

m = mid point of the class.

Steps:

1. Find the mid points of various classes.
2. Take deviations of these mid points from an assumed mean taking a common factor ($d^1 x$)
3. Multiply the frequencies of each class with these deviation $fd^1 x$
4. Square the deviations and multiply with respective frequencies of each class $fd^1 x^2$
5. Put in the formula

The only difference between discrete series and continuous series is to find the mid points of the various classes.

Example: Find the Standard Deviation of following data:

Class	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60
f	10	12	15	17	30	37	20	12

Solution:

Calculation of Standard Deviation

Class	midpoint(x)	f	$d^1 \times \frac{x - 37.5}{5}$	$d^1 x^2$	$fd^1 x$	$fd^2 x$
20-25	22.5	10	-3	9	-30	90
25-30	27.5	12	-2	4	-24	48
30-35	32.5	15	-1	1	-15	15
35-40	37.5	17	0	0	0	0
40-45	42.5	30	1	1	30	30
45-50	47.5	37	2	4	74	148
50-55	52.5	20	3	9	60	180
55-60	57.5	12	4	16	48	192
N=153					$\Sigma fd^1 x$ = 143	$\Sigma fd^2 x$ = 143

$$\sigma = \sqrt{\frac{\Sigma fd^1 x^2}{N} - \left| \frac{\Sigma fd^1 x}{N} \right|^2} \times i = \sqrt{\frac{703}{153} - \left| \frac{143}{153} \right|^2} \times 5$$

$$\sqrt{4.59 - 0.87} \times 5$$

$$\sqrt{3.72} \times 5 = 1.92 \times 5 = 9.6$$

Coefficient of Standard Deviation

Coefficient of standard deviation is a relative measure of dispersion and is used for comparing two or more series expressed in different units.

$$\text{Coefficients of standard deviation} = \frac{\sigma}{\bar{x}}$$

Coefficient of standard deviation expressed in percentage is called coefficient of variation.

$$\text{Coefficient of variation (CV)} = \frac{\sigma}{\bar{x}} \times 100$$

It may be pointed out that although any measure of dispersion can be used in conjunction with any average in computing relative dispersion the most common of them is arithmetic mean as the average and standard deviation, the measure of dispersion. When relative dispersion is stated in terms of arithmetic mean and standard deviation the resulting percentage is known as coefficient of variation or coefficient of variability.

Example: A Survey of TVs of two model gives the following result.

Results Life	No. of TVs	
	Videocon	Western
0-2	5	2
2-4	16	7
4-6	13	12
6-8	7	19
8-10	5	9
10-12	4	1

Which model has got greater uniformity?

Solution: Calculation of Mean and Standard Deviation of Videocon and western.

Life	Mid	Dx (A-5)	dx ²	f	Videocon		f	Western	
					fdx	fdx ²		fdx	fdx ²
0-2	1	-4	16	5	-20	80	2	-8	32
2-4	3	-2	4	16	-32	64	7	-14	28
4-6	5	0	0	13	0	0	12	0	0
6-8	7	2	4	7	14	28	19	38	76
8-10	9	4	16	5	20	80	9	36	144
10-12	11	6	36	4	24	144	1	6	36
$\Sigma f = 50$					$\Sigma fdx = 6$	$\Sigma fdx^2 = 396$	$\Sigma f = 50$	$\Sigma fdx = 50$	$\Sigma fdx^2 = 316$

$$\text{Videocon Mean Life } \bar{x} = A + \frac{\sum fdx}{N} = 5 + \frac{6}{50} = 5.12 \text{ years}$$

Standard Deviation

$$SD = \sqrt{\frac{\sum fdx^2}{N} - \left| \frac{\sum fdx}{N} \right|^2} = \sqrt{\frac{396}{50} - \left| \frac{6}{50} \right|^2}$$

$$\sqrt{7.92 - .0114}$$

$$\sqrt{7.9056} = 2.8117$$

$$\text{Coefficient of Variation} = \frac{2.8117}{5.12} \times 100 = 54.92\%$$

$$\text{Western Mean life} = \bar{x} = 5 + \frac{58}{50} = 5 + 1.16 = 6.16 \text{ years}$$

$$SD = \sqrt{\frac{316}{50} - \left| \frac{58}{50} \right|^2}$$

$$\sqrt{6.32 - 1.3456} = \sqrt{4.9744} = 2.2304$$

$$\text{Coefficient of Variation} = \frac{2.2303}{6.16} \times 100 = 36.21\%$$

Since coefficient of variation of life of western TV is less, Western TV may be regarded as more uniform than Videocon TV.

Variance

The Variance of a set of value is the square of the standard deviation.

Symbolically $V = \sigma^2$ the concept of variance was used to describe the squares of standard deviation.

The concept of variance is highly important in advanced works. The significance of this term lies in the fact that it is capable of very exhaustive type of analysis.

Mathematical Properties of Standard Deviation

Standard deviation is regarded as the most reliable measure of dispersion due to the following important properties.

1. The most important property of standard deviation is that it is amenable to algebraic treatment. The combined standard deviation of two series is denoted by Σ_{12} and is computed by the following.

$$\sigma_{12} = \sqrt{\frac{N_1 (S_1^2 + d_1^2) + N_2 (S_2^2 + d_2^2)}{N_1 + N_2}}$$

Where $d_1 = \bar{x}_1 - \bar{x}_{12}$ difference between mean of first series and combined mean of both series.

Where $d_2 = \bar{x}_2 - \bar{x}_{12}$ difference between mean of second series and combined mean of both series.

2. The sum of square deviation taken from mean is minimum. It is for this reason that standard deviation is calculated from arithmetic mean rather than any other value.
3. Standard deviation is independent of the origin of measurement.

Merits of Standard Deviation:

The Standard Deviation possesses most of the characteristics which an ideal measure of dispersion should have. Thus

1. Standard deviation is rigidly defined and it has definite value always.
2. It is based on all deviations of data.
3. It is amenable to algebraic treatment and is less affected by fluctuations of sampling than most other measures of dispersions.
4. The difficulty about the algebraic signs which was experienced in mean deviation is not faced here as by squaring deviations become positive.
5. For comparing the variability of two or more distribution coefficient of variation is considered to be the most appropriate and this is based on mean and standard deviation.
6. Standard deviation is used in further statistical work, for example in calculation of skewness, correlation etc.

Demerits:

1. As compared to other measures of dispersion it is difficult to compute and is not easily understood.
2. It gives more weight to extreme items and less to those which are near mean. That is deviation when squared, big deviations get more prominence.
3. It cannot be computed for open end distribution.

Uses

Despite the drawbacks mentioned above the standard deviation is the best measure of dispersion and should be used wherever possible. Just as mean is the best measure of central tendency (leaving exceptional cases) standard deviation is the best measure of dispersion.

However since standard deviation gives greater weight to extreme items it does not find much favour with economists and businessmen who are more interested in the result of modal class.

Choice of a Measure of Dispersion

In actuality the choice for a particular measure of dispersion depends upon the nature of data, reliability of the measure and rapidity of computation.

1. Nature of data: If a distribution is not symmetrical, mean deviation should be preferred to standard deviation as standard deviation gives more weight to extreme value.

In situation where median is an ideal average mean deviation is the best measure of dispersion.

If a distribution has open ends, quartile deviation has an edge over mean and standard deviation.

In case of symmetrical distribution, standard deviation is the best measure of dispersion.

2. Reliability of the measure: If the objective of the study is reliability, standard deviation is the best measure as it possesses all the properties of a good measure of dispersion.

3. Rapidity of computation: If it is the rapidity of computation which is required, range is preferred over other measures of dispersion. However it is to be noted that reliability of the measure should not be sacrificed at the cost of computation.

4. Comparison: If the objective of the study is to compare two or more series expressed in different units, coefficient of variation is the right choice. Thus standard deviation is the best and most appropriate measure of dispersion.

LESSON - 2

MEASUREMENT OF SKEWNESS AND KURTOSIS

- ▣ SKEWNESS - INTRODUCTION
- ▣ DEFINITION
- ▣ TEST SKEWNESS
- ▣ MEASURES OF SKEWNESS
- ▣ KARL PEARSON'S COEFFICIENT
- ▣ KOWLEY'S COEFFICIENT
- ▣ KURTOSIS

SKEWNESS

Skewness like measures of central value and dispersion is a measure for the study of frequency distribution. Averages tell us about the central value of a distribution and measures of dispersion about the concentration of items around the central value. However these measures do not tell us whether the dispersal of values on either side of an average is symmetrical or not. Skewness refers to distortion from symmetry. Thus skewness measures the degree of departure from symmetry. When a distribution is symmetrical skewness is absent and the value of mean, median and mode coincide. The presence of skewness in a distribution pull the mean and median away from mode.

In order to study a frequency distribution it would be of great use to know whether it would give a symmetrical distribution and if not, to what extent it would deviate from symmetrical distribution. In fact the measures of central tendency, measures of dispersion should always be supplemented by the measures of skewness to give a full understanding of the frequency distribution.

DEFINITION

Some definitions of skewness are given below:

1. "When a series is not symmetrical it is said to be asymmetrical or skewed" - Crokton & Cowden.
2. "Skewness refers to the asymmetry or lack of symmetry in the shape of frequency distribution" - Morris Hamburg.
3. "A distribution is said to be 'skewed' when the mean and median fall at different points in the distribution and the balance (centre of gravity) is shifted to one side or the other to left or right - Garrett.
4. Measures of Skewness tell us the direction and the extent of skewness. In symmetrical distribution the mean, median and mode are identical. The more the mean moves away from the mode the larger is the asymmetry or skewness" - Simpson and Kafka.

Thus from the above definitions it is clear that the word skewness refers to lack of symmetry. If a distribution is normal or symmetrical there would be no skewness and when the distribution is skewed either it would be tilted to left or right.

If some observations in a frequency distribution are extremely large the mean is always greater than median and mode. The curve of such a distribution has a longer tail to the right. Such distributions are called positively skewed distribution. On the other hand, if some of the observations are extremely small in a distribution, the mean is always smaller than median and mode. Such distributions are called relatively skewed and the curve of such distributions has longer tail to the left of the maximum point. The following 3 figures would give an idea about the shape of symmetrical and asymmetrical curves.

Fig. 1 Gives the shape of an ideal symmetrical curve. It is bell shaped. The spread of frequency in the curve is same on the both sides of the centre point of the curve.

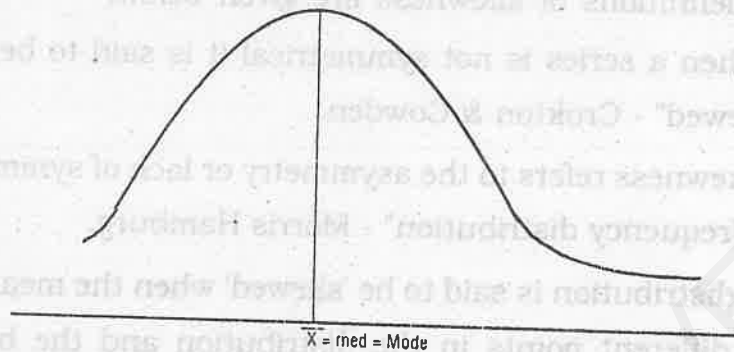


FIG. 1

The figure shows the value of mean, median and mode are identical and thus there is no skewness. A distribution which is not symmetrical may be towards the left of the centre point or towards the right of the centre point or median; accordingly they are called either positively skewed or negatively skewed distribution.

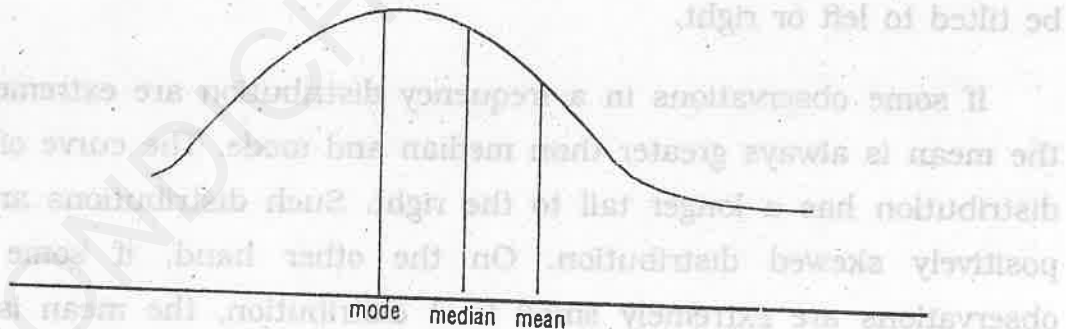


FIG. 2

Figure 2 gives the shape of moderately skewed curve. It is skewed to the right. In it the value of mean would be more than value of median and mode. Median would have a value higher than the value of mode. Such curves are called positively skewed curve.

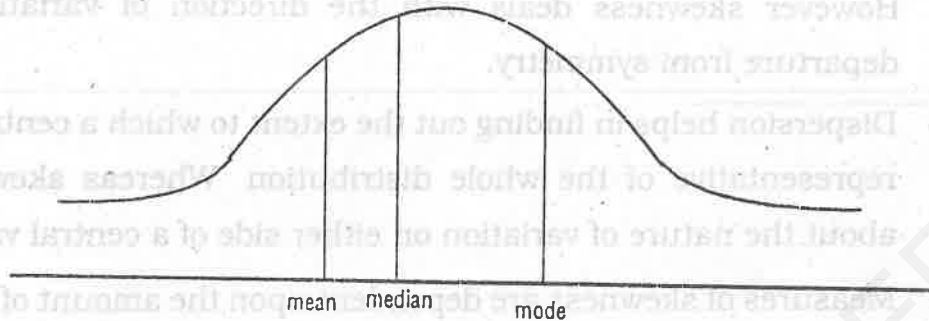


FIG. 3

Figure 3 also gives the shape of moderately skewed curve. This curve is skewed to the left and in it the value of mode would be greater than the value of median and the value of median would be greater than the value of the mean.

Thus in a positively skewed distribution the frequencies are spread out over a great range of values on the high value end of the curve than they are on the low value end. However in a negatively skewed distribution the position is reversed i.e. the excess tail is on the left-hand side.

It should be taken note that in a moderately symmetrical distribution the interval between mean and median is approximately one third of the interval between the mean and the mode. This relationship helps in measuring the degree of skewness.

From the above discussion we conclude

$$(i) \text{ in a symmetrical distribution } \bar{X} = M = Z$$

(mean) (median) (mode)

$$(ii) \text{ in a positively skewed distribution } \bar{x} > m > z$$

$$(iii) \text{ in a negatively skewed distribution } z > m > \bar{x}$$

Difference between Dispersion and Skewness

1. Dispersion deals with the spread of values around central value.
Skewness deals with symmetry of distribution of a central value.

2. Dispersion deals with the amount of variation rather than direction. However skewness deals with the direction of variation or the departure from symmetry.
3. Dispersion helps in finding out the extent to which a central value is representative of the whole distribution. Whereas skewness tells about the nature of variation on either side of a central value.
4. Measures of skewness are dependent upon the amount of dispersion not the vice versa.
5. Dispersion is by far the most important characteristic of distribution in business and economic series whereas though skewness is an important characteristic for defining the precise pattern of distribution, is rarely calculated for the same.

TEST OF SKEWNESS

There are certain tests with which it is possible to ascertain whether skewness exists or not. They are:

1. Absence of symmetry is denoted as skewness. Therefore in an asymmetrical distribution the value of mean, median and mode would not coincide. Mean and median are pulled wide apart from mode either to the right or to the left. That is mean in one end and mode on the other median being in the middle. We have already seen while discussing mode in the lesson of measure of central tendency that moderately asymmetrical distribution.

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

$$\text{or mode} = \text{mean} - 3 (\text{mean} - \text{median})$$

That is, it is assumed that $\text{mode} = 3 \text{ median} - 2 \text{ mean}$. Thus in a moderately asymmetrical distribution, we can ascertain the value of mode if we know mean and median.

1. In an asymmetrical distribution median travels about $\frac{2}{3}$ rd of the distance travelled by mean from mode.

2. In a symmetrical distribution median lies exactly half way between the quartiles. Thus if the first and third quartiles are different from median it indicates asymmetry or presence of skewness.

Symbolically

$$Q_3 - M \neq M - Q_1,$$

$$\text{or } (Q_3 - M) - (M - Q_1) \neq 0$$

3. A skewed distribution when plotted on a graph paper would not give a symmetrical bell shaped curve. That is when cut along a vertical line through the centre the two halves are not equal.
4. The sum of positive deviation from median is not equal to the sum of negative deviations.
5. The corresponding pairs of deciles or percentile are not equidistant from median.
6. Frequencies are not equal at points of equal deviations from the mode.
7. Otherwise stated, when skewness is absent i.e. in case of symmetrical distribution, the following conditions are satisfied:
- (i) The graph of distribution of frequencies shows the normal bell shaped symmetrical curve.
 - (ii) The value of mean, median mode of the distribution coincide.
 - (iii) The sum of positive deviation from the median is equal to negative deviations from the same value.
 - (iv) Quartiles are equidistant from the median.
 - (v) Frequencies are equally distributed at points of equal deviations from mode on both sides.
 - (vi) The corresponding pairs of deciles or percentiles are equidistant from median.

MEASURES OF SKEWNESS

The measures of skewness serve two purposes:

- (i) Firstly, the measures of skewness indicate the direction of skewness i.e. it is positive or negative.
- (ii) They measure extent of skewness. Though the direction of skewness can be depicted graphically the extent can only be measured by measures of skewness.

Thus measures of skewness tell us the direction and extent of asymmetry in a series and permit us to compare two or more with regard to these. Measure of skewness can either be absolute or relative.

Absolute Measure of Skewness

Skewness can be measured in absolute terms by taking symbolically $sk = \bar{X} - \text{Mode}$.

If the value of mean is greater than mode, skewness will be positive and if vice-versa is the case skewness will be negative. The reason why the difference between mean and mode can be used to measure skewness is that in a symmetrical distribution mean, median mode are same, but mean moves away the mode when the observations are asymmetrical.

The absolute measure of skewness is measured under another two methods.

1. Karl Pearson's
2. Bowley.

Karl Pearson's method is based on the assumption that in a skewed distribution mean, median mode do not coincide. Thus distance between any two values among mean, median mode would give us the extent of skewness i.e.

$$\text{Skewness} = \text{Mean} - \text{Mode or}$$

$$\text{Mean} - \text{Median or}$$

$$\text{Median} - \text{Mode}$$

It should be kept in the mind that for a symmetrical distribution skewness is zero as mean, median, mode coincide.

Bowley's method of measuring skewness is based on the assumption that in skewed distribution the quartiles are not equidistant from median i.e. (Median - Q₁ ≠ Q₃ - Median).

Thus we arrive at measure of skewness by finding out the difference between any of the two value of Q₃ median and Q₁. Symbolically skewness

$$= Q_3 - \text{Median} - (\text{Median} - Q_1)$$

$$= Q_3 - 2 \text{ Median} - Q_1$$

$$= Q_3 + Q_1 - 2 \text{ Median}$$

It should be mentioned here that the methods discussed above are based on different principles. Thus the results obtained would be different.

However the absolute measure of skewness are unsatisfactory on two counts:

- i) It would be expressed in the units of value of the distribution, therefore cannot be compared with a comparable series expressed in different units.
- ii) Distribution vary greatly and the difference between the mean and mode might be considerable in one series and small in another series although the frequency curves may be similarly skewed.

If absolute difference were expressed in relation to some measure of the spread of value in their respective distributions, the measures would be relative and can be used directly for comparison. This leads us to the discussion of the relative measures of skewness.

Relative Measures of Skewness

For the purpose of comparison of two series expressed in two different units it is necessary to use relative measure of skewness. In order to derive a relative measure, absolute measure of skewness is divided by a measure

of dispersion. In no case absolute measure be divided by a measure of central tendency unlike relative measure of dispersion where relative measure of dispersion is found out by dividing absolute measure of dispersion by a measure of central tendency. This is because our objective is to study the asymmetry in relation to dispersal of items round the central value rather than to study the extent of skewness in relation to the size of items. The most commonly used measures of skewness are

1. Karl Person's coefficient of skewness.
2. Bowley's coefficient of skewness.

However a good measure of skewness should have three properties. It should

1. Be a pure number in the sense that its value should be independent of units of the series and also of the degree of variation in the series.
2. Have zero value when distribution is symmetrical.
3. Have some meaningful scale of measure to make the interpretation easy.

KARL PEARSON'S COEFFICIENT OF SKEWNESS

This method of measuring skewness also known as Personian coefficient of skewness is based on the difference between mean and mode expressed in terms of standard deviation. The formula for measuring skewness is coefficient of skewness.

$$= \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

It is a pure number independent of units. Its value usually lies between 1 and the computed value is interpreted as follows:

- (1) When the value is zero, the distribution is symmetrical and value of mean, median mode are one and the same.
- (2) When its value is negative, the distribution is negatively skewed and the tail of such a distribution is to the left of the maximum point.

- (3) When its value is positive there is positive skewness in the series and the distribution has a tail to the right of the maximum point.

The formula serves two purposes. It gives both the direction as well as the extent of skewness.

The above formula of measuring skewness is not very popular as it involves inconvenience in determining the position of the mode and most of the series are ill-defined. Therefore it is better to use median in place of mode as we know there exist a relationship between the three measures of central tendency for a moderately skewed frequency distribution. The relationship is

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$$

Substituting the value of mode in the formula we get

$$\begin{aligned} \text{Coefficient of Skewness} &= \frac{\text{Mean} - (3 \text{ Median} - 2 \text{ Mean})}{\text{Standard Deviation}} \\ &= \frac{\text{Mean} - (3 \text{ Median} - 2 \text{ Mean})}{\text{Standard Deviation}} \\ &= \frac{3 (\text{Mean} - \text{Median})}{\text{Standard Deviation}} \end{aligned}$$

Theoretically, the value of this coefficient varies between ± 3 . However in practice it is rare that coefficient obtained by above method exceeds ± 1 .

Example: Form the following data calculate Karl Pearson's Coefficient of Skewness.

Marks	12	24	36	48	60	72	84
Students	8	14	18	36	30	20	10

Calculation of Coefficient of Skewness

X - 48/12	fdx	fdx2X	Marks f	Students dx
12	8	-3	-24	72
24	14	-2	-28	56
36	18	-1	-18	18
48	0	0	0	0
60	30	1	30	30
72	20	2	40	80
84	10	3	30	90
Σfdx 136			Σfdx 30	Σfdx^2 346

$$\text{Coefficient of Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{SD}}$$

$$\text{Mean} = A + \frac{\Sigma fdx}{N} \times C$$

$$A = 48; \Sigma fdx = 30; N = 136 \quad C = 12$$

$$= 48 + \frac{30}{136} \times 12$$

$$= 48 + 2.647 = 50.647$$

By inspection we find modal value is 48.

Standard deviation

$$\text{SD} = \sqrt{\frac{\Sigma fdx^2}{2} - \left[\frac{\Sigma fdx}{N} \right]^2 \times C}$$

$$\Sigma fdx^2 = 346$$

$$\Sigma fdx = 30$$

$$N = 136$$

$$C = 12$$

$$\sqrt{\frac{346}{2} - \left[\frac{30}{136}\right]^2} \times 12$$

$$= \sqrt{2.544 - 0.0499}$$

$$= 1.5796 \times 12 = 18.955$$

$$\text{Coefficient of skewness} = \frac{\text{Mean} - \text{Mode}}{\text{SD}}$$

$$\frac{50.647 - 48}{18.955} = \frac{2.647}{18.955} = 0.14$$

Example: Calculate Karl Pearson's coefficient of skewness from the following data:

Class X	Mid Values f	dx N-35/10	freq	fdx	fdx ²
0-10	5	-3	5	15	45
10-20	15	-2	6	-12	24
20-30	25	-1	11	-11	11
30-40	35	0	21	0	0
40-50	45	1	35	35	35
50-60	55	2	30	60	120
60-70	65	3	22	66	198
70-80	75	4	18	72	288
			N = 148	Σfdx=195	Σfdx ² = 718

Karl Pearson's Coefficient of Skewness

$$SK_p = \frac{\text{Mean} - \text{Mode}}{\text{SD}}$$

$$\text{Mean} = A + \frac{\Sigma fdx}{N} \times C$$

$$= 35 + \frac{195}{148} \times 10$$

$$= 48.1757$$

By inspection mode lies in 40-45 class since in that class frequency is high i.e. 55.

$$\text{Mode} = L + \frac{\Delta}{\Delta} \times 2$$

$$= 40 + \frac{14}{14+5} \times 10$$

$$= 40 + 7.368$$

$$= 47.368$$

Standard Deviation = SD (δ)

$$SD = \sqrt{\frac{\sum f dx^2}{N} - \left[\frac{\sum f dx}{N} \right]^2} \times C$$

$$= \sqrt{\frac{718}{148} - \left[\frac{195}{148} \right]^2} \times 10$$

$$= \sqrt{4.850 - 1.736} \times 10$$

$$= \sqrt{3.114} \times 10$$

$$= 17.6465$$

Pearson's Coefficient of Skewness =

$$\text{Coefficient of Skewness} = \frac{\text{Mean} - \text{Mode}}{SD}$$

$$\text{Sq} \sqrt{\frac{48.1756 - 47.363}{17.6465}}$$

$$= \frac{0.8677}{17.6465} = 0.0458$$

BOWLEY'S COEFFICIENT OF SKEWNESS

Prof. Bowley's measure of skewness is based upon median and quartiles. In a symmetrical distribution first and third quartiles are supposed to be equidistant from median. Whereas it is not so in case of an asymmetrical distribution. Therefore the difference between the values of $Q_3 - M$ and $M - Q_1$ is employed for measuring skewness in a series. If the difference is positively skewed the top 25 percent of the values will tend to be farther from median than the bottom 25 per cent i.e Q_3 will be farther from median than Q_1 is from median and reverse for negative skewness.

According to Bowley

$$Sk_b = \frac{(Q_3 - \text{Med}) - (\text{Med} - Q_1)}{(Q_3 - \text{Med}) + (\text{Med} - Q_1)}$$

or

$$= \frac{Q_3 + Q_1 - \text{Median}}{Q_3 - Q_1}$$

The denominator is in fact twice the quartile deviation, so that the degree of skewness is again measured relative to dispersion of the distribution. The coefficient is also a pure number. Its theoretical limit are ± 1 . For symmetrical distribution its value would be 0. Whenever positional averages are called for, skewness should be measured by Bowley's method. Thus this measure is useful in open end distribution and where extreme values are present.

However this measure of skewness and its coefficient do not always give dependable results. In many cases the distribution may not be perfectly symmetrical yet the coefficient may be found to be zero. The reason for such fallacy is that quartiles are based on all the observations of a series. Thus this measure of skewness should be used cautiously and as far as

possible for comparison Karl Pearson coefficient of skewness should be used.

Example: Find Bowley's coefficient of skewness from the following frequency distribution.

x	75	110	125	150	175	200	225	250
y	34	4	48	100	125	80	50	22

Class	Mid Values	cf
X	f	
75	35	35
100	4	75
125	48	123
150	100	223
175	125	348
200	80	428
225	50	478
250	22	500

$$Sk_b = \frac{(Q_3 - Q_1) - 2 \text{ Median}}{(Q_3 - Q_1)}$$

$$Q_1 = \text{Size of } N_{th} / 4 \text{ item} = 500/4 = 125_{th} \text{ item}$$

$$Q_2 = \text{Size of } N_{th} / 2 \text{ item} = 500/2 = 250_{th} \text{ item}$$

$$Q_3 = \text{Size of } 3 \times 500/4 = 375_{th} \text{ item,}$$

$$\text{Thus } Q_1 = 175$$

$$Q_2 = 200$$

$$Sk_b = \frac{200 + 150 - 2(175)}{(200 - 150)} = \frac{0}{50} = 0$$

Example 2: From the following data calculate quartiles and find the coefficient of skewness.

Class	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
Frequency	5	9	14	20	25	15	5	4
Class X	mid values f							cf
0-10	5							5
10-20	9							14
20-30	14							25
30-40	20							48
40-50	25							73
50-60	15							88
60-70	8							96
70-80	7							100

Q_1 = Size of N_{th} /4 item = $100/4 = 25_{th}$ item lies in 30-40 class

Q_2 = Size of N_{th} /2 item = $100/2 = 50_{th}$ item lies in 50-60 class

Q_3 = Size of 3 (N/4) item = $100/4 = 75_{th}$ item, lies in 40-50 class

$$Sk_b = L + \frac{(Q_3 - Q_1) - 2 \text{ Median}}{(Q_3 - Q_1)}$$

$$= 30 + \frac{100/4 - 14}{14} \times 10$$

$$= 30 + \frac{25 - 14}{14} \times 10$$

$$= 30 + 7.86 = 37.86$$

$$Q_2 = L + \frac{N/2 - c.f}{f} \times i$$

$$= 50 + \frac{(100/4) - 48}{25} \times 10$$

$$= 50 + \frac{50 - 48}{25} \times 10 = 50 + .8 = 50.8$$

$$Q_3 = L + \frac{3(N/4) - c.f}{f} \times i$$

$$= 60 + \frac{3(100/4) - 48}{88} \times 10$$

$$= 60 + \frac{75 - 48}{88} \times 10 = 60 + .22 = 60.22$$

$$\text{Coefficient of Skewness} = \frac{60.22 + 37.86 - 2 \times 50.8}{60.22 - 37.86}$$

$$= \frac{98.08 - 101.6}{22.36} = \frac{3.52}{22.36}$$

$$= -0.157$$

Comparison between Pearson and Bowley Measures of Skewness

Complete absence of skewness i.e. symmetry indicated by zero in both the methods. They both are derived from two measures of central tendency expressed as a ratio to measures of variation.

However the results obtained by these two measures are not to be compared with one another especially the numerical value are not related to one another. Sometimes rare occasions with unusually shaped distributions it is possible for them to emerge with opposite signs also.

KURTOSIS

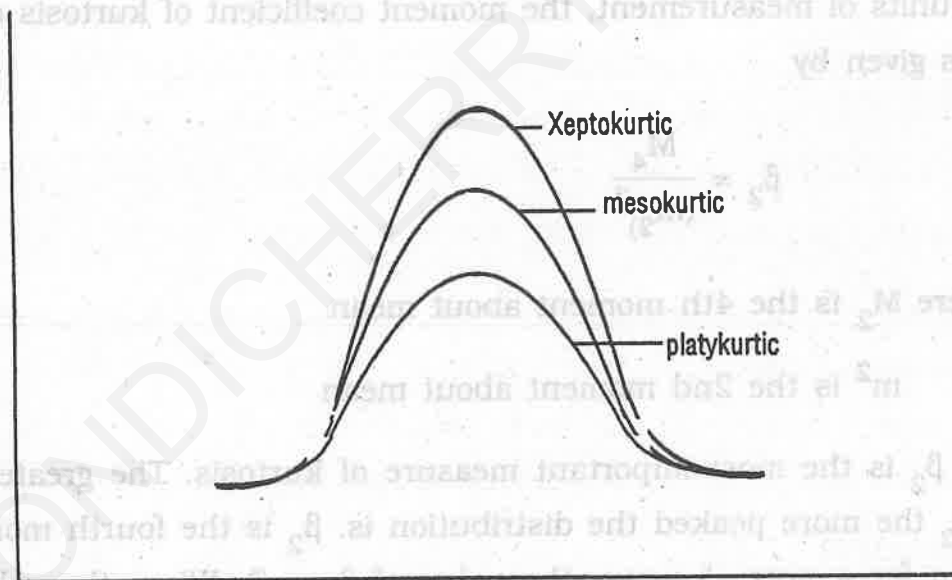
Kurtosis is yet another measure which tells us about the form of distribution. Kurtosis in Greek means bulginess. In statistics Kurtosis refers to the degree of flatness and peakedness in the region about the mode of the frequency curve. The degree of peakedness is measured in relation to normal curve, i.e. it tells us whether the distribution, if plotted on a graph paper would give a curve more flat than normal or more peaked than normal curve. Cowton and Cowden define Kurtosis as a measure that indicate the degree to which a curve of a frequency distribution is peaked or flat topped.

Simpson and Kafa have stated that the degree of Kurtosis of a distribution is measured relative to peakedness of a normal curve.

Spiegel defines Kurtosis as follows: "Kurtosis is the degree of peakedness of a distribution usually taken relative to a normal distribution".

If a curve is more peaked than a normal curve it is called leptokurtic. In such case the items are closely concentrated around mode. On the other hand, if a curve is more flat topped than normal curve, it is called platykurtic. The normal curve itself is known as mesokurtic.

The following diagram illustrates the difference between different curves.



Figure

W.S. Gissel described these curves in a very interesting manner and wrote that, Platykurtic curves are like platypus squat with short tails, leptokurtic curves are like Kangroos high with long tails noted for leaping.

Measures of Kurtosis

The tool designed to measure the height of a frequency distribution are called measures of kurtosis. The degree of kurtosis, present in a given

distribution is measured by 4th moment (M_4) about mean. The fourth moment about mean is defined as the arithmetic mean of the fourth power of deviation of the values from mean. It may be expressed in the symbolic form as -

$$M_4 = \frac{\sum (X_i - \bar{X})^4}{N}$$

The corresponding formula for a frequency distribution may be written as

$$M_4 = \frac{\sum (X_i - \bar{X})^4}{N}$$

However m_4 is not good measure of kurtosis because it is not free of the units of measurement, the moment coefficient of kurtosis denoted by β_2 is given by

$$\beta_2 = \frac{M_4}{(m_2)^2}$$

Where M_2 is the 4th moment about mean

m^2 is the 2nd moment about mean

β_2 is the most important measure of kurtosis. The greater the value of β_2 the more peaked the distribution is. β_2 is the fourth moment about mean for a normal curve, the value of $\beta_2 = 3$. When the value of β_2 is greater than the normal curves value 3 the curve is called a leptokurtic curve and when the value of β_2 is less than 3 the curve is less peaked than a normal curve i.e. platykurtic curve. In a normal curve or mesokurtic curve the value of kurtosis is 3.

Prof. R.A. Fisher has introduced another notation (read as gamma) for measures of the coefficient of kurtosis. γ_2 is defined as

$$\gamma_2 = \beta_2 - 3$$

for a normal curve $\gamma_2 = 0$ if γ_2 is positive the curve is leptokurtic and if γ_2 is negative the curve is platykurtic. Thus the value of β_2 or γ_2 are interpreted as.

- (i) When $\beta_2 < 3$ or γ_2 is negative the curve is leptokurtic.
- (ii) When $\beta_2 > 3$ or γ_2 is positive the curve is platykurtic.
- (iii) When $\beta_2 = 3$ or $\gamma_2 = 0$ the curve is a normal or mesokurtic curve.

Example: Find the kurtosis of the following series.

Class interval	0-10	10-20	20-30	30-40	40-50
Frequency	10	20	40	20	10

Solution:

Calculation of Kurtosis

Class interval	Mid values m	f	$\frac{m - 25}{dx} \cdot 10$	fdx	fdx ²	fdx ³	fdx ⁴
0-10	5	10	-2	-20	40	-80	160
10-20	15	20	-1	-20	20	-20	20
20-30	25	40	0	0	0	0	0
30-40	35	20	1	20	20	20	20
40-50	45	10	2	20	40	80	160

$$\text{Kurtosis} = M_4 / M^3_2$$

$$M_4 = \frac{\sum fd^4 x}{N} - 4 \frac{\sum fdx}{N} \frac{\sum fd^3 x}{N} + 6 \left(\frac{\sum fdx}{N} \right)^2$$

$$= \frac{(\sum fdx^2 x/N) - 3 \left(\frac{\sum fdx}{N} \right)^2}{N}$$

$$= \frac{360}{100} - 4 \frac{0}{100} \times \frac{0}{100} + 6 \left(\frac{0}{100} \right)^2 \frac{120}{100} - 3 \frac{(0)^4}{100}$$

$$= 3.6 - 4 \times 0 \times 0 + 6 \times 0 \times 1.2 - 3(0)^4$$

$$M_2 = \Sigma \frac{fdx^2}{N} - \left[\frac{\Sigma fdx}{n} \right]^2$$

$$= \frac{120}{100} - \left| \frac{0}{100} \right|^2 = 1.2$$

$$\beta_2 = \frac{M_4}{(M_2)^2} = \frac{3.6}{(1.2)^2} = \frac{3.6}{1.44} = 0.25$$

The curve is a platykurtic one.

Example: The following data are given to you. Find out whether the distribution is platykurtic.

$$N = 100 \quad \Sigma fdx = 50 \quad \Sigma fdx^2 = 1967.2 \quad \Sigma fdx^3 = 2925.8$$

$$\Sigma fdx^4 = 86650.$$

Solution: To study whether the distribution is platykurtic or not we have to calculate:

$$\beta_2 = \frac{M_4}{M_2^2}$$

$$M_4 = \frac{\Sigma fd^4 x}{N} - 4 \frac{\Sigma fdx}{N} \frac{\Sigma fd^3 x}{N} + 6 \frac{\Sigma fdx^2}{N} \frac{\Sigma fd^2 x}{N} - 3 \left(\frac{\Sigma fdx}{N} \right)^4$$

$$= \frac{86650.2}{100} - 4 \frac{50}{100} \times \frac{2925.8}{100} + 6 \left| \frac{50}{100} \right|^2 \times$$

$$\frac{1967.2}{100} - 3(50/100)^4$$

$$= 866.502 - 4 \times 5 \times 29.258 + (6 \times (5)^2 \times 19.672) - 3 \times (5)^4$$

$$= 866.502 - 58.516 + 29.508 - 0.1875 = 837.3065$$

$$M_2 = \frac{\Sigma fd^2 x}{N} - \left| \frac{\Sigma fdx}{N} \right|^2$$

$$= \frac{1967.2}{.100} - \left| \frac{50}{100} \right|^2$$

$$= 19.672 - 5 (2)^2 = 19.422$$

$$\text{therefore } \beta_2 = \frac{m_4}{m_2^2} = \frac{837.3065}{(19.422)^2} = 2.22$$

As the value of β_2 is less than 3, therefore the distribution is platykurtic.

Now that we have discussed dispersion, Skewness, Kurtosis it will not be out of place to compare and contrast them as all these measures are the measures to study the formation of a frequency distribution.

Dispersion studies the deviation of the items from the central value i.e. the scatteredness of the items round a central value or among themselves but it does not speak anything about which side the deviation are clustered more. They are clustered whether below or above average. However measures of skewness studies this point. They tell us about the cluster of deviations above and below a measure of central tendency. The deviation below and above the measure of central tendency in a normal distribution are equal but not in the case of asymmetrical or skewed distribution. Kurtosis, however, studies the concentration of items at central part of a series. If items are too much in centre, the curve becomes leptokurtic and if the concentration in the centre is comparatively little the curve becomes platykurtic.

Thus we find that measures of dispersion skewness and kurtosis study three different aspects of a frequency distribution. The span within which the value of a variable lies is focused by measure of dispersion. Measures of skewness throw light on the shape of a series and the size of variation on either side of a central value. Kurtosis studies the frequencies of a series at the central value.

UNIT - V

CORRELATION ANALYSIS

- ❑ INTRODUCTION
- ❑ MEANING AND DEFINITION
- ❑ USES OF CORRELATION
- ❑ TYPES OF CORRELATION
- ❑ METHODS OF STUDYING CORRELATION

INTRODUCTION

So far we have confined ourselves to variation of a single variable. In practice we come across a large number of problems involving the use of two or more than two variables. In daily life we are not only concerned in studying the characteristics of the variable in isolation but also jointly in order to examine the relationship of one variable with another variable. Problems of this type which involves functional relationship of one variable with that of another is studied by the method of correlation. For example the relationship between the age of husband and age of wife, price of a commodity and the amount demanded, heights and weights of a group of persons, income and expenditure of a group of persons, etc. are studied with the help of correlation analysis. The measure of correlation called correlation coefficient expresses the degree and direction of the relationship. Thus in correlation analysis we study the pattern of relationship between two variables as well as the closeness with which two variables vary.

MEANING AND DEFINITION

The term correlation (or covariation) indicates the relationship between two such variables in which with change in the values of one variable, the values of the other variable also changes. Correlation indicates the relationship between two variables so that movements in one variable tend to be accomplished by movement in the second variable.

According to Croxton and Cowden "when the relationship is of a quantitative nature, the appropriate statistical tool for discovering and expressing it in a brief formula is known as correlation".

Ya Lun Chau defines correlation as "an attempt to determine the degree of relationship between variables".

According to L.R. Connor, "If two or more quantities vary in sympathy so that movements in the one tend to be accompanied by corresponding movements in the other(s) then they are said to be correlated".

The above definitions make it clear that the term correlation refers to the study of relationship between two or more variables and it can be defined as the amount of similarities, direction and degree of variation in corresponding pairs of observations of two variables.

USES OF CORRELATION

The study of correlation is of immense use in practical life because of the reasons as follows:

1. With the help of correlation analysis one can measure the degree of relationship that exists between the variables in one figure.
2. From one variable we can estimate the other variable by the help of regression analysis only when we establish the variables are related.
3. Correlation analysis is very helpful in understanding the economic behaviour. It helps us in locating such variables on which the other variables depend. It aids in locating the critically important variables on which others depend; for example we can find out the factors responsible for price rise or low productivity.
4. Correlation study helps us in identifying such factors which can stabilize a disturbed economic situation.
5. Interrelationship studies between different variables are helpful tools in promoting research.
6. The effect of correlation is to reduce the range of uncertainty in the decision making effort. In social sciences, particularly in business

world, forecasting is an important phenomenon and correlation studies help us to make relatively more dependable forecast.

Thus correlation studies are very widely used as the basic tool for analysis and interpretation of statistical data relative to two or more variables. However, though the word correlation is used in the sense of mutual dependence of two or more variables, yet it is not at all necessary that it should be always so. Even a very high degree of correlation between two variables does not necessarily indicate a cause and effect relationship between them. The correlation may be any one or a combination of the following reasons:

1. The correlation may be due to pure chance especially in a small sample. We may get a high degree of correlation between two variables in a sample but in the universe there may not be any relationship between the variables at all.
2. Both the correlated variables may be influenced by one or more other variables. For example when prices of rice and jute are increasing we may find high degree of correlation. In reality they are not related or neither of them are cause or effect. It may be due to their production which is affected by rainfall.
3. Both the variables may be mutually influencing each other so that neither can be designated as the cause and the other effect. For example, the demand of a commodity may go down as a result of rise in price. It may also be so that demand of the commodity has gone up due to anticipated shortage in future and has resulted in price rise.
4. There might be a situation of nonsense or spurious correlation between the two variables under study. One may find high degree of correlation between number of divorces per year and exports of television sets. Obviously there cannot be any relationship between number of divorces and television export. It should be understood that correlation is a relationship between related variables only.

The above points make it clear that correlation is only a mathematical relationship and it does not necessarily signify a cause and effect relationship between the variables. Further it should also be remembered that existence of correlation is no guarantee that the relationship would always be of the same type.

TYPES OF CORRELATION

Correlation may be classified into three categories, namely,

- 1) Positive or negative
- 2) Simple, multiple and partial
- 3) Linear or non-linear

Positive or Negative Correlation

It is otherwise called as direct (positive) or inverse (negative) correlation. The correlation between two variables is said to be direct or positive when they move in the same direction so that an increase or decrease in the value of one variable is associated with increase or decrease in the value of the other. On the other hand if the variables move in the opposite direction i.e., increase in one variable is associated with decrease in the other and vice versa the correlation is said to be inverse or negative. The following example would illustrate the difference between positive and negative correlation.

Positive Correlation

	(1)					(2)			
X	10	15	20	25	X	17	13	7	5
Y	29	35	40	49	Y	34	22	15	7

Negative Correlation

	(1)					(2)			
X	20	30	40	50	X	60	40	30	20
Y	50	40	35	20	Y	100	120	125	137

Simple, Partial and Multiple Correlation

The distinction between simple, partial and multiple correlation is based upon the number of variables studied. In simple correlation we study only two variables. When three or more variables are studied it is a case of multiple correlation. In partial correlation though more than two factors are involved, correlation is studied only between two factors and the other two factors are assumed to be constant.

Linear and Non-Linear Correlation

There exists a linear correlative between two variables if the amount of change in one variable tends to bear a constant ratio to the amount of change in the second variable. The distinction between linear and non linear correlation is based upon the constancy of the ratio of change between the variables. Thus if with 10% increase in demand there is 20% increase in price there is a linear relationship between the variables. This relationship is of the type $Y = a + bx$ (straight line equation). If the corresponding values of two such series are plotted in a graph paper a straight line would be obtained. On the other hand the correlation between two variables is said to be non-linear (curvilinear) when the ratio of variation in related variables is fluctuating.

The following example illustrates linear and non-linear correlations.

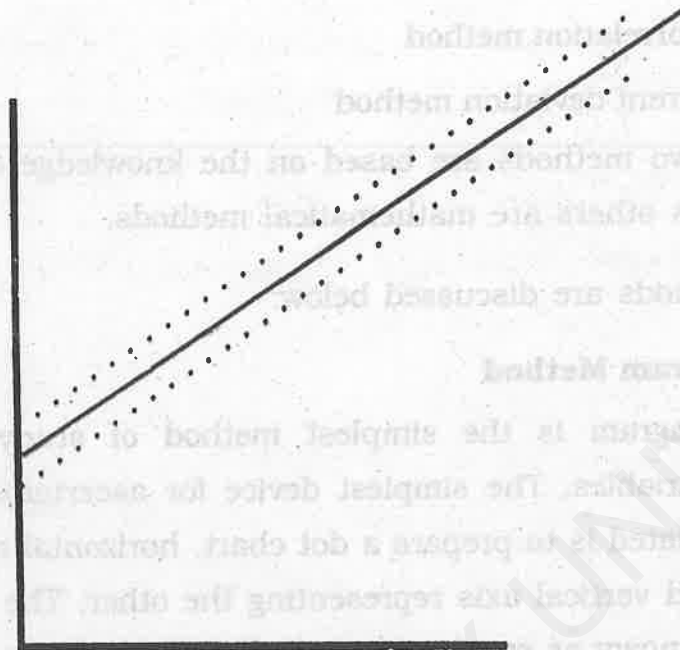
Linear Correlation

X	2	4	6	8	10
Y	70	140	210	280	350

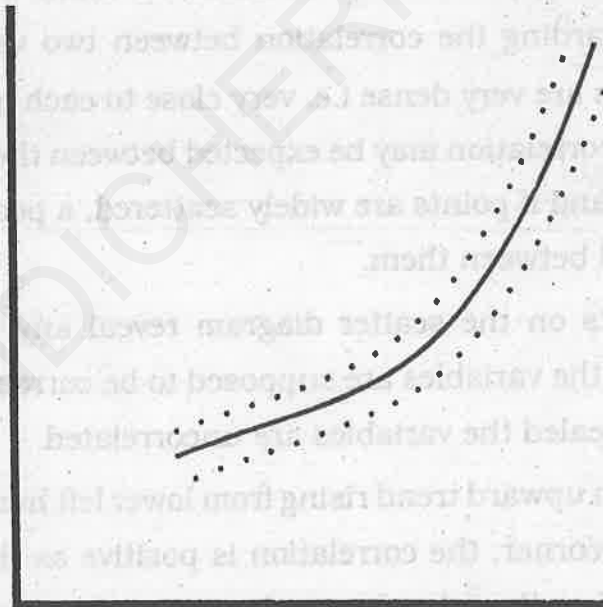
Non-Linear Correlation

X	2	4	6	8	10
Y	5	8	12	15	25

The following two diagrams also will illustrate the difference between linear and non-linear (curvilinear) correlation.



Graph: 5.1 Positive Linear Correlation



Graph: 5.2 Linear Correlation

METHODS OF STUDYING CORRELATION

The various methods by which correlation studies are made are as follows:

- i. Scatter diagram method
- ii. Graphic method

- iii. Karl Pearson's coefficient of correlation method
- iv. Rank correlation method
- v. Concurrent deviation method

The first two methods are based on the knowledge of diagrams and graphs, whereas others are mathematical methods.

These methods are discussed below:

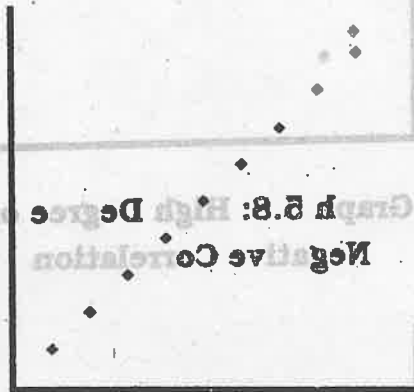
1. Scatter Diagram Method

The scatter diagram is the simplest method of studying relationship between two variables. The simplest device for ascertaining whether two variables are related is to prepare a dot chart, horizontal axis representing one variable and vertical axis representing the other. The diagram of dots so obtained is known as scatter diagram. From the scatter diagram we can form a fairly good, though rough idea about the relationship between two variables. The following points may be borne in mind in interpreting the scatter diagram regarding the correlation between two variables.

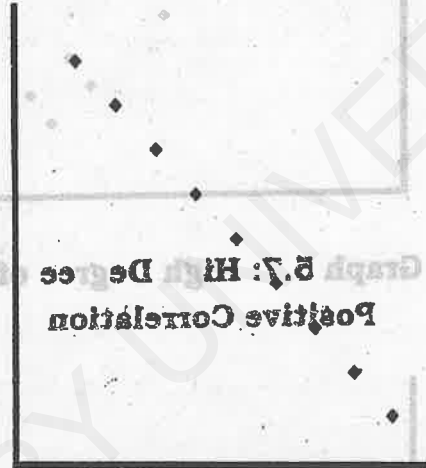
- (i) If the points are very dense i.e. very close to each other, a fairly good amount of correlation may be expected between the two variables; on the other hand if points are widely scattered, a poor correlation may be expected between them.
- (ii) If the points on the scatter diagram reveal any trend (upward or downward) the variables are supposed to be correlated whereas if no trend is revealed the variables are uncorrelated.
- (iii) If there is an upward trend rising from lower left hand corner to upper right hand corner, the correlation is positive as they move in same direction. If on the other hand, the points depict a downward trend from upper left hand corner to lower right hand corner the correlation is negative, as in this case the variables move in the opposite direction.
- (iv) In particular, if all the points lie on a straight line starting from the left bottom and going up towards the right top, the correlation is perfect and positive, and if all the points lie on a straight line starting

left top and coming down to right bottom, the correlation is perfect and negative.

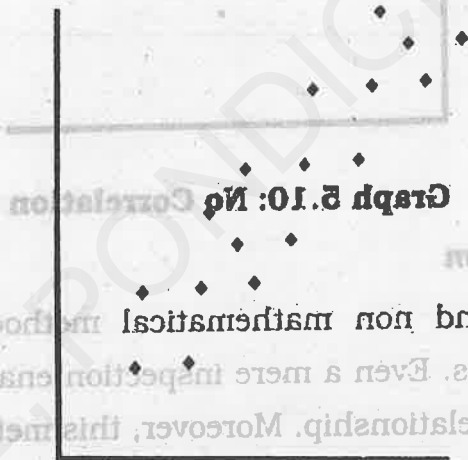
The following diagrams of the scattered data depict different types of correlation.



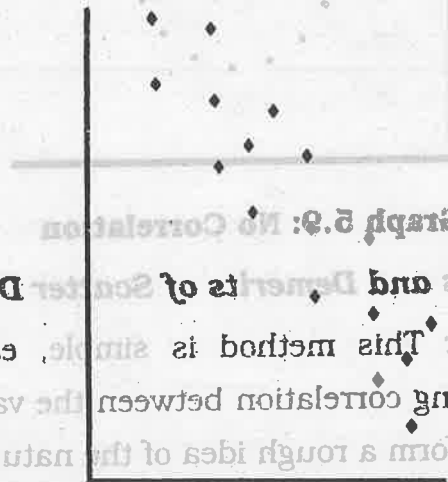
Graph 5.3: Perfect Positive Linear Correlation



Graph 5.4: Perfect Negative Linear Correlation

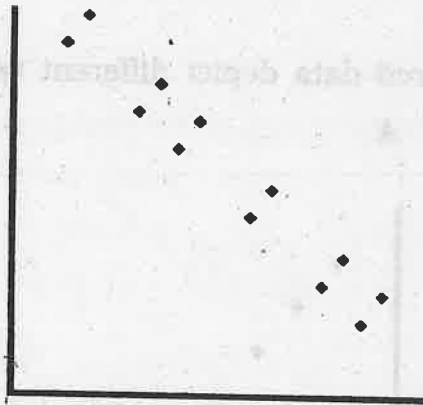


Graph 5.5: Low Degree of Positive Correlation

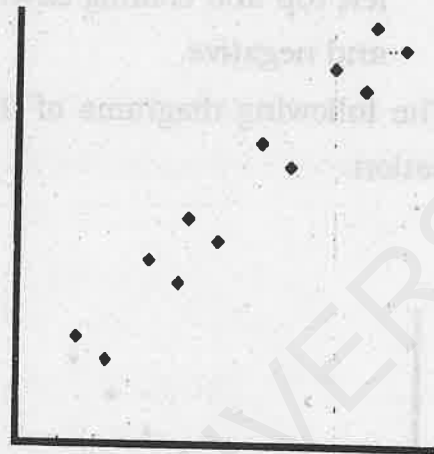


Graph 5.6: Low Degree of Negative Correlation

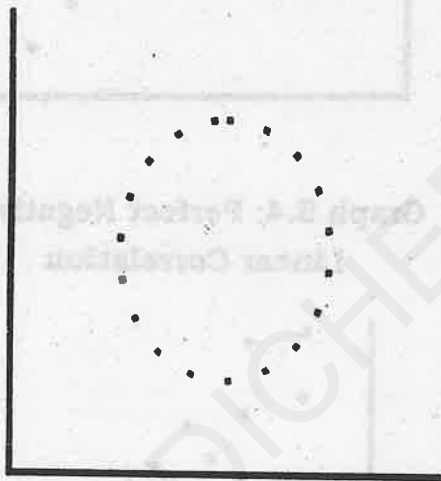
Demerits: The major limitation of the method is that it gives a visual picture of the relationship. It may tell about the nature of relationship but it fails



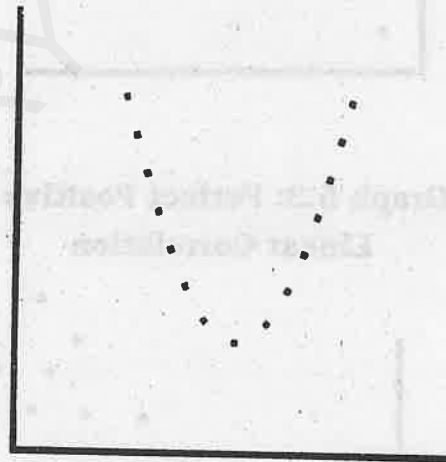
Graph 5.7: High Degree of Positive Correlation



Graph 5.8: High Degree of Negative Correlation



Graph 5.9: No Correlation



Graph 5.10: No Correlation

Merits and Demerits of Scatter Diagram

Merits: This method is simple, easy and non mathematical method of studying correlation between the variables. Even a mere inspection enables us to form a rough idea of the nature of relationship. Moreover, this method is not affected by extreme observations unlike the mathematical formula of ascertaining correlation. Drawing a scatter diagram usually is the first step in investigating the relationship between the variables.

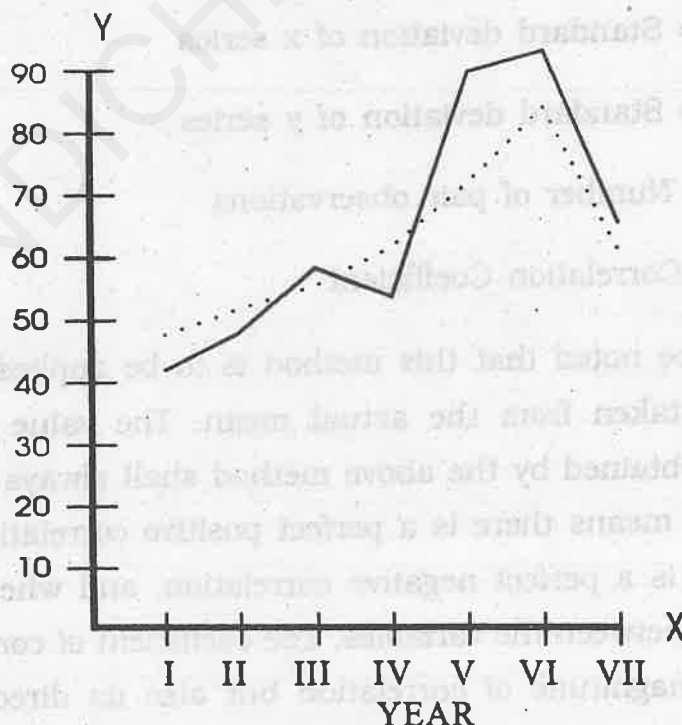
Demerits: The major limitation of the method is that it gives a visual picture of the relationship. It may tell about the nature of relationship but it fails

to tell about the extent or exact degree of relationship. This method also is not amenable to any further mathematical treatment.

2. Graphic Method

It is the simplest method of ascertaining the presence of correlation between two variables. In this method individual values of the two variables are plotted on the graph paper. By examining the direction and closeness of the two curves so drawn we can infer whether or not the variables are related. If both the curves move in one direction, there is correlation, if the curves do not move in the same direction there is no correlation. The following figure shows how the variables can be plotted in the graph paper from a given data.

Year	I	II	III	IV	V	VI	VII
X	42	45	58	55	89	90	66
Y	46	49	54	59	65	76	60



Graph 5.11

Merits and Demerits of Graphic Method

Like scatter diagram method, it is possible to ascertain the nature and extent of correlation from a graph easily. But the exact degree of correlation cannot be ascertained like from a scatter diagram. This method is generally used where data for a period of time is available and where the exact degree of correlation is not required.

3. Karl Pearson's Coefficient of Correlation

It is a mathematical method for measuring the intensity or magnitude of relationship between two series of variables. Of the several mathematical methods of measuring correlation, Karl Pearson's coefficient of correlation is most widely used in practice. The Pearson's coefficient of correlation is denoted by 'r'. The formula for computing 'r' is

$$r = \frac{\sum xy}{N \sigma x \sigma y}$$

where $X = (X - \bar{X})$

$Y = (Y - \bar{Y})$

σx = Standard deviation of x series

σy = Standard deviation of y series

N = Number of pair observations

r = Correlation Coefficient

It should be noted that this method is to be applied only where the deviations are taken from the actual mean. The value of coefficient of correlation as obtained by the above method shall always lie between ± 1 . When $r = +1$ it means there is a perfect positive correlation, when $r = -1$ it means there is a perfect negative correlation, and when $r = 0$ there is no relationship between the variables. The coefficient of correlation not only describes the magnitude of correlation but also its direction. The above formula can be transformed into another formula where standard deviations of the two series need not be calculated and which is easier to apply.

Symbolically
$$= \frac{\Sigma xy}{N \sigma_x \sigma_y}$$

$$\text{or } r = \frac{\sum xy}{N \sqrt{\frac{\sum x^2}{N}} \sqrt{\frac{\sum y^2}{N}}}$$

where $X = (X - X')$

$$N - Y = Y$$

$$Y = (Y - Y')$$

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$

We can illustrate the application of the formula in the following example.

Illustration: Calculate Karl Pearson's Coefficient of correlation from the following data.

Roll No. of the students	1	2	3	4	5	6	7	8
Marks in Statistics	65	66	67	67	68	69	70	72
Marks in Accountancy	67	68	65	68	72	72	69	71

Solution: Let marks in Statistics be denoted by x and Accountancy by y .

[illegible]

$$r = \frac{\Sigma xy}{N \sigma_x \sigma_y} \quad (1st \text{ formula})$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} \quad (2nd \text{ formula})$$

$$x' = (X - \bar{X}) \quad y' = (Y - \bar{Y})$$

$$\bar{x} = \frac{\Sigma x}{N} = \frac{552}{8} = 68 \quad \bar{y} = \frac{\Sigma y}{N} = \frac{552}{8} = 69$$

Calculation of r by first formula:

$$\sigma_x = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{36}{8}} = \sqrt{4.5} = 2.121$$

$$\sigma_y = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{44}{8}} = \sqrt{5.5} = 2.345$$

$$r = \frac{24}{8 \times 2.121 \times 2.345} = + 0.603$$

Calculation of r by second formula:

$$r = \frac{24}{\sqrt{36 \times 44}} = + 0.603$$

Direct Method of calculating correlation

The coefficient of correlation can be calculated directly without finding out the deviation of various values from the mean of the series with the help of following formula:

$$r = \frac{\Sigma XY - \Sigma X \Sigma Y / N}{\left[\Sigma X^2 - (\Sigma X)^2 / N \right] \left[\Sigma Y^2 - (\Sigma Y)^2 / N \right]}$$

$$\text{or} = \frac{N \Sigma xy - \Sigma x \Sigma y}{\sqrt{N \Sigma x^2 - (\Sigma x)^2} \sqrt{N \Sigma y^2 - (\Sigma y)^2 / N}}$$

The above problem can be solved as follows:

R.N.	X	X ²	Y	Y ²	XY
1	65	4225	67	4489	4355
2	66	4356	68	4624	4488
3	67	4489	65	4225	4355
4	67	4489	68	4624	4556
5	68	4624	72	5184	4896
6	69	4761	72	5184	4868
7	70	4900	69	4761	4830
8	72	5184	71	5041	5112

$$\Sigma X = 544 \quad \Sigma X^2 = 37028 \quad \Sigma Y = 552 \quad \Sigma Y^2 = 38132 \quad \Sigma XY = 37560$$

$$r = \frac{8 \times 37560 - (544 \times 552)}{\sqrt{8(37028) - (544)^2} \times \sqrt{8(38132) - (552)^2}}$$

$$r = \frac{300480 - 300288}{\sqrt{296224 - 295936} \times \sqrt{305056 - 304704}}$$

$$r = \frac{192}{16.97 \times 18.76}$$

$$r = \frac{192}{318.35} = +0.603$$

When deviations are taken from assumed mean

When the figures of both the series of variables are big as well as their actual means are in fraction, the calculation of correlation by the methods discussed above would involve too many calculations and the calculation will be tedious. In such case, to make the calculation easier we can make use of assumed mean and modify the formula accordingly for finding out correlation.

The modified formula for calculation of correlation coefficient is as follows:

$$r = \frac{\sum dxdy - (\sum dx)(\sum dy)}{n \sqrt{\sum dx^2 - \frac{(\sum dx)^2}{N}} \sqrt{\sum dy^2 - \frac{(\sum dy)^2}{N}}}$$

where $\sum dx$ and $\sum dy$ are the sum of deviations of the x and y series respectively from their assumed means. $\sum dx^2$ and $\sum dy^2$ are the sum of the squares of deviation of x and y series respectively from their assumed means. $\sum dxdy$ is the sum of the product of corresponding deviations of x and y series from their assumed means.

The following example will illustrate the use of above formula.

Illustration: Calculate the coefficient of correlation for the ages of husband and wife.

Age of husband	24	28	29	30	31	32	34	36	37	40
Age of wife	19	23	24	25	26	27	29	30	31	33

x	y	dx (x - 32)	dy (y - 26)	dx ²	dy ²	dxdy
24	19	-8	-7	64	49	56
28	23	-4	-3	16	9	12
29	24	-3	-2	9	4	6
30	25	-2	-1	4	1	2
31	26	-1	0	1	0	0
32	27	0	1	0	1	0
34	29	2	3	4	9	6
36	30	4	4	16	16	16
37	31	5	5	25	25	25
40	34	8	7	64	49	56

$\sum x = 321$	$\sum y = 267$	$\sum dx = 1$	$\sum dy = 7$	$\sum dx^2 =$	$\sum dy^2 =$	$\sum dxdy =$
				203	163	179

Note: This formula is the one discussed above for assumed mean. The deviations are taken from the assumed mean. The deviations are also multiplied by the frequencies.

$$r = \frac{\sum dx dy}{\sqrt{\sum dx^2 - \frac{(\sum dx)^2}{N}}} \times \frac{\sum dy}{\sqrt{\sum dy^2 - \frac{(\sum dy)^2}{N}}}$$

Steps:

$$r = \frac{179 - \frac{1 \times 7}{10}}{\sqrt{203 - \frac{(1)^2}{10}}} \times \frac{7}{\sqrt{163 - \frac{(7)^2}{10}}}$$

$$r = \frac{179 - 0.7}{14.24 \times 12.57}$$

$$= \frac{178.3}{178.99}$$

$$= +0.996$$

Correlation of Grouped Data

When the number of observations is large the data are often classified into two way frequency distribution, otherwise called as bivariate frequency distribution since it shows the frequency distribution of two related variables. The class intervals of y are listed in the column headings and those of x are listed in rows of the table. The frequencies for each cell of the table are determined by tallying the frequency of the distribution of a single variable.

The formula for calculating the coefficient of correlation in such case is:

$$r = \frac{N \sum f dx dy - \sum f dx \sum f dy}{\sqrt{N \sum f dx^2 - (\sum f dx)^2} \times \sqrt{N \sum f dy^2 - (\sum f dy)^2}}$$

(OR)

$$r = \sum f dx dy - \frac{\sum f dx}{N} \times \frac{\sum f dy}{N}$$

Note: This formula is the same as the one discussed above for assumed mean. The only difference is that here the deviations are also multiplied by the frequencies.

Steps:

- (i) Take the step deviations of x and y series from assumed means and denote them by dx and dy.
- (ii) Multiply dx and dy and the frequency of the cell. The value fdx dy may be put at the top on right or left hand side of the cell.
- (iii) Add all values of fdx dy as calculated in step (ii) and thus obtain the value of $\Sigma fdx dy$.
- (iv) Multiply the frequency of x series with the step deviation and total all such product to get Σfdx . Similarly get Σfdy .
- (v) Take the square of the step deviations of x series and multiply them with the frequency and total the products to get Σfdx^2 , similarly get Σfdy^2 .
- (vi) Substitute the values so obtained in the formula given above to get the value of r.

The following example would clarify the above points.

Illustration: From the following table given below calculate the coefficient of correlation between family income and food expenditure of 100 families.

Food Expenditure in %	Family Income				
	2000-3000	3000-4000	4000-5000	5000-6000	6000-7000
10 - 15	—	—	—	03	07
15 - 20	—	04	09	04	03
20 - 25	07	06	12	05	—
25 - 30	03	10	19	08	—

Bivariate Table

$$r = \frac{\sum fxdy - \frac{\sum fdx \sum fdy}{N}}{\sqrt{\sum fdx^2 - \frac{(\sum fdx)^2}{N}}} \sqrt{\sum fdy^2 - \frac{(\sum fdy)^2}{N}}$$

$$r = \frac{-48 - 0 \times 100}{\sqrt{120 - \frac{(0)^2}{100}}} \sqrt{200 - \frac{(100)^2}{100}}$$

$$r = \frac{-48}{10.95 \times 10} = \frac{48}{109.5}$$

$$= -0.4363$$

Merits and Demerits of Karl Pearson's Coefficient of Correlation

Karl Pearson's method is the most popular method for measuring the degree of relationship among all the mathematical methods used for the same. The chief merit of this coefficient is that it gives the degree of the relationship among the variables as well as the direction of the correlation.

However it suffers from some limitations also. These are:

1. The correlation coefficient always assumes linear relationship even though it may not be there.
2. It is liable to be misinterpreted as a high degree of correlation does not necessarily mean very close relationship. Thus great care should be exercised in interpreting the values.
3. The values of the coefficient is unduly affected by the extreme items.
4. As compared to other methods it is tedious to calculate.

4. Rank Correlation Method

The Karl Pearson's coefficient of correlation as discussed above can not be used in cases where the direct quantitative measurement of the phenomenon under study is not possible, for example efficiency, honesty intelligence etc. In such cases it may be possible to arrange various items of a series in serial order but the quantitative measurement of their value is difficult. There are many such attributes which are incapable of

quantitative measurement. If it is desired to have a study of association between two such attributes, say intelligence and beauty, the Karl Pearson's coefficient of correlation cannot be calculated as these attributes cannot be assigned definite values. In such cases one can rank or array the different items and apply Spearman's rank correlation method for finding out the degree of correlation. The formula for computing Spearman's rank correlation is

$$r_s = 1 - \frac{6 \sum d^2}{N(N^2 - 1)} \text{ or}$$

$$r_s = 1 - \frac{6 \sum d^2}{N^3 - N}$$

where r_s denotes the Spearman's rank correlation

d denotes the difference of the ranks of the paired attributes of a single item.

N stands for the number of pairs.

The values of this coefficient, interpreted in the same way as Karl Pearson's coefficient of correlation ranges between +1 to -1 when r_s is +1 it indicates complete agreement in the order of ranks between the two attributes. When r_s is -1 it indicates complete disagreement in the order of the ranks as they are in opposite direction. This shall be clear from the following example:

R_1	R_2	d ($R_1 - R_2$)	d^2
1	1	0	0
2	2	0	0
3	3	0	0
			$\sum d^2 = 0$

$$r_s = 1 - \frac{6 \sum d^2}{N^3 - N}$$

$$r_s = 1 - \frac{6 \times 0}{3^3 - 3} = 1 - 0 = 1$$

R_1	R_2	d ($R_1 - R_2$)	d^2
1	3	-2	4
2	2	0	0
3	1	2	4
			$\Sigma d^2 = 8$

$$r = 1 - \frac{6 \Sigma d^2}{N^3 - N}$$

$$r = 1 - \frac{6 \times 8}{3^3 - 3} = 1 - 2 = -1$$

There are two types of problems of calculating this coefficient:

- (a) when actual ranks are given
- (b) when actual ranks are not given

In each of these two types of problem a difficulty arises when ranks of two individuals are the same. Such problems need a modification in the formula given above.

(a) When actual ranks are given:

In this situation the following steps are involved.

- (i) Compute the difference of ranks ($R_1 - R_2$) and denote them by d .
- (ii) Compute d and total them to get Σd^2
- (iii) Use the formula given below to get the coefficient

$$r_s = 1 - \frac{6 \Sigma d^2}{N^3 - N}$$

The following example will illustrate the above method.

Illustration: The ranking of ten students in Statistics and Economics are as follows:

Statistics	3	5	8	4	7	10	2	1	6	9
Economics	6	4	9	8	1	2	3	10	5	7

Use Spearman's formula to find out the rank correlation coefficient.

Solution:

R_1 Statistics	R_2 Economics	d Rank difference	d^2 Square of rank difference
3	6	-3	9
5	4	+1	1
8	9	-1	1
4	8	-4	16
7	1	+6	36
10	2	+8	64
2	3	-1	1
1	10	-9	81
6	5	+1	1
9	7	+2	4
			$\Sigma d^2 = 214$

$$r_s = 1 - \frac{6 \Sigma d^2}{N^3 - N}$$

$$r_s = 1 - \frac{6(214)}{10^3 - 10}$$

$$r_s = 1 - \frac{1284}{10(99)} = 1 - 1.3 = -0.3$$

Illustration: Ten competitors in a beauty contest were ranked by three judges in the following order.

First Judge	1	6	5	9	2	3	4	10	7	8
Second Judge	3	5	8	7	4	9	2	1	6	10
Third Judge	6	4	8	9	1	3	2	10	7	5

Use the method of rank correlation to determine which pair of judges has the nearest approach to common tastes in beauty.

Solution: With a view to find out which pair of judges have the nearest approach to common taste in beauty, we will compare the rank correlation between the judgements of

- (i) 1st and 2nd judge
- (ii) 1st and 3rd judge
- (iii) 2nd and 3rd judge

The ranks given by the three judges would be denoted by R_1 , R_2 and R_3 .

R_1	R_2	R_3	$R_1 - R_2$ $d_{1,2}$	$R_1 - R_3$ $d_{1,3}$	$R_2 - R_3$ $d_{2,3}$	d^2_{12}	d^2_{13}	d^2_{23}
1	3	6	-2	-5	-3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	8	-3	-3	0	9	9	0
9	7	9	2	0	-2	4	0	4
2	4	1	-2	1	3	4	1	9
3	9	3	-6	0	6	36	0	36
4	2	2	2	2	0	4	4	0
10	1	10	9	0	-9	81	0	81
7	6	7	1	0	-1	1	0	1
8	10	5	-2	3	5	4	9	25
						<u>148</u>	<u>54</u>	<u>166</u>

We have $n = 10$

Spearman's rank correlation coefficient are given by

$$\begin{aligned} r_{1,2} &= 1 - \frac{6 \sum d^2}{N^3 - N} \\ &= 1 - \frac{6 \times 148}{10^3 - 10} \\ &= 1 - \frac{888}{990} = 1 - 0.896 = 0.106 \end{aligned}$$

$$\begin{aligned} r_{1,3} &= 1 - \frac{6 \sum d^2}{N^3 - N} \\ &= 1 - \frac{6 \times 54}{10^3 - 10} \\ &= 1 - \frac{324}{990} = 1 - 0.327 = 0.673 \end{aligned}$$

$$\begin{aligned} r_{2,3} &= 1 - \frac{6 \sum d^2}{N^3 - N} \\ &= 1 - \frac{6 \times 166}{10^3 - 10} \\ &= 1 - \frac{996}{990} = 1 - 1.006 = -0.006 \end{aligned}$$

Since $r_{1,3}$ is the highest the pair of 1st and 3rd judges has the nearest approach to common taste in beauty.

(b) When ranks are not given

Rank formula can be used even if we are dealing with variables which can be measured quantitatively. If we are given the actual data (not the ranks) we have to convert the data into ranks. The highest (or smallest) value is given rank 1 i.e., either ranking is to be done by descending or ascending order. Whatever may be the order of ranking it has to be uniformly followed in case of both the variables.

Illustration: Calculate Spearman's rank correlation coefficient between advertisement cost and sales from the following data:

Advertisement cost (in '000 Rs)	39	68	65	92	85	76	27	98	36	79
Sales (in lakhs)	47	59	59	86	63	68	60	91	51	84

Solution: Let x denote the advertisement cost ('000 Rs) and y denote the sales (lakhs).

x	y	R_x	R_y	d	d^2
39	47	8	10	-2	4
68	54	6	8	-2	4
65	59	7	7	0	0
92	86	2	2	0	0
85	63	3	5	-2	4
76	68	5	4	1	1
27	60	10	6	4	16
98	91	1	1	0	0
36	51	9	9	0	0
79	84	4	3	1	1
				$\Sigma d = 0$	$\Sigma d^2 = 30$

Here $n = 10$

Therefore

$$\begin{aligned}
 r_s &= 1 - \frac{6 \Sigma d^2}{N^3 - N} \\
 &= 1 - \frac{6 \times 30}{10^3 - 10} \\
 &= 1 - \frac{180}{990} = 1 - 0.181 = 0.819
 \end{aligned}$$

Equal Ranks

In some cases it may be found necessary to rank two or more individuals or entries as equal. In such cases common ranks are assigned to the repeated items. This common ranks are the arithmetic mean of ranks which these items would have got if they were different from each other and the next will get the rank next to rank used in computing the common rank. Thus if two individuals are ranked equal at fifth place, they are each given the rank $5 + 6/2$, that is 5.5, if 4 are ranked equal at 6th place they are given the rank $6 + 7 + 8 + 9/4 = 30/4 = 7.5$ and the next rank would be 10th where equal ranks are assigned to some entries, an adjustment in the above formula is made calculating rank correlation coefficient.

The adjustment consists of adding $1/12 (m^3 - m)$ to the value of $\sum d^2$ where m stands for the number of items whose ranks are common. If there are more than one such group of items in common, this value is added as many times as the number of such groups. The formula can thus be written as

$$r_s = 1 - \frac{6 \left\{ \sum d^2 + \frac{1}{12} (m^3 - m) + \frac{1}{12} (m^3 - m) + \dots \right\}}{N^3 - N}$$

Illustration: A psychologist wanted to compare two methods A and B of teaching. He selected a random sample of 22 students. He grouped them into 11 pairs so that the students in a pair here approximately equal scores on an intelligence test. In each pair one student was taught by method A and the other by method B and examined after the course. The marks obtained by them are tabulated below.

Pair	1	2	3	4	5	6	7	8	9	10	11
A	24	29	19	14	30	19	27	30	20	28	11
B	37	35	16	26	23	27	19	20	16	11	21

Find the rank correlation coefficient.

Solution:

A	B	Rank of A	Rank of B	d	d ²
24	37	6	1	5	25
29	35	3	2	1	1
19	16	8.5	9.5	-1	1
14	26	10	4	6	36
30	23	1.5	5	-3.5	12.25
19	27	8.5	3	5.5	30.25
27	19	5	8	-3	9.00
30	20	1.5	7	-5.5	30.25
20	16	7	9.5	-2.5	6.25
28	11	4	11	-7	49.00
11	21	11	6	5	25.00
				$\Sigma d = 0$	$\Sigma d^2 = 225$

In the A series the value 30 occurs twice. The common rank assigned to each of those values is 1.5 the arithmetic mean of 1 and 2, the ranks which these observations would have taken if they were different. The next value 29 gets the next rank ie. 3. Again the value 19 occurs twice. The common rank assigned to it is 8.5. The arithmetic mean of 8 and 9 and the next value 14 gets the rank 10. Similarly in B series the value 16 occurs twice and the common rank assigned to each is 9.5, the arithmetic mean of 9 and 10. The next value 11 gets the rank 11.

Hence we see that in the A series item 19, 30 are repeated, each occurs twice and in B series the item 16 is repeated. Thus in each of the three cases $m = 2$. Hence on applying the correction factor $m^3 - m/12$ for each repeated items we get

$$r_s = 1 - \frac{6 \left[\Sigma d^2 + \frac{2^3 - 2}{12} + \frac{2^3 - 2^3 - 2}{12} + \frac{2^3 - 2}{12} \right]}{11^3 - 11}$$

$$= 1 - \frac{6 \times 226.5}{11 \times 120}$$

$$= 1 - 1.0225 = -0.0225.$$

Merits and Demerits of the Rank Method

Merits:

1. This method is very simple to calculate and to understand.
2. Where the data are of a qualitative nature like beauty, honesty, intelligence etc. this method can be employed usefully.
3. When the ranks of different item values in the variables only are given this is the only method for finding out degree of correlation.
4. If it is desired to use this formula when actual values are given ranks can be ascertained and correlation can be found out.
5. Since in this method Σd or sum of the differences provides a check on calculation.
6. It can be interpreted in the same way like Karl Pearson's coefficient.

Demerits:

1. This method cannot be used for finding out correlation in a grouped frequency distribution.
2. It can be conveniently used only when n is small say 30 or less. If it exceeds 30 the calculations becomes quite tedious and require a lot of time.
3. As all the information concerning the variables is not utilised this method lacks precision as compared to Pearson's method.

5. Concurrent Deviation Method

This method is one of the ways of ascertaining the coefficient of correlation by an extremely simple calculation. It is based on the direction of change or variation in the two paired variables. In this method correlation is calculated between the direction of deviation and not their magnitude. In majority of Pearson's coefficient with much less calculations.

To calculate the coefficient of concurrent deviations the deviations are not calculated from an average or assumed or moving average but only their direction from the preceding item and only the direction of the deviation (i.e., positive or negative) and not the extent of deviation are considered. The formula for calculation of coefficient of concurrent deviation is given below.

Coefficient of concurrent deviation or

$$r_c = \pm \sqrt{\pm (2c - n)/2}$$

where r_c stands for coefficient of concurrent deviation.

c stands for the number of pairs of concurrent deviation and n for number of pair observations. The value of this coefficient of correlation also varies between +1 to -1. The plus and minus signs given in the formula should be carefully noted. If the value of $(2c - n)/n$ is negative its square cannot be calculated and so a minus sign is placed before the sign of the root so that the square root may be calculated and the minus sign may be kept before the value of the coefficient of correlation.

The steps in the calculation of this coefficient are

1. Examine the fluctuation of each series and find whether each item increases or diminishes in comparison with the item just preceding. All increases are noted as plus and all decreases as minus. If there is neither increase or decrease the direction is zero. The direction of change on both series is denoted by dx and dy respectively.
2. Multiply dx and dy and determine the value of C which would be the number of positive product of duty ie $(-x-)$ or $(+x+)$
3. Count number of paired observation 'n'
4. Use the formula given below to obtain the value of the coefficient or

r_c

$$r_c = \pm \sqrt{\pm \frac{2c - n}{n}}$$

we will apply the formula in the following example to illustrate the steps.

Illustration: The following are the marks obtained by a group of 10 students in Economics and Statistics.

Students	1	2	3	4	5	6	7	8	9	10
Economics	10	36	98	25	75	85	91	65	68	34
Statistics	49	39	92	60	68	62	86	58	53	47

Calculate r_c by the method of concurrent deviation.

Solution:

Students	Marks in Economics	Deviation from preceding item (dx)	Marks in Statistics	Deviation from preceding item (dy)	dx dy
1	10	.	49	.	.
2	36	+	39	-	-
3	98	+	92	+	+
4	25	-	60	-	+
5	75	+	62	+	+
6	85	+	62	-	-
7	91	+	86	-	+
8	65	-	58	-	-
9	68	+	53	-	-
10	34	-	47	-	+
	<u>N = 9</u>		<u>N = 9</u>		<u>C = 6</u>

$$r_c = \pm \sqrt{\pm \frac{2c - n}{n}}$$

$$r_c = \pm \sqrt{\pm \frac{2(6) - 9}{9}}$$

$$r_c = \sqrt{\frac{12 - 9}{9}}$$

$$= \sqrt{3/9}$$

$$= 0.58.$$

Merits and Demerits of Concurrent Deviation Method

Merits:

1. As compared to other methods it is the simplest and easiest.
2. When the number of items is very large, this method may be used to form a quick idea about the degree of relationship before making use of complicated methods.

Demerits:

1. The chief demerit is that it is only a very rough indicator of correlation.
2. It does not differentiate between small and big deviation i.e., an increase from 10 to 11 is given the same weight as to 10 to 10,000, i.e. it takes into consideration the direction of change and not the magnitude. It should be however, remembered that the results obtained by this method are not very different from those obtained by the use of Karl Pearson's coefficient in the case of short term oscillation only.

UNIT - VI

REGRESSION ANALYSIS

- ❑ INTRODUCTION
- ❑ MEANING AND DEFINITION
- ❑ DIFFERENCE BETWEEN CORRELATION AND REGRESSION ANALYSIS
- ❑ METHODS OF STUDYING REGRESSION 1. GRAPHIC, 2. ALGEBRAIC
- ❑ REGRESSION LINE, REGRESSION COEFFICIENTS
- ❑ USE OF REGRESSION ANALYSIS
- ❑ COEFFICIENT OF DETERMINATION

INTRODUCTION

The dictionary meaning of the word regression is 'stepping back' or returning to the average value. Francis Galton in the later half of Nineteenth century had used this word for the first time. Galton studied the average relationship between the height of fathers and their sons graphically and called the line describing the relationship, the line of regression. But to day the word regression is used in statistics with a much wider perspective. Regression analysis, in general sense, means the estimation or prediction of the unknown value of one variable from the known value of the other variable. It is one of the very important statistical tools which is extensively used in almost all sciences – natural, social and physical. It is specially used in business and economics to study the relationship between two or more variables that are related causally and for estimation of demand and supply curves, cost functions, production and consumption function.

MEANING AND DEFINITION

Prediction or estimation is one of the major problems in almost all spheres of human activity. The estimation or prediction of future production, consumption, prices, investments, sales, profit, income etc. are of paramount importance to a businessman or economist. Population

estimates and population projections are indispensable for efficient planning of an economy. Regression analysis is one of the very scientific techniques for making such predictions.

At this stage we shall examine some definition of the term regression:

- (1) "Regression is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data" – M.M. Blair.
- (2) "One of the most frequently used techniques in economics and business research to find a relation between two or more variables that are related causally, is regression analysis" – Taro Yanane.
- (3) "Regression analysis attempts to establish the nature of the relationship between variables – i.e. to study the functional relationship between the variable and thereby provide a mechanism for prediction or forecasting" – Ya Lum Cdou.
- (4) The term "regression analysis refers to the methods by which estimates are made of the values of a variable from a knowledge of the values of one or more other variables and to the measurement of the errors involved in this estimation process" – Morris Hamburg.

The above definitions make it clear that regression analysis is done for estimating or predicting the unknown value of one variable from the known value of the other variable. The variable which is used to predict the variable of interest is called as the independent variable or "explanatory variable" and the variable we are trying to predict is called as "explained variable" or "dependent variable". The independent variable is called X and dependent variable Y. The regression study which confines itself to a study of only two variables is called simple regression. The regression analysis which studies more than two variables at a time is called multiple regression.

It should be noted that the terms dependent and independent refer to the mathematical or functional meaning of dependence – they do not imply that there is necessarily any cause and effect relationship. What it meant

is simply that estimates of values of dependent variable Y may be obtained for given values of independent variable X from a mathematical function involving X and Y are dependent on value of X. The X variable may or may not be causing change in the Y variable.

DIFFERENCE BETWEEN CORRELATION AND REGRESSION ANALYSIS

Both correlation and regression analysis help us in studying the relationship between two variables; however they differ in their approach and objective.

1. The correlation coefficient measures the degree of covariability between the variables whereas the objective of regression analysis is to study the nature of relationship between the variables which will be able to help us in prediction of value of one variable on the basis of the other.
2. Correlation between two series is not necessarily a cause and effect relationship. A high degree of positive correlation between two variables does not mean that one is the effect of the other. There may be no cause and effect relationship and yet they may be correlated. Regression, however, presumes one variable as a cause and the other as its effect. It should be noted that the presence of association does not imply causation, but existence of causation always implies association.
3. The coefficient of correlation varies between ± 1 The regression coefficient have the same high as the correlation coefficient.
4. Correlation coefficient cannot exceed unity whereas one of the regression coefficients can have a value higher than unity but the product of the two regression coefficients cannot exceed unity because r is the square root of the product of the two regression coefficients.
5. There may be nonsense correlation between two variables which is purely due to chance and has no practical relevance. However there is nothing like nonsense regression.

6. Correlation coefficient is independent of scale and origin. Regression coefficients are independent of change of origin but not scale.

METHODS OF STUDYING REGRESSION

Regression can be studied by two methods:

- (1) Graphically
- (2) Algebraically

1. Graphic Method

Graphic method is otherwise known as scatter diagram method. It is possible to find out the actual relationship between two variables with the help of a scatter diagram. A scatter diagram contains one point for each pair of values of X and Y variable. Usually in the diagram the independent variable is taken on horizontal axis (i.e. X axis) and dependent variable on vertical axis (Y axis). If the points form a straight line then there is a perfect correlation and the value of one variable can be estimated given the value of other. But mostly in economic and commercial problems, perfect correlation is a rarity, so the problem is to draw line on graph in such a way that dots are best represented by it.

This line is to be drawn by inspection and care must be taken to draw it in such a way as to be the best fit. The following points should be kept in mind while drawing this line:

1. The line should be as close as possible to all the points on the graph.
2. Almost an equal number of points should be there on either side of the line.
3. An attempt should be made to draw the line in such a way that the points on its either side are equidistant from it.

When both the variables are made as dependent as well as independent, we get two regression lines representing the dependence of one on the other. We can draw two regression lines to predict the values of X and Y variables. The regression line which is used to predict the value of Y for a value of X is called "regression line of Y on X". Similarly the

regression line used to predict the value of X for a value of Y is called "regression line of X on Y". If the coefficient of correlation between X and Y are perfect i.e. either +1 or -1 there will be only one regression line as the variation in the two series in such cases always increases or decreases by a constant figure.

Merits and demerits of the method: This method is very simple and easy. It does not take much time to draw the estimating line. But as it is drawn by free hand different persons may draw different lines for the same data.

Method of least square: In order to avoid the difficulties associated with the drawing of regression line by graphic freehand method, a mathematical relationship is established between the movement of the variables and algebraic equations are obtained to represent the relative movements of X and Y series. One such method is the method of least square. In this method we minimise the sum of squares of the deviations between the given values of a variable and its estimated value given by the line of best fit. Line or regression of Y on X is the line which gives the best estimate for the value of Y for a specified value of X and similarly the line of regressions of X and Y gives the best estimate for the value of X for a specified value of Y. In the method of least square the line of best fit is obtained by the equation of straight line $Y = a + bX$ and that in the method of least squares this line is obtained with the help of the following normal equation:

$$\Sigma Y = na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

If the values of X and Y variables are substituted in the above equation we get the values of a and b and thus get the regression line of Y on X. Here Y is the dependent variable and X is the independent variable. To get the regression line of X and Y we will have to assume X as the dependent variable and y as the independent variable.

Very often the question comes to mind as to why should there be two regression lines to obtain the values of Y and X and why one regression line does not serve the purpose. The answer is simple and it is that one regression line cannot minimize the sum of square of deviation for both X and y series unless the relationship between them indicates perfect positive or negative correlation.

Regression lines: If we take the case of two variables X and Y, we shall have two regression lines as regression of X on Y and regression of Y on X. The regression line of Y on X gives the most probable values for Y for given values of X and regression line of X on Y gives the most probable values of X for given value of Y. However when there is a perfect correlation ($r = \pm 1$) the regression lines will coincide i.e. we will have only one line. The nearer the lines, higher the degree of correlation and the farther the two regression lines from each other, the lesser is the degree of correlation.

It should be noted that the regression lines cut each other at the point of average of X and Y i.e. if from the point of intersection a perpendicular is drawn on X-axis we got mean value of X and if perpendicular is drawn on Y axis we get the mean value of Y.

It is important to note that the regression lines are drawn on least squares assumption which stipulates that the sum of squares of the deviations of the observed Y value from the fitted line shall be minimum. The total of the squares of the deviations of various points is minimum only from the line of best fit. The deviation from the points from the line of best fit can be measured in two ways vertical i.e. parallel to Y axis and horizontal i.e. parallel to X axis for minimising the total of the squares separately it is essential to have two regression line. The regression line of Y on X is drawn in such a way that it minimises total of squares of the vertical deviation and regression line of X on Y minimises the total squares of the horizontal deviations. This can be best appreciated with the help of the following example.

Illustration:

Height of fathers (in inches)	65	63	67	64	68	62	70	66	68	67	69	71
Height of sons (in inches)	68	66	68	65	69	66	68	65	71	67	68	70

The regression equations corresponding to these variables are

$$X = -3.38 + 1.036 Y \longrightarrow (1)$$

$$Y = 35.82 + 0.476 X \longrightarrow (2)$$

By assuming any value of Y we can find out corresponding values of X from equation (1)

for example if Y is 63 X is 61.89

if Y is 67 X is 65.63

we can plot these points on the graph and obtain regression line on X and Y.

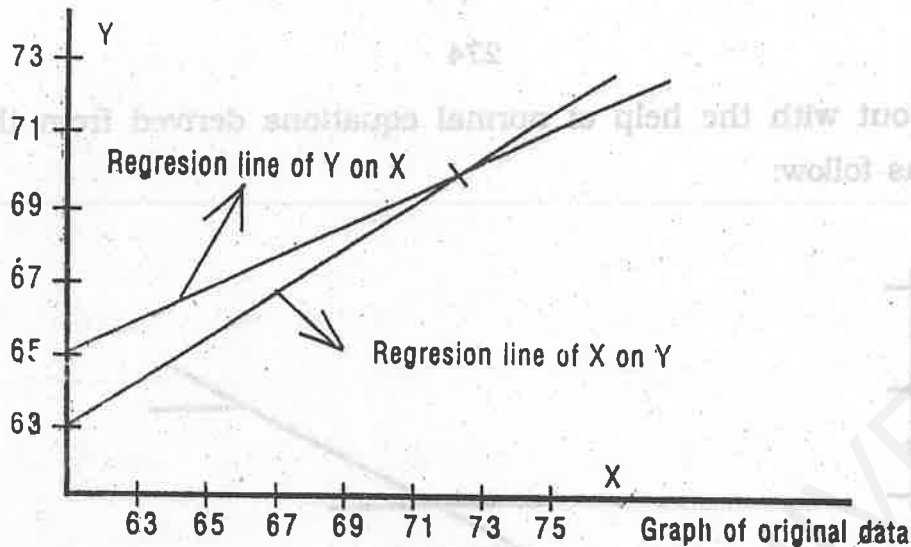
Similarly by arranging any value to X in equation (2) we can obtain corresponding value of Y.

Thus if X is 65 Y would be 66.76

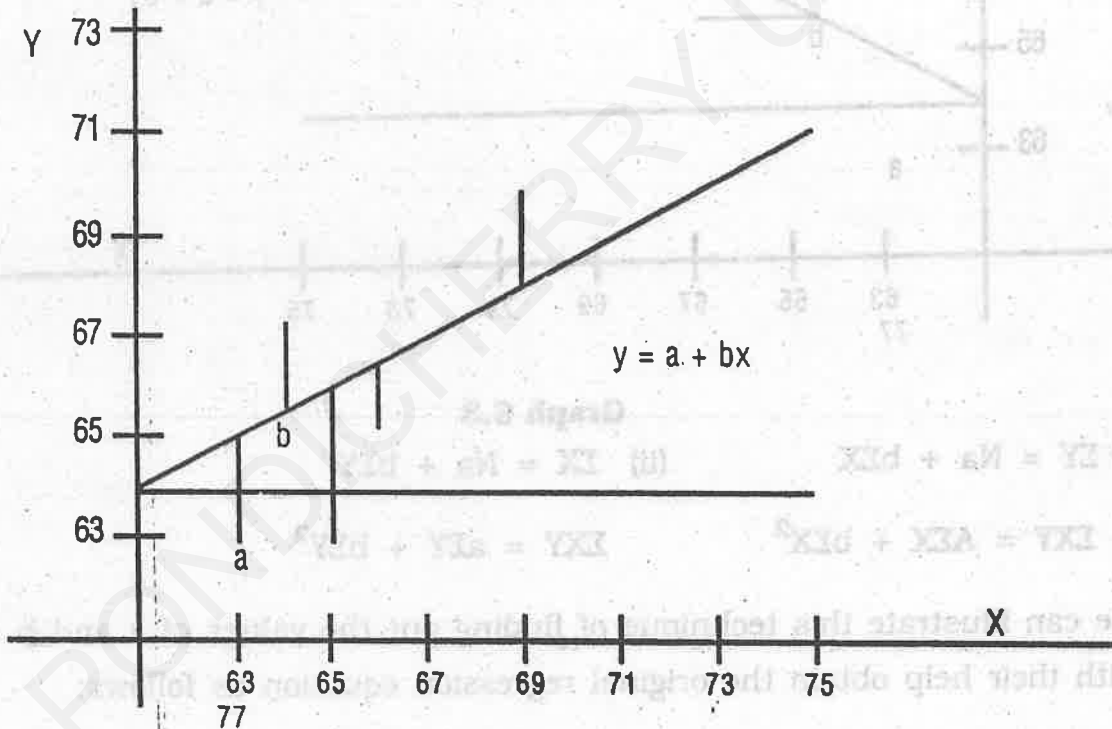
X is 70 Y would be 69.14

The graph of original data and these lines would be as follows:

Regression equation: Regression equations are algebraic expressions of the regression lines. Since there are two regression lines, there are two regression equations. Regression line of X on Y gives the best possible mean values of X for given value of Y and similarly the regression line of Y on X gives the best possible mean values of Y for given value of X.



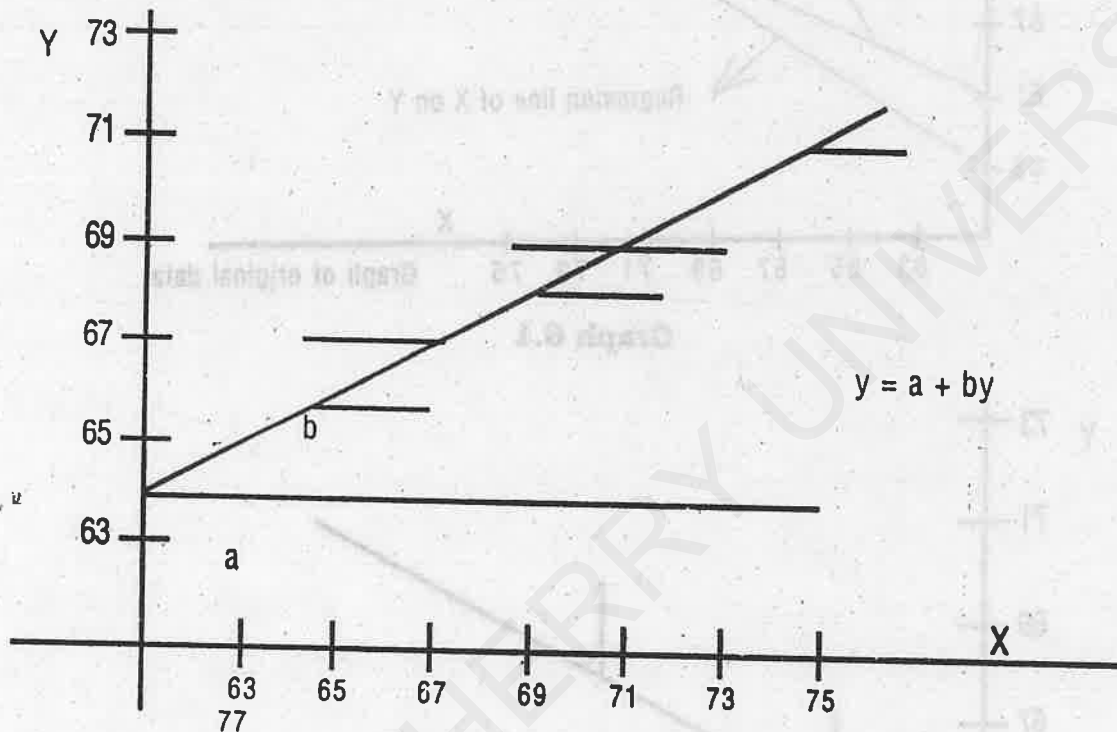
Graph 6.1



Graph 6.2

The regression equation of X on Y is $X = a + bY$ and the regression equation of Y on X is $Y = a + bX$. These are the equations of a straight line. In these equations the values of 'a' and 'b' are constant. The parameter 'a' indicates the level of the line of regression (the distance of the line of regression i.e. the distance of the line above or below the origin). The parameter 'b' determines the slope of the line i.e. the corresponding changes in X in relation to per unit change in Y or increase in the value of a, b can

be found out with the help of normal equations derived from the main equation as follow:



Graph 6.3

(i) $\Sigma Y = Na + b\Sigma X$

(ii) $\Sigma X = Na + b\Sigma Y$

$$\Sigma XY = A\Sigma X + b\Sigma X^2$$

$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

We can illustrate this technique of finding out the values of a and b and with their help obtain the original regression equation as follows:

Illustration: From the following data obtain the two regression equations using the method of least square:

X	1	6	5	2	1	1	7	3
Y	6	1	0	6	1	2	1	5

Solution:

X	Y	X ²	Y ²	XY
1	6	1	36	6
6	1	36	1	6
5	0	25	0	0
2	6	4	36	12
1	1	1	1	1
1	2	1	4	2
7	1	49	1	7
3	5	9	25	15

$$\Sigma X = 26 \quad \Sigma Y = 22 \quad \Sigma X^2 = 126 \quad \Sigma Y^2 = 104 \quad \Sigma XY = 49$$

Regression equation Y on X

$$Y = a + bX$$

To get the values of a and b the following two normal equations are used:

$$\Sigma Y = na + b\Sigma X$$

$$\Sigma XY = \Sigma X + b\Sigma X^2$$

Substituting the values

$$22 = 8a + b(26) \longrightarrow (1)$$

$$49 = a(26) + b(126) \longrightarrow (2)$$

Multiplying equation (1) by 13 and equation (2) by 4 deducting (4) from (3) we get

$$286 = 104a + 338b \longrightarrow (3)$$

$$196 = -104a + 504b \longrightarrow (4)$$

$$\hline 90 = -166b$$

$$\text{or } b = -0.54 \text{ (approx)}$$

Substituting 'b' in equation (1) we get

$$22 = 8a + 26(-0.54)$$

$$= 8a - 14.04$$

$$8a = 22 + 14.04$$

$$a = 36.04/8$$

$$= 4.51 \text{ (approx)}$$

$Y = 4.51 - (0.54)X$. This is the regression equation of Y on X.

Similarly the regression equation of X on Y is $X = a + bY$. The two normal equations are

$$\Sigma X^2 = Na + b\Sigma Y$$

$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

Substituting the values in the above equations

$$26 = 8a + b(22) \longrightarrow (1)$$

$$4a = 22a + b(104) \longrightarrow (2)$$

Multiplying equation (1) by 11 and (2) by 4 we get

$$286 = 88a + 242b \longrightarrow (3)$$

$$196 = 88a + 416b \longrightarrow (4)$$

Deducting (4) from (3) we get

$$90 = -174b \text{ or}$$

$$b = -0.52 \text{ (approx)}$$

Substituting b value in equation (1) we get

$$26 = 8a + 22(-0.52)$$

$$26 = 8a + (-11.44) \quad \text{or}$$

$$a = 26 + 11.44 / 8$$

$$= 4.68$$

Therefore $X = 4.68 - (0.52) Y$. This is the regression equation of X on Y.

Deviations taken from Arithmetic Means of X and Y

The method of obtaining regression equations by the method of least square that we have seen above is very tedious. The work can be simplified to a larger extent if instead of obtaining the regression equations with the help of original values we take the deviations of X and Y series from their respective means. It simplifies the calculation and gives us the same result as given by the method of least square.

The regression equations in such a case are written as follows:

Regression equations of X on Y: $(X - X') = r \sigma_x / \sigma_y (Y - Y')$

X' is the mean of X series and Y' is the mean of Y series.

r is the coefficient of correlation between X and Y series and σ_x and σ_y are the standard deviations of X and Y series respectively.

$r \sigma_x / \sigma_y$ is called the regression coefficient of X on Y and denoted by b_{xy} . Thus

$$\begin{aligned} b_{xy} &= r \frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{N \sigma_x \sigma_y} \times \frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{N \sigma_y^2} \\ &= \frac{\Sigma xy}{n \Sigma y^2 / n} = \frac{\Sigma xy}{\Sigma y^2} \end{aligned}$$

Thus instead of finding out the values of r , σ_x , σ_y we can directly find out the value of b_{xy} by dividing the product of the deviations of X and Y series from their respective means by the sum of the squares of the deviations of the Y series from its mean.

Similarly regression equation of Y on X

$$(Y - Y') = r \sigma_y / \sigma_x (X - X')$$

$r \sigma_y / \sigma_x$ is called the regression coefficient of Y on X or 'byx'.

$r \sigma_y / \sigma_x$ can be calculated in the same way as above and its

value will be $\Sigma xy / \Sigma x^2$

Thus the two regression equations can be rewritten as follows:

(i) Regression equation of X on Y

$$X - X' = \Sigma xy / \Sigma y^2 (Y - Y')$$

(ii) Regression equation of Y on X

$$Y - Y' = \Sigma xy / \Sigma x^2 (X - X')$$

REGRESSION COEFFICIENT

As discussed above bxy and byx are called as the regression coefficient of regression equation X on Y and Y on X. The regression coefficients bxy and byx possess some important properties. They are

1. The underroot of the product of two regression coefficients gives us the value of correlation coefficients. Symbolically

$$r = \sqrt{b_{xy} \times b_{yx}}$$

Proof:

$$b_{xy} = r \sigma_x / \sigma_y \text{ and } b_{yx} = r \sigma_y / \sigma_x$$

$$b_{xy} \times b_{yx} = r \sigma_x / \sigma_y \times r \sigma_y / \sigma_x = r^2$$

$$r = \sqrt{b_{xy} \times b_{yx}}$$

Then the geometric mean of bxy and byx gives the value of coefficient of correlation.

2. Both the regression coefficients will have signs i.e., either they will be positive or negative. The reason is that so far as standard deviation is concerned they are always positive. Only coefficient of correlation can be either positive or negative. If regression coefficient are negative r will also be negative. For example if $b_{xy} = -1.3$ and $b_{yx} = -0.65$

$$r = \sqrt{-1.3 \times -0.65} = -0.92 \text{ not } 0.92$$

3. Since the correlation coefficient cannot exceed one, one of the regression coefficients must be less than one or more; in other words both the coefficients cannot be greater than one.

The following example would illustrate the use of the above method of obtaining regression equation.

Illustration: From the following data, obtain the two regression equations and their correlation coefficient.

Sales	46	42	44	40	43	41	45
Purchase	40	38	36	35	39	37	41

Solution: Let us denote the sales by the variable X and purchase by the variable Y.

X	X'	x ²	Y	Y'	y ²	XY
46	+3	9	40	+2	4	+6
42	-1	1	38	0	0	0
44	+1	1	36	-2	4	-2
40	-3	9	35	-3	9	9
43	0	0	39	+1	1	0
41	-2	4	37	-1	1	2
45	+2	4	41	+3	9	6
$\Sigma x = 301$		$\Sigma x^2 = 28$	$\Sigma y = 266$		$\Sigma y^2 = 28$	$\Sigma xy = 21$

$$X' = \Sigma x / n = 301 / 7 = 43$$

$$Y' = 266 / 7 = 38$$

Regression equation of Y on X

$$Y - Y' = b_{yx}(X - X') \text{ where } b_{yx} = \Sigma xy / \Sigma x^2$$

Putting the values

$$Y - 38 = 21/28 (X - 43)$$

$$Y = 0.75 (X - 43) + 38$$

$$= 0.75X + 5.73$$

Regression equation of X on Y

$$X - X' = b_{xy} (Y - Y') \text{ where } b_{xy} = \Sigma xy / \Sigma y^2$$

$$X - 43 = 21/28$$

$$= 0.75 (Y - 38)$$

$$= 0.75Y + 14.50$$

We have

$$r^2 = b_{yx} \cdot b_{xy} = 0.75 \times 0.75$$

$$r = 0.75$$

Since both the regression coefficients are positive r must be positive.
Hence $r = 0.75$.

Illustration: A panel of judges A and B graded seven debators and independently awarded the following marks:

Debator	Marks by A	Marks by B
1	41	33
2	35	40
3	29	29
4	31	32
5	45	39
6	39	36
7	32	29

An eighth debator was awarded 37 marks by judge A while judge B was not present. If judge B had also been present, how many marks do you expect him to award to the eighth debator for assuming that the same degree of relationship exists in their judgement.

Solution: Let the marks awarded by judge 'A' be denoted by the variable X and marks awarded by judge 'B' be variable Y.

Debator	X	Y	X	Y	x^2	y^2	xy
			$x - x'$	$y - y'$			
1	41	33	5	-1	25	1	-5
2	35	40	-1	6	1	36	-6
3	29	29	-7	-5	49	25	35
4	31	32	-5	-2	25	4	10
5	45	39	9	5	81	25	45
6	39	36	3	2	9	4	6
7	32	29	-4	-5	16	25	20
<hr/>							
	$\Sigma X = 252$	$\Sigma Y = 238$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma x^2 = 206$	$\Sigma y^2 = 120$	$\Sigma xy = 105$

$$X' = \Sigma X / n = 252 / 7 = 36$$

$$Y' = \Sigma Y / n = 238 / 7 = 34$$

The equation for regression line of Y on X is given as

$$Y - Y' = b_{yx} (X - X') \quad \text{where } b_{yx} = \Sigma xy / \Sigma x^2$$

$$Y - 34 = 105 / 206 (X - 36)$$

$$Y = 0.51(X - 36) + 34$$

$$Y = 0.51X - 18.36 + 34$$

$$Y = 0.51X + 15.64$$

When X is 37

$$Y = 0.51 \times 37 + 15.64$$

$$= 18.87 + 15.64$$

$$= 34.51$$

Hence if the judge B had been present he would have given 35 marks to the eighth debator.

Deviation from Assumed Mean

Where the actual means of the two series X and Y are in fraction, the method of 'A' finding out regression equations discussed above becomes very tedious. In such cases the deviations are taken from the assumed mean. When deviations are taken from assumed mean the entire procedure of finding regression equations remains the same - the only difference is we take deviation from the assumed mean. However some simplification is possible. The regression equation of X on Y is

$$X - X' = r \sigma_x / \sigma_y (Y - Y')$$

Now the value of $r \sigma_x / \sigma_y$ or b_{xy} will be obtained as follows:

$$b_{xy} = r \frac{\sigma_x}{\sigma_y} = \frac{\sum dx dy - \frac{\sum dx \sum dy}{N}}{\sigma_x \sigma_y} \times \frac{\sigma_x}{\sigma_y}$$

$$\text{or } b_{xy} = \frac{\sum dx dy - \frac{\sum dx \sum dy}{N}}{\sigma_y^2}$$

$$\text{or } b_{xy} = \frac{\sum dx dy - \frac{\sum dx \sum dy}{N}}{\sum dy^2 - \frac{(\sum dy)^2}{N}}$$

Similarly the regression equation of Y on X is

$$Y - Y' = r \sigma_y / \sigma_x (X - X')$$

$$\text{where } r \frac{\sigma_y}{\sigma_x} = \frac{\sum dx dy - \frac{\sum dx \sum dy}{N}}{\sigma_y \sigma_x} \times \frac{\sigma_y}{\sigma_x}$$

$$= \frac{\Sigma dx dy - \frac{\Sigma dx \Sigma dy}{N}}{\sigma_x^2}$$

$$b_{xy} = \frac{\Sigma dx dy - \frac{\Sigma dx \Sigma dy}{N}}{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}}$$

Once the value of b_{xy} and b_{yx} are found out without calculating r , σ_x , σ_y , they can be inserted in the formula of this regression equation.

It should be noted that in both the cases the numerator is the same, the only difference is in the denominator.

However in case of grouped data given in a two way frequency table the formula is slightly changed. In a correlation table the given frequencies are also taken, therefore the values are to be multiplied by the frequencies. Another important change is that if step deviation has to be taken the values given by the formula has to be multiplied by i_x/i_y in case of regression equation of X on Y and by i_y/i_x in case of regression equation Y on X where i_x and i_y are the class intervals of X and Y series. This is necessary because unlike coefficient of correlation, regression coefficient are affected by the change of scale, though they are not affected by change of origin when the regression coefficient are calculated from correlation, table their values are obtained as follows:

$$b_{xy} = r \frac{\sigma_x}{\sigma_y} = \frac{\Sigma f dx dy - \frac{\Sigma f dx \Sigma f dy}{N}}{\Sigma f dx^2 - \frac{(\Sigma f dx)^2}{N}} \times \frac{i_x}{i_y}$$

$$b_{yx} = r \frac{\sigma_y}{\sigma_x} = \frac{\Sigma f dx dy - \frac{\Sigma f dx \Sigma f dy}{N}}{\Sigma f dx^2 - \frac{(\Sigma f dx)^2}{N}} \times \frac{i_y}{i_x}$$

The following example would illustrate the application of the formula given below:

Illustration: From the following data obtain regression equation taking the deviation from assumed mean:

X	78	89	97	69	59	79	68	61
Y	125	137	156	112	107	136	123	108

Solution:

X	dx	dx ²	Y	dy	dy ²	dx dy
	(X - A)			(Y - A)		
	A = 69			A = 112		
78	9	81	125	13	169	117
89	20	400	137	25	625	500
97	28	784	156	44	1936	1232
69	0	0	112	0	0	0
59	-10	100	107	-5	25	50
79	10	100	136	24	576	240
68	-1	1	123	11	121	-11
61	-8	64	108	-4	16	32
$\Sigma x=600 \quad \Sigma dx=48 \quad \Sigma dx^2=1530 \quad \Sigma y=1004 \quad \Sigma dy=108 \quad \Sigma dy^2=3468 \quad \Sigma dx dy=2160$						

Regression equation of Y on X

$$Y - Y' = by_x (X - X')$$

$$by_x = \frac{\Sigma dx dy - \frac{\Sigma dx \Sigma dy}{N}}{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}}$$

Substituting the values

$$by_x = \frac{2160 - \frac{48(108)}{8}}{1530 - \frac{(48)^2}{8}}$$

$$= \frac{2160 - 648}{1530 - 288} = \frac{1512}{1242} = 1.22 \text{ (approx)}$$

Therefore the equation of line of regression of Y on X

$$Y - Y' = by_x (X - X')$$

$$Y - 125.5 = 1.22 (X - 75)$$

$$Y = 1.22 X - 91.5 + 125.5$$

$$= 34 + 1.22 X$$

Regression equation X on Y

$$X - X' = b_{xy} (Y - Y')$$

$$b_{xy} = \frac{\sum dx dy - \frac{\sum dx \sum dy}{N}}{\sum dy^2 - \frac{(\sum dy)^2}{N}}$$

$$b_{xy} = \frac{2160 - \frac{48(108)}{8}}{3468 - \frac{(108)^2}{8}}$$

$$= \frac{2160 - 648}{3468 - 1458} = \frac{1512}{2010} = 0.75 \text{ (approx)}$$

Therefore the equation of line of regression of X on Y

$$X - X' = b_{xy} (Y - Y') \text{ where } X' = 75, Y' = 125 \text{ and } b_{xy} = 0.75$$

$$X - 75 = 0.75 (Y - 125)$$

$$X - 75 = 0.75 Y - 93.75$$

$$X = 0.75 Y - 93.75 + 75$$

$$= 0.75 Y - 18.75$$

Illustration: Following is the distribution of students according to their heights and weights

Height (in inches)	Weights (in lbs)			
	90 - 100	100 - 110	110 - 120	120 - 130
50 - 55	4	7	5	2
55 - 60	6	10	7	4
60 - 65	6	12	10	7
65 - 70	3	8	6	3

Calculate the two coefficients of regression and obtain the two regression equations.

Table - 6(A)

Hight in inches		Weight 2 lbs	n	90-100	100-110	110-120	120-130	f	fdx	fdx ²	fdxdy
				95	105	115	125				
			dy/dx	-2	-1	0	1				
50-55	52.5	-1	4	7	5	2	18	-18	18	13	
55-60	57.5	0	6	16	7	4	27	0	0	0	
60-65	62.5	1	6	12	10	7	35	35	35	-17	
65-70	67.5	2	3	8	6	3	20	40	80	22	
			f	19	37	28	16	100	Σfdx= 57	Σfdx ² = 133	Σfdxdy= -26
			fdy	-38	-37	0	16	Σfdy= -59			
			fdy ²	75	37	0	16	Σfdy ² = 129			
			fdxdy	-16	-21	0	-11	Σfdxdy= -26			

Table 6A

$$X' = A + \frac{\sum fdy}{N} \times C$$

$$X' = 115 - \frac{59}{100} \times 10 = 109.1$$

$$Y' = A + \frac{\sum fdx}{N} \times C$$

$$Y' = 57.5 + \frac{57}{100} \times 5 = 60.35$$

Regression equation of Y on X

$$Y - Y' = b_{yx} (X - X')$$

$$b_{yx} = \frac{\sum fdx dy - \frac{\sum fdx \sum fdy}{N}}{\sum fdx^2 - \frac{(\sum fdx)^2}{N}} \times \frac{i_y}{i_x}$$

$$b_{yx} = \frac{-26 - \frac{(-59)(57)}{100}}{133 - \left[\frac{57}{100}\right]^2} \times \frac{100}{5} = 0.145$$

$$Y - 109.1 = 0.145 (X - 60.35)$$

$$Y = 0.145 X + 103.35$$

Regression equation of X on Y

$$X - X' = b_{xy} (Y - Y')$$

$$b_{xy} = \frac{\sum fdx dy - \frac{\sum fdx \sum fdy}{N}}{\sum fdy^2 - \frac{(\sum fdy)^2}{N}} \times \frac{i_x}{i_y}$$

$$b_{xy} = \frac{-26 - \frac{(-59)(57)}{100}}{129 - \left[\frac{-59}{100}\right]^2} \times \frac{5}{10} = 0.044$$

$$X - 60.35 = 0.044 (Y - 109.1) = 0.044 Y - 4.8$$

$$X = 0.044Y + 55.55$$

USES OF REGRESSION

Regression analysis is a branch of statistical theory that is widely used in almost all the scientific disciplines. The use of regression analysis is done for estimating or predicting the unknown value of one variable from the known value of other variable. In economics as well as in the field of business this tool of statistical analysis is widely used. Businessmen are interested in predicting future production, consumption, investment prices, profits, sales etc. In fact the success of a businessman depends much on the correctness of various estimates. In our day to day life also with the help of regression analysis we can estimate or predict the effect of one variable on the other. In particular, regression analysis attempts to accomplish the following:

1. Regression analysis helps to estimate the dependent variable from the value of an independent variable
2. With the help of regression coefficient we can calculate the correlation coefficient. The square of correlation coefficient (r), called coefficient of determination, measures the degree of association or correlation that exists between the two variables.
3. Regression analysis is also used as a measure of error involved in using the regression line as the basis for estimation.

The use of regression concept has varied applications in physical as well as in social sciences.

COEFFICIENT OF DETERMINATION

Coefficient of determination (r^2) is defined as the ratio of the explained variance to the total variance.

$$\text{Coefficient of determination} = \frac{\text{Explained Variation}}{\text{Total Variance}}$$

Coefficient of determination is nothing but square of coefficient of correlation i.e., r^2 . One very convenient and useful way of interpreting the value of coefficient of correlation between two variables is to use the square of coefficient of correlation which is nothing but coefficient of determination. The maximum value of coefficient of determination is unity, because 'r' is never more than one and the explained variation of a dependent variable by an independent variable can never be more than total variation.

It should be noted that the fact that a correlation between two variables has a value of $r = 0.66$ and correlation between two other variables has a correlation of $r = 0.30$ does not demonstrate that first correlation is twice as strong as the second. The relationship between the two given values of r can be better understood by computing r^2 . When $r = 0.60$, $r^2 = 0.36$ and when $r = 0.3$, $r^2 = 0.09$. This implies that in first case 36% of the total variation is explained and in second case 9% of the total variation is explained.

The coefficient of unexplained variance to total variance is called as coefficient of non determination. Thus the coefficient of non determination is one minus coefficient of determination. The coefficient of non determination is denoted by K^2 and its square root is called as coefficient of alienation.

$$\text{Symbolically } K^2 = 1 - r^2$$

$$\text{Coefficient of non determination} = 1 - \frac{\text{Explained variance}}{\text{Total variance}}$$

$$\text{Coefficient of alienation} = \sqrt{1 - \frac{\text{Explained variance}}{\text{Total variance}}}$$

The coefficient of determination, however is often misinterpreted. It may be interpreted from the very term that X stands in a determining or causal relationship to Y or r^2 speaks of how much variation X is able to explain of the total variance on Y. However it is a neutral term whether causally is there between X and Y is to be determined on the basis of

evidences other than quantitative measurement, on the top of it r^2 being a square it is always a positive number. It cannot tell whether the correlation is positive or negative. Thus the square root of r^2 , $\sqrt{r^2} = \pm r$ is frequently computed to indicate the direction of the relationship in addition to indicating the degree of relationship.

Limitations of Regression Analysis

With the help of regression the estimates are made but it should be noted that for estimation, unless the assumption that has been taken remains unchanged since the equation was computed, estimates may go wrong. Another point to be remembered is that the relationship shown in the scatter diagram may not be the same if the equation is extended beyond the values in computing the equation. For example, if we find a close linear relationship between yield of a crop and amount of fertilizer applied, it would not be logical to extend this equation beyond the limits of the experiment for it is quite likely that if the amount of fertilizer were increased indefinitely, the yield would eventually decline as too much fertilizer is applied.

UNIT - VII**LESSON - 1**

INTERPOLATION AND EXTRAPOLATION

- ❑ INTRODUCTION
- ❑ DEFINITION
- ❑ SIGNIFICANCE OF INTERPOLATION AND EXTRAPOLATION
- ❑ ASSUMPTIONS OF INTERPOLATION AND EXTRAPOLATION
- ❑ METHODS OF INTERPOLATION AND EXTRAPOLATION

INTRODUCTION

Many a time in practical work we come across situations where we have to estimate a value which is not available in a given series or predict a future value. For example the census of population of India takes place every 10 years i.e., we have figures for 1931, 1941, 1951, 1961, 1971, 1981, 1991. Now if we require the population of 1989 and 1996 what should we do? This can be done either by guessing or by analysing the data and inserting a value in between a series of data or projecting the forward. The process of inserting a value in between the series of data is called interpolation and projecting forward or backward on the basis of series of values given is known as extrapolation. Thus interpolation supplies us with the missing link and extrapolation helps in forecasting.

DEFINITION

The following definitions give a formal expression to the basic idea of interpolation and extrapolation.

"Interpolation consists in reading a value which lies between two extreme points. Extrapolation means reading a value, that lies outside the two extreme points" - W.M. Harper.

"Interpolation is the estimation of a most likely estimate in a given condition. The technique of estimation of a past figure is termed as

interpolation while estimating a probable figure for the future is called extrapolation" - Hirach.

There is no difference between interpolation and extrapolation so far as the methods are concerned but for distinguishing past from the future we give them two different names. Interpolation relates to past whereas extrapolation gives us the forecast for the future.

SIGNIFICANCE OF INTERPOLATION AND EXTRAPOLATION

The tools of interpolation and extrapolation are of great practical significance. Their utility can be well imagined in the following situations.

- (1) Estimation of Intermediate values:** It often happens in business or economic situations, a particular type of information is collected at regular intervals. It is likely that at some future date it may be felt that data for the intermediate period is necessary. The only alternative is to use the technique of interpolation. The most common examples are intercensus population figures, mid-term figures of industrial production etc.
- (2) Non-availability of data or loss of data:** In case of non-availability of data extrapolation helps to estimate the value of some past period and to fulfil the gap in the data caused on account of any loss or destruction of data and interpolation becomes useful.
- (3) Derivation of median and mode:** In case of continuous frequency distribution the interpolation technique is used to derive the formula for computation of median and mode.
- (4) Bringing uniformity in the data:** Sometimes data pertaining to a particular phenomenon are grouped by different agencies in different types of groups which make them unfit for comparison. To bring uniformity in groups interpolation technique is used.
- (5) Making of forecast:** For a number of phenomena estimates for future have to be made. Extrapolation is a scientific technique that can be used for estimating or projecting certain phenomena. However the accuracy of

interpolation depends on (1) knowledge of the possible fluctuations of the figures to be obtained by a general inspection of the fluctuations at dates for which they are given and (2) knowledge of course of events with which the figures are connected.

Assumptions

The following assumptions are made while making use of the techniques of interpolation and extrapolation.

1. There is no sudden jump in the figures from one period to another within the period under consideration. In other words the given data do not refer to abnormal periods such as famine, war, drought, epidemic, etc. which may result in sudden change in the series.
2. The rate of change of figures from one period to another is uniform.

The limitations however of these techniques may be

1. A number of consecutive missing values in a series cannot be estimated; at any rate they cannot be reliable.
2. Unless there is fairly good number of observations, the technique of interpolation and extrapolation fail to deliver desired result.
3. Unless the gaps between the values are equal these methods cannot be used.

Methods of Interpolation

Broadly speaking the various methods of interpolation can be divided under two heads.

1. Graphic method and
2. Algebraic method

Under algebraic head there are several methods. The following are some of the important and more popular methods:

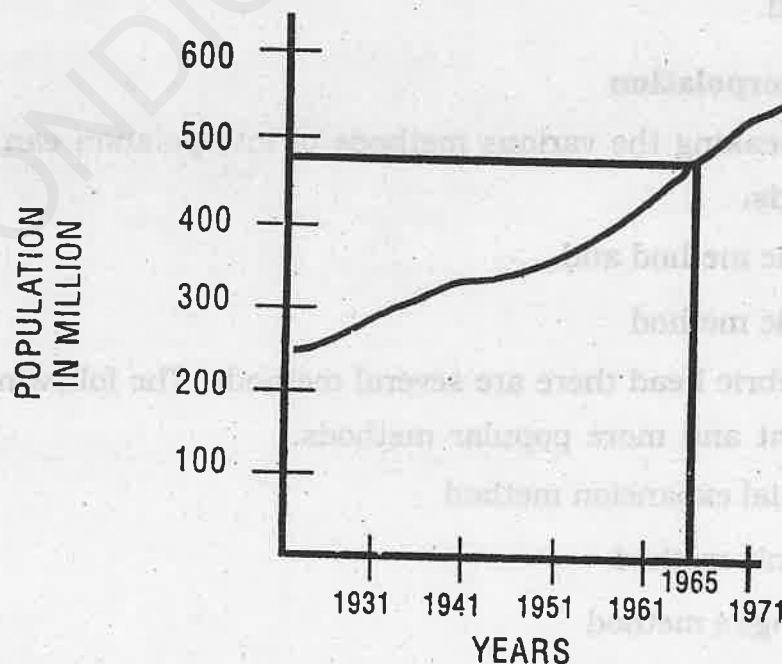
1. Binomial expansion method
2. Newton's method
3. Lagrange's method
4. Parabolic curve method

Graphic Method: This is the simplest of all types of interpolation methods. When this method is used the given data are plotted in a graph paper and the plotted points are joined. If there are only two values a straight line is obtained; otherwise a curve will be obtained. On the X-axis the years are taken and in the Y-axis the values of the variables. For the year the value is to be interpolated, a perpendicular line is drawn on the line or curve, whatever may be, depending on the number and values of the variables. From the point where the perpendicular drawn meets the curve, another perpendicular is drawn on the Y-axis. The corresponding value in Y-axis is the required value of the variable for the desired year. The following example will illustrate it.

Illustration: From the following data determine the population for the year 1965.

Year	Population in millions
1931	251
1941	279
1951	319
1961	439
1971	548

Solution:



Graph 7.1.1

Thus population for the year 1965 will be 472. This method is the simplest method of interpolation but suffers with the limitations that different smooth curves can be drawn through the point; thus the value arrived at is not free from subjectivity. In addition, the larger the volume of figures, the narrower the scale has to be in the graph paper and consequently greater will be the error of approximation. The method also will not be helpful for future projection.

Binomial expansion method

This method is simple to understand and requires very little calculation. But it is applicable only in those situations where the following two conditions are satisfied:

1. It can be used only when the independent variable X advances by equal interval say 5, 10, 15, 20, 25 etc. If the increase is not uniform this method is not applicable; for example if X is 5, 8, 13, 17, 24 this method cannot be applied.
2. The value of X for which Y is to be interpolated is one of the class limits of X series. For example if

X	5	10	15	20	25
Y	30	32	?	38	40

We can determine the values of Y corresponding to $X = 15$ but not when X is 12 & 19. The same is true for extrapolation i.e., we can extrapolate the values of $X = 30$ but $X \neq 27$.

When this method is applied the formula for the binomial theorem is expanded and equated with zero.

$$(Y - 1)^n = Y^n - nY^{n-1} + \frac{n(n-1)}{2!} Y^{n-2} - \frac{n(n-1)(n-2)}{3!} Y^{n-3} + \frac{n(n-1)(n-2)(n-3)}{4!} Y^{n-4} = 0$$

or

$$n_{cn} Y^n - nc_{n-1} Y^{n-1} + nc_{n-2} Y^{n-2} + nc_{n-3} Y^{n-3} \dots nc_0 Y^{n-0} = 0$$

where n is the number of known values of Y the expansion of the binomial for some values of n is given as

Number of known values of Y	Formula	Expansion
2	Δ_0^2 or $(Y-1)^2 = 0$	$Y_2 - 2Y_1 + Y_0 = 0$
3	Δ_0^3 or $(Y-1)^3 = 0$	$Y_3 - 3Y_2 + Y_1 - Y_0 = 0$
4	Δ_0^4 or $(Y-1)^4 = 0$	$Y_4 - 4Y_3 + 6Y_2 - 4Y_1 + Y_0 = 0$
5	Δ_0^5 or $(Y-1)^5 = 0$	$Y_5 - 5Y_4 + 10Y_3 - 10Y_2 + 5Y_1 - Y_0 = 0$
6	Δ_0^6 or $(Y-1)^6 = 0$	$Y_6 - 6Y_5 + 15Y_4 - 20Y_3 + 15Y_2 - 6Y_1 + Y_0 = 0$

The expansion of the binomial formula as shown appears to be a little difficult and complex. However this can be done by a simple procedure as follows:

- (1) The first subscript of Y will be the number equivalent of which we have to find the binomial expansion. Thus if $(Y-1)^4 = 0$ is to be expanded the first Y will be Y^4 . After that each Y 's subscript will be reduced by 1 till it reaches Y_0 i.e., Y_4, Y_3, Y_2, Y_1 and Y_0 .

- (2) The plus and minus signs are to be placed alternatively starting from first which is a plus

$$+Y_4, -Y_3, +Y_2, -Y_1, +Y_0$$

- (3) The numerical coefficient would be determined as follows:

- (a) The first numerical coefficient will be always 1. Thus in this case the term becomes $1 Y_4$ but the value 1 is not written, so it is Y_4 .

- (b) The second coefficient would be the product of numerical coefficient of first value and its subscript divided by one. Thus in the present illustration it would be $1 \times 4 / 1 = 4$, so the term would be $4 Y_3$.

(c) The third coefficient would be the product of the numerical coefficient of second value and its subscript divided by 2. Thus in our illustration it would be $4 \times 3 / 2 = 6$ and the third term would be $6 Y_3$

(d) Likewise 4th coefficient would be $6 \times 2 / 3 = 4$, the 4th term would be $4 Y_4$ and so on. The coefficients can be found out by referring Pascal's Triangle given below.

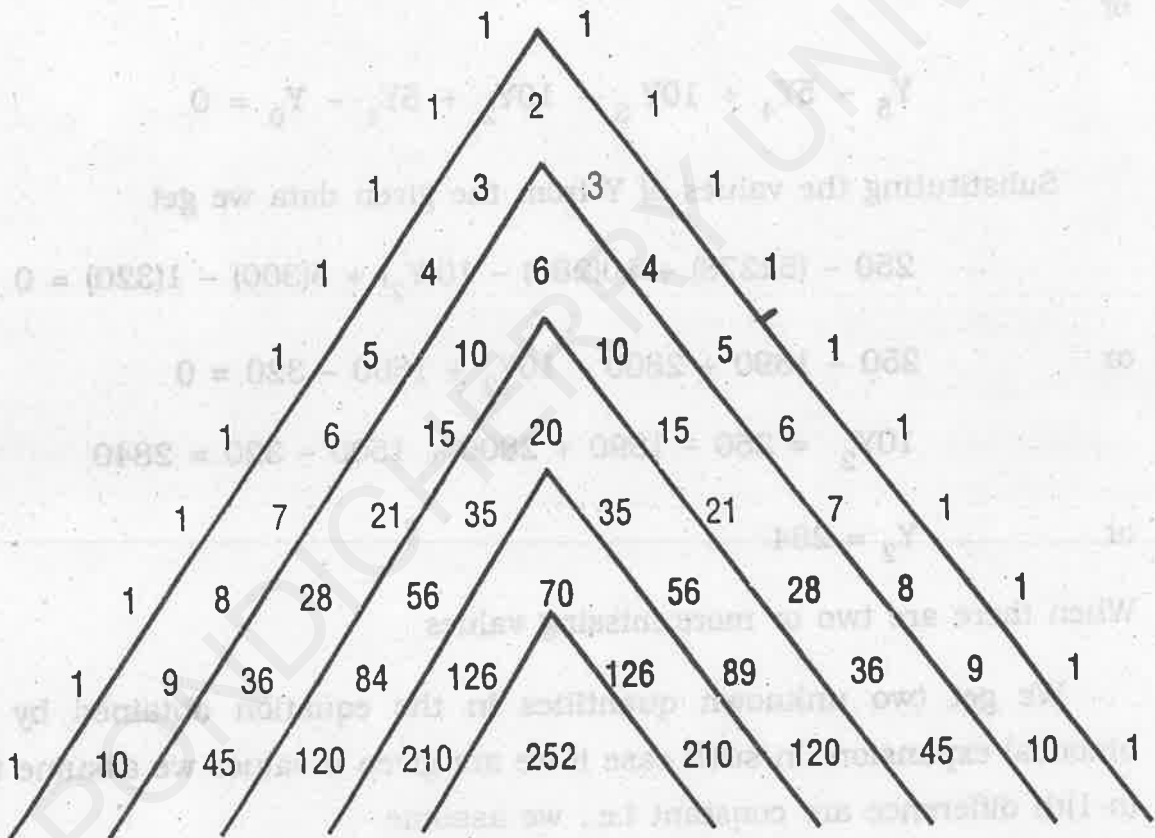


Fig. 7.1.2 Pascal's Triangle

Illustration: Find the missing figure in the following data:

Year	1970	1975	1980	1985	1990	1995
Sales of Umbrella	320	300		280	278	250

Solution

	X_0	X_1	X_2	X_3	X_4	X_5
Year	1970	1975	1980	1985	1990	1995
Sales of Umbrella	320	300	?	280	278	250
	Y_0	Y_1	Y_2	Y_3	Y_4	Y_5

Since the known value of Y are 5 therefore $\Delta_0^5 = 0$ or $(Y - 1)^5 = 0$

or

$$Y_5 - 5Y_4 + 10Y_3 - 10Y_2 + 5Y_1 - Y_0 = 0$$

Substituting the values of Y from the given data we get

$$250 - (5 \times 278) + 10(280) - 10(Y_2) + 5(300) - 1(320) = 0$$

or

$$250 - 1390 + 2800 - 10Y_2 + 1500 - 320 = 0$$

$$10Y_2 = 250 - 1390 + 2800 + 1500 - 320 = 2840$$

or

$$Y_2 = 284$$

When there are two or more missing values

We get two unknown quantities in the equation obtained by the binomial expansion. In such case if we are given n values we assume that (n-1)th difference are constant i.e., we assume

$$\Delta^{n-1} \Delta^{n-2} \Delta^{n-3} \dots \text{are constant.}$$

If (n-1)th difference is constant nth difference is zero.

i.e.,

$$\Delta_{Y_1}^n = 0, \Delta_{Y_2}^n = 0 \text{ and so on.}$$

The following example will illustrate the procedure.

Illustration: From the following data of profits of a firm (in lakh rupees) interpolate the missing figure.

Year	1965	1970	1975	1980	1985	1990	1995
Profit (in lakhs)	20	22	26	?	36	?	43
	Y_0	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6

Solution: As five figures are known we shall assume that the fifth order difference will be zero. In the problem there are two unknown figures hence two equations will be required to determine them

$$\Delta_0^5 = Y_5 - 5Y_4 + 10Y_3 - 5Y_2 + Y_0 = 0 \text{ and}$$

$$\Delta_1^5 = Y_6 - 5Y_5 + 10Y_4 - 5Y_3 + Y_1 = 0$$

Substituting the values

$$Y_5 - 5(35) + 10Y_2 - 10(26) + 5(22) - 20 = 0 \text{ ---- (1)}$$

$$43 - 5Y_5 + 10(35) - 10(Y_3) + 2(26) - 220 = 0 \text{ ---- (2)}$$

or

$$Y_5 + 10Y_2 = 345 \text{ Substituting equation (1) from (2) we get}$$

$$\frac{5Y_5 - 10Y_2 = 501}{4Y_5 = 156} \text{ or } Y_5 = 156/4 = 39$$

Substituting the values of Y_2 in the above equation

$$39 + 10Y_2 = 345$$

$$10Y_2 = 345 - 39 \text{ or } Y_2 = 306/10 = 30.6$$

Thus missing values corresponding to 1980 and 1990 are 30.6 and 39 respectively.

3. Newton's method

The newton's method can be classified into four different heads:

- (i) Newton's Advancing Difference Method
- (ii) Newton's Gauss (forward) Method
- (iii) Newton's Gauss (backward) Method
- (iv) Newton's Divided Difference Method

(i) Newton's Advancing Difference Method

The advancing difference method is applicable where the independent variable X advances by equal interval like 10, 20, 30, 40 etc. However unlike binomial method it is not necessary that the value of X to be interpolated should be one of the class intervals. For example value of $X = 25$ or 33 can be interpolated. Similarly, we can extrapolate $X = 59$.

The formula for interpolation is

$$Y_x = Y_0 + x \Delta_0^1 + \frac{x(x-1)}{2!} \Delta_0^2 + \frac{x(x-1)(x-2)}{3!} \Delta_0^3 + \frac{x(x-1)(x-2)(x-3)}{4!} \Delta_0^4 + \dots$$

where Y_0 represents the value of Y at origin.

Y_x represents the figure to be interpolated and $\Delta_0^1, \Delta_0^2, \Delta_0^3$ are the first, second and third leading differences of each column to various degree of difference. The value of x is obtained as follows:

$$x = \frac{\text{The value to be interpolated} - \text{The value at origin}}{\text{Difference between two adjoining values}}$$

When applying this method the difference between the various value of Y are to be calculated. The difference are indicated by the sign Δ . Thus the first difference would be Δ^1 second Δ^2 and Δ^3 likewise. The following is the table of differences.

TABLE OF DIFFERENCE

Table No.1

X	Y	1st Δ Δ^1		2nd Δ Δ^2		3rd Δ Δ^3		4th Δ Δ^4	
X_0	Y_0	$Y_1 - Y_0$	Δ^1_0	$\Delta^1_1 - \Delta^1_0$	Δ^2_0	$\Delta^2_1 - \Delta^2_0$	Δ^3_0	$\Delta^3_1 - \Delta^3_0$	Δ^4_0
X_1	Y_1	$Y_2 - Y_1$	Δ^1_1	$\Delta^1_2 - \Delta^1_1$	Δ^2_1	$\Delta^2_2 - \Delta^2_1$	Δ^3_1		
X_2	Y_2	$Y_3 - Y_2$	Δ^1_2	$\Delta^1_3 - \Delta^1_2$	Δ^2_2				
X_3	Y_3	$Y_4 - Y_3$	Δ^1_3						
X_4	Y_4								

Illustration: Given the following annual premium charged by LIC of India for a policy of Rs.1000. Calculate the premium payable at the age 26.

Age in Years 20 25 30 35 40

Premium (Rs.) 23 27 32 39 48

Solution: Applying Newton's method

Table No.2

Age X	Premium Y		1st Diff.		2nd Diff.		3rd Diff.		4th Diff.	
20	23	Y_0	+4	Δ^1_0	+1	Δ^2_0				
25	27	Y_1	+5	Δ^1_1	+2	Δ^2_1	1	Δ^3_0		
30	32	Y_2	+7	Δ^1_2	+2	Δ^2_2	0	Δ^3_1	-1	Δ^4_0
35	39	Y_3	+9	Δ^1_3						
40	48	Y_4								

$$Y_x = Y_0 + x \Delta_0^1 + \frac{x(x-1)}{2!} \Delta_0^2 + \frac{x(x-1)(x-2)}{3!} \Delta_0^3 + \frac{x(x-1)(x-2)(x-3)}{4!} \Delta_0^4$$

where $x = \frac{26 - 20}{5} = 1.2$

$$\begin{aligned} Y_{26} &= 23 + 1.2 \times 4 + \frac{(1.2)(1.2-1)}{1 \times 2} \times 1 + \frac{(1.2)(1.2-1)(1.2-2)}{3 \times 2} \times 1 \\ &\quad + \frac{(1.2)(1.2-1)(1.2-2)(1.2-3)}{4 \times 3 \times 2} \times -1 \\ &= 23 + 4.8 + 0.12 + (-0.032) + 0.0144 \\ &= 27.90 \end{aligned}$$

Thus the premium payable at the age 26 is Rs. 27.90.

(ii) Newton's Gauss (forward) Method

This method of interpolation is applicable when

- (a) The independent variable x advances by equal intervals and
- (b) The figure to be interpolated is in the middle of the series.

In this method the value preceding the value to be interpolated is denoted by X_0 and previous values are denoted by X_{-1} , X_{-2} , X_{-3} and value after X_0 are denoted by X_1 , X_2 , X_3 Similarly difference are denoted according to Y 's rotation

$$\Delta^1 Y_0, \Delta^2 Y_{-1}, \Delta^3 Y_{-1}, \Delta^4 Y_{-3} \dots$$

The formula for interpolation is

$$\begin{aligned} Y_x &= Y_0 + x \Delta^1 Y_0 + \frac{x(x-1)}{1 \times 2} \Delta^2 Y_{-1} + \frac{x(x-1)(x-2)}{1 \times 2 \times 3} \Delta^3 Y_{-1} \\ &\quad + \frac{x(x-1)(x-2)(x-3)}{1 \times 2 \times 3 \times 4} \Delta^4 Y_{-2} + \dots \end{aligned}$$

where $x = \frac{\text{Item to be interpolated} - \text{Preceding item}}{\text{Difference between adjusting items}}$

Illustration: The following are the annual sales of a concern for the last few years. Estimate sales for the year 1993.

Year	1988	1990	1992	1994	1996
Sales in '000 Rs.	23	26	30	35	42

Solution:

TABLE OF DIFFERENCE

Table No.3

X Year	Y Sales		1st Δ Δ^1		2nd Δ Δ^2		3rd Δ Δ^3		4th Δ Δ^4	
1988	23	Y_2	+3	$\Delta^1_{.2}$	+1	$\Delta^2_{.2}$	0	$\Delta^3_{.2}$	+1	$\Delta^4_{.2}$
1990	26	Y_1	+4	$\Delta^1_{.1}$	+1	$\Delta^2_{.1}$	1	$\Delta^3_{.1}$		
1992	30	Y_0	+5	$\Delta^1_{.0}$	+2	$\Delta^2_{.0}$				
1994	35	Y_1	+7	Δ^1_1						
1996	42	Y_2								

$$\text{The value of } x = \frac{1993 - 1992}{1994 - 1992} = \frac{1}{2} = 0.5$$

Substituting the values in Newton's Gauss forward formula we get:

$$\begin{aligned}
 Y_x &= 30 + (0.5 \times 5) + \frac{(0.5)(0.5-1)}{1 \times 2} \times 1 + \frac{(0.5)(0.5-1)(0.5-2)}{1 \times 2 \times 3} \times 1 \\
 &\quad + \frac{(0.5)(0.5-1)(0.5-2)(0.5-3)}{1 \times 2 \times 3 \times 4} \times 1 \\
 &= 30 + 2.5 + (-0.125) + 0.0625 + (-0.0391) \\
 &= 32.3984
 \end{aligned}$$

(iii) Newton's Gauss (Backward) Method

This method is applicable when

- (a) X series advances by equal intervals

(b) When the figures to be interpolated is near the end of the series

The formula is

$$Y_x = Y_0 - X\Delta^1 Y_{-1} - \frac{(x+1)X}{1 \times 2} \Delta^2 Y_{-1} + \frac{(x+1) X (x-1)}{1 \times 2 \times 3} \Delta^3 Y_{-2} \\ + \frac{(x+1) X (x-1) (x-2)}{1 \times 2 \times 3 \times 4} \Delta^4 Y_{-2}$$

where Y_0 is the figure succeeding the missing figure and

$$x = \frac{\text{Item succeeding the item to be interpolated} - \text{Item to be interpolated}}{\text{Difference between adjoining items}}$$

Illustration: The following table gives some values of x and y variable. Interpolate the values of y when x is 45.

X	10	20	30	40	50	60
Y	22	26	32	41	47	51

Solution: The figure to be interpolated is towards the end of the table hence Newton's Gauss (backward) method is used.

TABLE OF DIFFERENCE

Table No.4

[illegible]

of X series advances by equal intervals

$$Y_x = Y_0 - X \Delta^1 Y_{-1} + \frac{(x+1)X}{1 \times 2} \Delta^2 Y_{-1} + \frac{(x+1)X(x-1)}{1 \times 2 \times 3} \Delta^3 Y_{-2} \\ + \frac{(x+1)X(x-1)(x-2)}{1 \times 2 \times 3 \times 4} \Delta^4 Y_{-2}$$

$$= 45 - 3 - 0.375 + 0.125$$

$$= 41.75$$

(iv) Newton's Divided Difference Method

This method is to be used when the value of the independent variable x advances by unequal intervals. The formula is

$$Y_x = Y_0 + (X - X_0) \Delta^1 0 + (X - X_0)(X - X_1) \Delta^2 0 \\ + (X - X_0)(X - X_1)(X - X_2) \Delta^3 0 + \dots$$

where $\Delta^1 0$, $\Delta^2 0$, $\Delta^3 0$ are first, second and third leading divided differences respectively. The divided difference are obtained in the manner as follows.

Table No.5

X	Y	Divided Difference		
		1st Δ Δ^1	2nd Δ Δ^2	3rd Δ Δ^3
X_0	Y_0	$\frac{Y_1 - Y_0}{X_1 - X_0} \quad \Delta^1_0$	$\frac{-\Delta^1_1 - \Delta^1_0}{X_2 - X_0} \quad \Delta^2_0$	$\frac{-\Delta^2_1 - \Delta^2_0}{X_3 - X_0} \quad \Delta^3_0$
X_1	Y_1	$\frac{Y_2 - Y_1}{X_2 - X_1} \quad \Delta^1_1$	$\frac{-\Delta^1_2 - \Delta^1_1}{X_3 - X_1} \quad \Delta^2_1$	
X_2	Y_2	$\frac{Y_3 - Y_2}{X_3 - X_2} \quad \Delta^1_2$		
X_3	Y_3			

Illustration: The observed values of a function are respectively 168, 120, 72, 63 at four positions 4, 9, 12, 16 of independent variable. What best estimate can you give for the value of the function at the position 6 of the independent variable?

Solution: Since the independent variable is advancing by unequal interval Newton's divided difference method is used.

TABLE-6

X		Y		1st Δ Δ^1	2nd Δ Δ^2	3rd Δ Δ^3
4	X ₀	168	Y ₀	$\frac{120-168}{5}$ -9.6 Δ^1_0	$\frac{-16-(-9.6)}{12-4}$ 0.8 Δ^2_0	$\frac{-1.96-(-.8)}{16-4}$ =0.23
9	X ₁	120	Y ₁	$\frac{72-120}{3}$ -16 Δ^1_1	$\frac{2.256-(-16)}{16-9}$	
12	X ₂	72	Y ₂	$\frac{63-72}{4}$ -2.25 Δ^1_2		
16	X ₃	63	Y ₃			

$$Y_x = Y_0 + (X - X_0) \Delta^1_0 + (X - X_0)(X - X_1) \Delta^2_0$$

$$+ (X - X_0)(X - X_1)(X - X_2) \Delta^3_0$$

$$= 168 + (6-4)(-9.6) + (6-4)(6-9)(-0.8) + (6-4)(6-9)(6-12)(.23)$$

$$= 168 + (-19.2) + (2X - 3X - 0.8) + (2X - 3X - 6X) (.23)$$

$$= 168 - 19.2 + 4.8 + 8.28$$

$$= 161.88$$

Langrang's method

Like Newton's divided difference method this formula is also used when x series does not advance by equal interval. This also can be used when x series advances by equal intervals. The Langrang's formula is

$$\begin{aligned}
 Y_x = & Y_0 \frac{(X - X_1)(X - X_2)(X - X_3)(X - X_4) \dots (X - X_n)}{(X_0 - X_1)(X_0 - X_2)(X_0 - X_3)(X_0 - X_4) \dots (X_0 - X_n)} \\
 & + Y_1 \frac{(X - X_0)(X - X_2)(X - X_3)(X - X_4) \dots (X - X_n)}{(X_1 - X_0)(X_1 - X_2)(X_1 - X_3)(X_1 - X_4) \dots (X_1 - X_n)} \\
 & + Y_2 \frac{(X - X_0)(X - X_1)(X - X_3)(X - X_4) \dots (X - X_n)}{(X_2 - X_0)(X_2 - X_1)(X_2 - X_3)(X_2 - X_4) \dots (X_2 - X_n)} \\
 & + Y_n \frac{(X - X_0)(X - X_1)(X - X_2) \dots (X - X_{n-1})}{(X_n - X_0)(X_n - X_1)(X_n - X_2) \dots (X_n - X_{n-1})}
 \end{aligned}$$

where Y_x is the figure to be interpolated, X is the value of x series for which Y_x is to be obtained. $X_0, X_1, X_2, X_3 \dots X_n$ are given values of X variable. Y_0, Y_1, Y_2, Y_3, Y_n are the corresponding given values of Y variable. The following example would illustrate the formula.

Illustration: The following table gives the normal weight of a baby during the first six months of life

Age in months	0	2	3	5	6
Weight in lbs	5	8	9	12	14

Estimate the weight of a baby at the age of 4 months.

Solution:

X	Y
0 X_0	5 Y_0
2 X_1	8 Y_1
3 X_2	9 Y_2
5 X_3	12 Y_3
6 X_4	14 Y_4

By substituting the values in Lagrang's formula we get

$$\begin{aligned}
 & \frac{(4-2)(4-3)(4-5)(4-6)}{(0-2)(0-3)(0-5)(0-6)} \times 5 + \frac{(4-0)(4-3)(4-5)(4-6)}{(2-0)(2-3)(2-5)(2-6)} \times 8 \\
 & + \frac{(4-0)(4-2)(4-5)(4-6)}{(3-0)(3-2)(3-5)(3-6)} \times 9 + \frac{(4-0)(4-2)(4-3)(4-6)}{(5-0)(5-2)(5-3)(5-6)} \times 12 \\
 & + \frac{(4-0)(4-2)(4-3)(4-5)}{(6-0)(6-2)(6-3)(6-5)} \times 14 \\
 & = \\
 & \frac{2x1x - 1x - 2}{-2x - 3x - 5x - 6} \times 5 + \frac{4x 1x - 1x - 2}{2x - 1x - 3x - 4} \times 8 + \frac{4x 2x - 1x - 2}{3x 1x - 2x - 3} \times 9 \\
 & = \frac{4x 2x 1x - 2}{5x 3x 2 x1} \times 12 + \frac{4x 2x 1x - 1}{6x 4x 3x 1} \times 14 \\
 & \frac{4}{180} \times 5 + \frac{8}{-24} \times 8 + \frac{16}{18} \times 9 + \frac{-16}{-30} \times 12 + \frac{-8}{72} \times 14 \\
 & = 0.11 + (-2.67) + 8 + 6.4 - 1.55 \\
 & = 10.29
 \end{aligned}$$

Thus the weight of the baby at the age of 4 months will be 10.29 pounds.

Parabolic Curve Method

This method of interpolation is also known as method of simultaneous equation. This method is based on the assumption that the values of X and Y are interdependent and the values of X are known. The variable Y is taken as dependent variable and X as independent. Consequently for a given X we can find out the value of Y. The equation of this curve is

$$Y = a + bx + cx^2 + dx^3 + ex^4 \dots kx^n$$

This equation represents a parabolic curve of nth degree and a,b,c,d etc are the constants. The power to which this equation is to be raised

depends upon the number of known quantities. The curve is raised to the power 'one' less than the number of known quantities. For example if known quantities are 4 we would take a curve of 3rd order.

$$Y = a + bx + cx^2 + dx^3$$

Illustration: Estimate the profits for the year 1993 from the following data:

Year	1991	1992	1994
Profit	8.5	12	10

Solution: Value of X by taking deviation from 1993 and the value of Y are

Year	1991	1992	1993	1994
X	-2	-1	0	1
Profit	8.5	12	Y_0	10
Y				

Since the known values are 3 we raise a parabola of second order.

$$Y = a + bx + cx^2$$

Substituting X and Y we get

$$8.5 = a - 2b + 4c \quad \text{-----} \quad (1)$$

$$12 = a - b + c \quad \text{-----} \quad (2)$$

$$Y = a \quad \text{-----} \quad (3)$$

$$10 = a + b + c \quad \text{-----} \quad (4)$$

Adding (2) and (4) we get

$$22 = 2a + 2c \quad \text{-----} \quad (5)$$

Multiplying equation (4) by 2 and adding it with equation (1)

$$20 = 2a + 2b + 2c$$

$$8.5 = a - 2b + 4c$$

$$28.5 = 3a + 6c \quad \text{-----} \quad (6)$$

Multiplying equation (5) by (3) and subtracting equation (6) from it we get

$$66 = 6a + 6c$$

$$28.5 = 3a + 6c$$

$$\underline{37.5 = 3a}$$

$$\text{or } a = 12.5$$

Thus profit for the year 1993 would be 12.5 lakhs.

EXTRAPOLATION

As pointed out earlier extrapolation refers to estimating a value beyond the given values of Y . It is a futuristic estimation. There is no specific formula for extrapolation. All the various methods discussed in interpolation can be adopted for extrapolation. The choice of method would depend on the nature of data given. The conditions applicable with different formula for interpolation hold good for extrapolation also.

Thus if we are given population figure of India for 1941, 49, 51, 61, 71, 81 and we have to extrapolate 1991 we will use binomial expansion method. If the population figure has to be extrapolated for 86 we can use Newton's methods of advancing difference.

LESSON - 2

TIME SERIES ANALYSIS

- ❑ INTRODUCTION
- ❑ DEFINITION
- ❑ UTILITY OF TIME SERIES ANALYSIS
- ❑ VARIATION IN TIME SERIES
- ❑ TECHNIQUES OF MEASUREMENT

INTRODUCTION

Time series refers to the chronologically ordered values of a variable over successive time periods. Thus time series refers to such a series in which time is one variable. The analysis of such figures chronologically arranged over the years successively are called analysis of time series.

The essential requirements of a time series are:

1. It must consist of a homogenous set of data
2. Data should be available for a sufficiently long period say 7 to 10 years or relevant time period
3. The time gap between various values must, as far as possible be equal.
4. The gaps, if any, in the data must be made up by interpolation.

DEFINITION

The following are some definitions which will clarify the concept of time series.

"A time series consists of data arranged chronologically" – Croxton and Cowden.

"A set of data depending on the time is called time series" – Kenny and Keeping.

A time series may be defined as a sequence of values of some variable corresponding to successive points in time" – W. Herisch.

"Time series consists of statistical data which are collected, recorded and observed over successive increments" – Patterson.

"A time series may be defined as a sequence of repeated measurements of a variable made periodically through time".

Utility of Time Series

The following are the uses of time series analysis:

1. It helps in the analysis of past behaviour of a variable – That is the effect of various factors like technological, economic etc., on a variable over a period of years can be studied by the help of time series analysis.
2. It helps in forecasting – The analysis of past condition becomes the basis for forecasting the behaviour of the variable in future. This helps in making future plans of action.
3. It helps in evaluating the achievements – The review and evaluation of progress in any field of economic and business activity is largely done on the basis of time series data.
4. It helps in making comparative studies – As in time series the data are chronologically arranged, it facilitates comparison between one period of time with that of the other. It provides a scientific basis for making comparisons.
5. It helps in dissecting the data into various components. These components called by various names as seasonal variation, cyclical variation and irregular variation throw light on the economic behaviour pattern.

Variation in Time Series

The term time series is used to refer to any group of statistical information collected at regular intervals of time. There are four kinds of changes or variation involved in time series analysis. They are

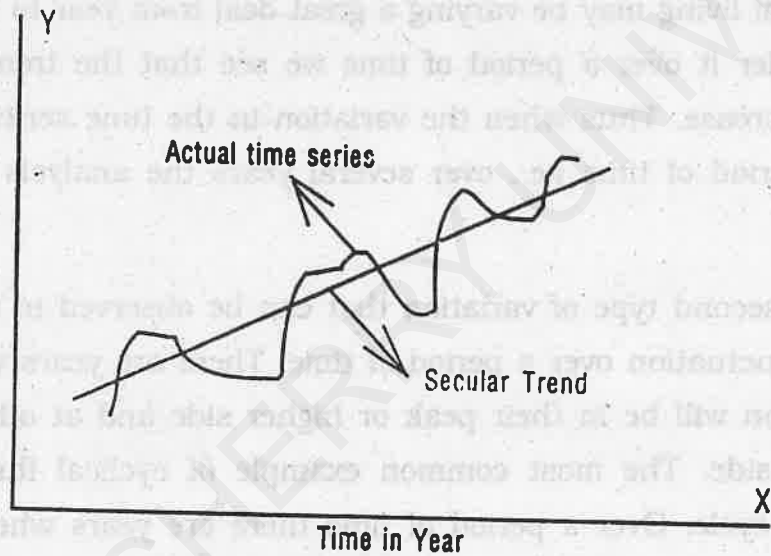
1. Secular trend
2. Cyclical fluctuation
3. Seasonal variation
4. Irregular variation

With the secular trend the value of the variable tends to increase or decrease over a long period of time. The steady increase in the cost of living recorded by the consumer price index is an example of secular trend. The cost of living may be varying a great deal from year to year. But when we consider it over a period of time we see that the trend is towards a steady increase. Thus when the variation in the time series is studied for a long period of time i.e., over several years the analysis is called trend analysis.

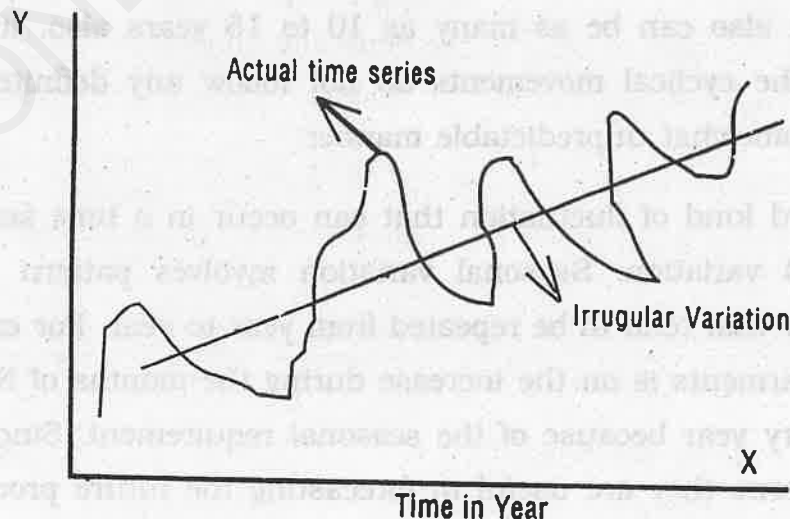
The second type of variation that can be observed in a time series is cyclical fluctuation over a period of time. There are years when numerical information will be in their peak or higher side and at other time will be in lower side. The most common example of cyclical fluctuation is the business cycle. Over a period of time there are years when the business cycle has a peak above the trend line and at other times the activity is likely to slump, touching the low point below the trend line. The time between touching the peak and falling to the low points is usually 3 to 5 years, but it also can be as many as 10 to 15 years also. It should be noted that the cyclical movements do not follow any definite trend but move in a somewhat unpredictable manner.

The third kind of fluctuation that can occur in a time series data is the seasonal variation. Seasonal variation involves pattern of changes within a year that tend to be repeated from year to year. For example sale of woollen garments is on the increase during the months of November to January every year because of the seasonal requirement. Since there are regular patterns they are useful in forecasting the future production run etc.

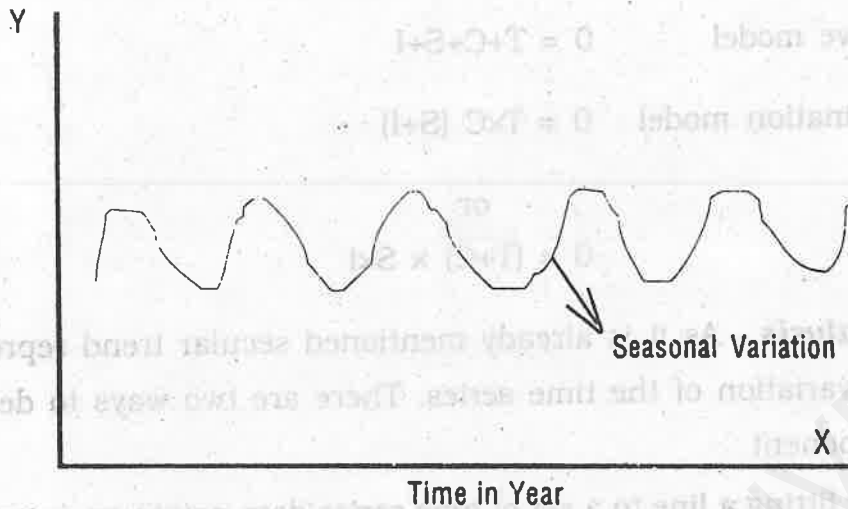
Irregular variation is the fourth type of change that can be observed in a time series data. These variations may be due to (1) random fluctuations, no one of which is significantly important to warrant singling out for individual treatment (ii) non-recurring irregular influences that exerts a significant one time impact on the behaviour of time series and as such must be explicitly recognised. The events like flood, strike, war, earthquake etc., which influence the time series data. The above four variations are shown diagrammatically below.



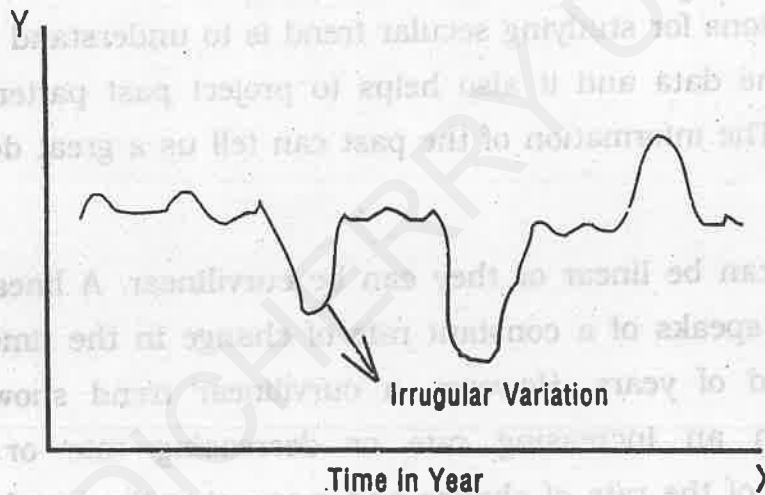
Graph 7.2.1



Graph 7.2.2



Graph 7.2.3



Graph 7.2.4

The above four variations are generally considered as interacting in multiplicative manner to produce observed values of the overall time series

Multiplicative model: $O = T \times C \times S \times I$

O = observed values of time series

T = trend component

C = cyclical component

S = seasonal component

I = Irregular component

Other type of models that are possible are

Additive model $O = T+C+S+I$

Combination model $O = T \times C (S+I)$

or

$$O = (T+C) \times S \times I$$

Trend Analysis - As it is already mentioned secular trend represents the long term variation of the time series. There are two ways to describe the trend component

- (1) by fitting a line to a set of time series data points on a graph
- (2) by fitting a trend line by the method of least square

The reasons for studying secular trend is to understand the historical pattern in the data and it also helps to project past patterns or trends into future. The information of the past can tell us a great deal about the future.

Trends can be linear or they can be curvilinear. A linear trend or a straight line speaks of a constant rate of change in the time series data over a period of years. However, a curvilinear trend shows the trend increasing in an increasing rate or decreasing rate or vice versa. Computation of the rate of change and measuring the trend is a process known as fitting a curve to the data. Basically there are four methods for fitting the trend in time series. These are

- (i) free hand method
- (ii) method of semi average
- (iii) method of moving average
- (iv) method of least square

Fitting a trend curve involves assuming that a given time series exhibits a certain trend movement. Measuring a trend actually means computing the constants of the equation that we have chosen to be representative of the trend in the data.

(i) Free hand method

This method is otherwise called as graphic method in the sense that the trend line is determined by inspecting the graph of the series. According to this method the trend values are determined by drawing free hand straight line through the time series data that is judged by the analyst to represent adequately the longterm movement in the series. Once the free hand line is drawn, a trend equation for the line is approximated. This is done by first reading of the trend values of the first and the last period from the chart with reference to the freehand line. For this purpose the first period is usually considered the origin. Thus the trend value of first period is the value of 'a' for the equation. The difference between the trend value of first period and last period divided by number of years gives the 'b' value of the equation. This method of finding the trend values and trend line equation is very simple, easy and direct but it suffers from the limitations like for the same series of data different people may draw different lines; even one person may draw different trend lines in different times. In addition there is no formal statistical criterion to judge the adequacy of such a line. For this reason mostly the freehand curve is not recommended for fitting a trend line.

(ii) Semi average method

To determine the trend values by semi average method, the series in question is first divided into two equal segments then the arithmetic mean of each part is computed. Then a straight line is drawn through the two arithmetic means plotted in the graph to get the trend line. Each average provides the trend value for the middle time period of the corresponding segment. When the time series has even number of years dividing the total time period into two parts is not difficult, but when it has odd number of years there are three methods for separating the series.

- (a) Add half of the value of middle period to the total values of each part.
- (b) Add the total value of middle period to the total value of each part.
- (c) Drop the value of middle period from the computation of the averages.

With semi average method the middle time unit is considered as the origin and the values of Y intercept and slope of the straight line are derived by applying the following equations

$$a = \frac{S_1 + S_2}{t_1 + t_2}$$

$$b = \frac{S_2 - S_1}{t_1 (n - t_2)}$$

where t_1 and t_2 refer to the number of time units for first and second segment in the series, S_1 and S_2 refer to the corresponding partial sum respectively and 'n' is the total number of periods in the series.

Illustration: Compute the trend by semi average method for the data relating to number of persons registered in an employment exchange in Pondicherry during 1981 to 1995.

Year	No. of persons in thousands
1981	10.5
1982	15.3
1983	13.5
1984	12.9
1985	11.1
1986	15.9
1987	16.0
1988	16.5
1989	16.0
1990	16.4
1991	19.9
1992	21.7
1993	18.7
1994	18.6
1995	21.5

Solution:

Years	X	No. of persons	Semi-average	Trend Value
1981	-7	10.5		11.6
1982	-6	15.3		12.3
1983	-5	13.5		12.9
1984	-4	12.9	13.6	13.6
1985	-3	11.1		14.3
1986	-2	15.9		14.9
1987	-1	16.0		15.6
1988	0	16.5		16.3
1989	1	16.0		17.0
1990	2	16.4		17.6
1991	3	19.9		18.3
1992	4	21.7	19.0	19.0
1993	5	18.7		19.6
1994	6	18.6		20.3
1995	7	21.5		21.0

The series contains 15 years and is divided into two parts with 7 years in each and the middle year being dropped. The arithmetic mean for the first half is 13.6 and that for second is 19.0. From this two average values we get

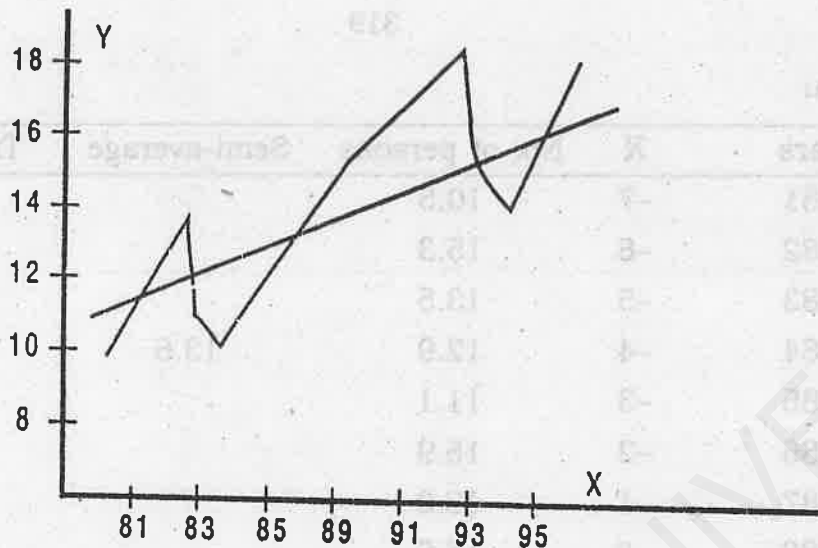
$$a = \frac{95.2 + 132.8}{7 + 7} = 16.3$$

$$b = \frac{132.8 - 95.2}{7(15 - 7)} = 0.67$$

Thus the trend equation becomes

$$Y = 16.3 + 0.67x$$

Trend value of each year can now be determined substituting the X value for that period. The straight line trend can be drawn through the two point 13.6 and 19.



Graph 7.2.5

This method of determining the trend is not a subjective one. The slope of the trend line depends on the difference between the averages that are computed from the original values with each average typical of the level of that segment of the data. The major drawback of this method is that the arithmetic means may be unduly affected by the extreme values in the series. Thus in the presence of extreme values in the time series data the trend line may not be the true representation of the secular movements of the series. Therefore the trend values obtained by this method are not accurate enough for the purpose of forecasting the future trend.

(iii) Method of moving average

Another method that is used for determining the trend component in a time series is the method of moving average. This method may be considered as an artificially constructed time series in which each periodic figure is replaced by the mean of the value of that period and those of a number of preceding and succeeding periods. The computation of moving average is simple and straightforward.

The steps in calculating the moving averages are

1. Compute the moving totals for the number of years asked. If it is a three yearly moving average the value of first three years value is added and written in the center year i.e., against second year. This is

first three year moving total. Then the first year value is deleted and fourth year value is included to form the second three year moving total which is centred at the third year i.e., the total of the value of 2nd, 3rd and 4th year is written against 3rd year. In a similar way the computation moves through the end of the series.

2. The moving average is obtained by dividing the moving totals calculated in the previous step by the number of years moving average asked. That is if it is a three yearly moving average the 3 yearly moving total is divided by 3.

However it is to be noted that in computing moving average for an even number of periods, the procedure is slightly complicated. For example calculation of 4 yearly moving average starts with adding up the first four years' value in the series to form four yearly moving total. The second moving total is obtained by dropping the value of first period from and adding the value of fifth period to the first four year moving total and so on until all the moving totals have been obtained from the series. Then each moving total is divided by the number of year average (as in this case it is 4). Thus the moving totals are divided by four and 4 yearly moving average is obtained. However the moving totals and moving average so obtained fall between two periods i.e., in this case it falls between 2nd and 3rd year. Whereas data that are typical of a period should be written against the particular year. As the moving averages are written not against a particular year but rather in between two years, we need to centre it i.e., from the averages which is written between 2nd and 3rd as well as 3rd and 4th year we can get a centred average against 3rd year by adding the two averages and dividing it by two. Thus the first centred moving average will fall against 3rd year and second centred moving average will be against 4th year and so on if we are calculating 4 yearly moving average.

A moving average of equal length period eliminates the periodic fluctuations and the moving average of equal length will be linear if the series changes on the average by a constant per time unit and its fluctuations are periodic.

Illustration: (odd period of moving average)

Find the trend value of the following data by the moving average method (Take 3 years cycle).

Year	Population in million
1989	512
1990	519
1991	538
1992	548
1993	560
1994	575
1995	590
1996	599

Solution:

Year	Population in million	3 Yearly moving total	3 Yearly moving average
1989	512		
1990	519	1569	523
1991	538	1605	535
1992	548	1646	549
1993	560	1683	561
1994	575	1725	575
1995	590	1764	588
1996	599		

Illustration: (Even period of moving average)

From the following series of observations find out 4 yearly moving average.

Year	1989	90	91	92	93	94	95	96
Annual sales	3	7	1	6	4	9	8	3
(Rs.'0000)								

Solution:

Year	Sales in '0000	4 yearly moving total not centred	4 yearly moving average	4 yearly moving average centred
1989	3			
1990	7			
	————→	17	4.25	
1991	1			
	————→	18	4.5	4.375
1992	6			
	————→	19	3.8	6.4
1993	4			
	————→	27	6.75	7.17
1994	9			
	————→	24	6	6.39
1995	8			
1996	3			

Moving average method constitutes a satisfactory trend for a series that is basically linear and that is regular in duration and amplitude is a useful technique for analysing the time series data. However the limitations of moving average method are that in computing moving averages we lose some years at the beginning and end of the series. Another drawback of the method is that the moving average is not represented by any mathematical formula and therefore is not capable of objective future projection. Since the major objective of trend analysis is that of forecasting, moving average is not used as a trend measure.

(iv) Method of least square

The earlier discussed methods for trend analysis have certain defects particularly providing a satisfactory projection of the future. To overcome this defect a convenient method is to follow a mathematical approach. The device of getting an objective fit of a straight line to a series of data is the least square method. The line fitted by the method means that the estimates of the constants a and b are the best linear unbiased estimates of those constants.

To determine the value of a and b in a linear equation by least square method we are required to solve the following two normal equations simultaneously

$$\Sigma y = Na + b\Sigma x$$

$$\Sigma xy = a \Sigma x + b \Sigma x^2$$

In the case of time series analysis the solution of a and b from these equation is simplified by using the middle of the series as the origin. Since the time units in a series are usually of uniform duration and are consecutive numbers then the middle point is taken as the origin the sum of time units ie Σx will be zero. As a result the above two normal equations reduce to

$$\Sigma y = Na$$

$$\Sigma xy = b \Sigma x^2$$

Therefore we can get

$$a = \frac{\sum y}{N} \text{ and } b = \frac{\sum xy}{\sum x^2}$$

It should be noted that in computing the trend it is convenient to use the middle of the series as the origin and get Σx as zero. If the series contains an odd number of periods the origin is the middle of the given period. If an even number of periods is involved the origin is set between the two middle periods.

Illustration: Given below are the figures of production (in thousand tons) of a sugar factory.

Year	1990	91	92	93	94	95	96
Production in '000 tons	40	47	49	42	45	50	52

Fit a straight line trend by least square method and tabulate the trend value.

Solution:

Year	X (taking 93 as origin)	Y	X^2	XY	Trend Value
1990	-3	40	9	-120	42.35
1991	-2	47	4	-94	43.71
1992	-1	49	1	-49	45.07
1993	0	42	0	0	46.43
1994	1	45	1	45	47.79
1995	2	50	4	100	49.15
1996	3	52	9	156	50.51
$\Sigma x = 0$		$\Sigma y = 325$	$\Sigma x^2 = 28$	$\Sigma xy = 38$	

$$N = 7; \Sigma y = 325; \Sigma xy = 38; \Sigma x^2 = 28$$

$$a = \Sigma y / N = 325 / 7 = 46.43$$

$$b = \Sigma xy / \Sigma x^2 = 38 / 28 = 1.36$$

Illustration: Fit a straight line trend equation by the least square method and estimate the trend values.

Year	1988	89	90	91	92	93	94	95
Value	80	90	92	83	95	99	94	106

Solution: In this case since N is 8 (even number) we have the origin to the time which is the arithmetic mean of two middle terms 1991 and 1992.

Year	X (Deviation from origin 91.5)	Y (Deviation multiplied by 2)	XY	X ²	Y _c	
1988	-3.5	-7	80	-560	49	87.475
1989	-2.5	-5	90	-450	25	88.875
1990	-1.5	-3	92	-276	9	90.275
1991	-0.5	-1	83	-83	1	91.675
1992	0.5	1	95	95	1	93.075
1993	1.5	3	99	297	9	94.475
1994	2.5	5	94	470	25	95.875
1995	3.5	7	106	742	49	97.275
		$\Sigma x=0$	$\Sigma y=739$	$\Sigma xy=235$	$\Sigma x^2=168$	

$$Y = a + bx; \quad a = \Sigma y/N; \quad b = \Sigma xy/\Sigma x^2$$

$$a = 739/8 = 92.375$$

$$b = 235/168 = 1.40$$

The merits of this method is that it is a mathematical method and it does not have any subjectivity. Trend values can be obtained for all the given time periods in the series, which is not possible in other methods like moving average and it is useful for future prediction. However this method is tedious, time consuming as well as very rigid. In this method only long term variation can be studied and the impact of cyclical, seasonal and irregular variations are ignored.

Measurement of Seasonal Index

By seasonal variation in a time series we mean the variations of regular and periodic nature with period less than one year. If the variation is for longer than one year period they will be considered under cyclical variation. The variations can be measured in terms of absolute amounts of change through additive process or in terms of relative seasonal factor in rates or percentages. The elimination of seasonal variation from time series is known as deseasonalisation. The deseasonalisation helps in the process of

decomposition of a time series into various components viz., trend, seasonal variation, cyclical variation and irregular or random variation.

The following are important methods of studying seasonal variations:

1. Method of simple average
2. Ratio to trend method
3. Ratio to moving average method
4. Link relative method

1. Method of simple average

The procedure of computing seasonal index by this method is

- (1) Arrange the data by year, months or quarters as the case may be.
- (2) Compute the arithmetic average for each period i.e., month, year or quarter.
- (3) Obtain the overall average X for a month, or quarter from all monthly or quarterly averages obtained in step (2).

$$\text{ie. } X = \frac{X_1 + X_2 + X_3 + \dots + X_{12}}{12} \quad \text{or} \quad X = \frac{X_1 + X_2 + X_3 + X_4}{4}$$

- (4) Seasonal indices for different months or quarters are obtained by expressing each monthly (quarterly) averages as a percentage of overall average X

Seasonal index for k th month (or quarter)

$$\frac{\text{Monthly Average}}{\text{Total Average}} \times 100 \quad \text{i.e.} \quad \frac{X_k}{X} \times 100$$

where $k = 1, 2, \dots$ to 12 & $k (1..to 4)$

Illustration: Use the method of monthly average to determine the quarterly indices from the following data of production of a certain commodity for the year 1993, 1994, 1995 and 1996.

Year	1st quarter	2nd quarter	3rd quarter	4th quarter
1993	33	44	40	37
1994	42	47	42	39
1995	44	48	39	36
1996	50	54	42	34

Solution:

Year	1st quarter	2nd quarter	3rd quarter	4th quarter
1993	33	44	40	37
1994	42	47	42	39
1995	44	48	39	36
1996	50	54	42	34
Total	169	193	163	146
Average	42.25	48.25	40.75	36.50
Seasonal index	100.74	115.05	97.17	87.03

$$\text{The average of averages} = \frac{42.25 + 48.25 + 40.75 + 36.5}{4}$$

$$= 41.938$$

$$\text{Seasonal index} = \frac{\text{quarterly average}}{\text{grand average}} \times 100$$

$$\text{Thus seasonal index of 1st quarter} = \frac{42.25}{41.938} \times 100 = 100.74$$

$$\text{Thus seasonal index of 2nd quarter} = \frac{48.25}{41.938} \times 100 = 115.05$$

$$\text{Thus seasonal index of 3rd quarter} = \frac{40.75}{41.938} \times 100 = 97.17$$

$$\text{Thus seasonal index of 4th quarter} = \frac{36.50}{41.938} \times 100 = 87.03$$

Ratio to Trend Method

This method is an improvement over the average method. This method also is otherwise known as percentage to trend method. This method assumes that seasonal variation for various seasons (month for monthly data, quarter for quarterly data) is a constant factor of trend. This method takes into cognizance of the effect of trend on time series. The steps involved in computation of seasonal indices by this method are as follows:

1. Obtain the trend values by least square method by season wise.
2. Divide the original data by corresponding trend values and multiply these ratios by 100. Thus we express the original data as a percentage to trend value.
3. The effect of cyclical and irregular movements are eliminated by the process of averaging the percentage for each unit of time. Either arithmetic mean or median can be used for averaging. Thus season wise figures for various years are averaged and this average gives the preliminary indices of seasonal variation.
4. The preliminary indices obtained in step (3) are adjusted to a total of 1200 for monthly data or 400 for quarterly data by multiplying each of them by a constant factor 'k' given by

$$k = \frac{1200}{\text{Sum of monthly indices}} \text{ and } \frac{400}{\text{Sum of quarterly indices}}$$

Illustration: Find the seasonal variation by ratio to trend method from the data given below.

Year	1st quarter	2nd quarter	3rd quarter	4th quarter
1991	40	50	46	44
1992	44	62	60	54
1993	50	68	64	58
1994	64	86	78	72
1995	90	102	96	92

Calculation of trend by least square method.

Year	X (Deviation from middle year 93)	X ²	Y Yearly total	Yearly average	XY	Trend value
1991	-2	4	180	45	-90	42
1992	-1	1	220	55	-55	54
1993	0	0	240	60	0	66
1994	1	1	300	75	75	78
1995	2	4	380	95	190	90
<u>N = 5</u>		<u>$\Sigma X^2 = 10$</u>		<u>$\Sigma Y = 330$</u>	<u>$\Sigma XY = 120$</u>	

The equation of straight line trend is $Y = a + bx$

$$a = \Sigma y / N \quad b = \Sigma xy / \Sigma x^2$$

$$a = 330/5 = 66 \quad b = 120/10 = 12$$

$$\text{Quarterly increment} = 12/4 = 3$$

Now we calculate quarterly trend value; consider 1991 the trend value of the middle quarter is 42 i.e., half of 2nd and half of 3rd quarter. Quarterly increment is 3 so the trend value of 2nd quarter is $42 - 3/2 = 40.5$ and 3rd quarter is $42 + 3/2 = 43.5$ trend value of first quarter is $40.5 - 3 = 37.5$ and 4th quarter is $43.5 + 3 = 46.5$.

We thus get quarterly values as shown below:

Year	1st quarter	2nd quarter	3rd quarter	4th quarter
1991	37.5	40.5	43.5	46.5
1992	49.5	52.5	55.5	58.5
1993	61.5	64.5	67.5	70.5
1994	73.5	76.5	79.5	82.5
1995	85.5	88.5	91.5	94.5

The given values are expressed as percentage of the corresponding trend value. Thus the value for first quarter will be $(40/37.5 \times 100) = 106.67$ for 2nd quarter $(50/40.5 \times 100) = 123.46$ etc.

Quarterly values as percentage to trend value would be

Year	1st quarter	2nd quarter	3rd quarter	4th quarter
1991	106.67	123.46	105.75	94.62
1992	88.89	118.09	108.11	92.31
1993	81.30	105.43	94.81	82.69
1994	87.07	112.42	91.11	87.27
1995	105.26	115.25	109.92	97.35
Total	469.19	574.65	504.7	454.24
Average	93.84	114.93	100.94	90.85
SI adjusted	93.71	114.77	100.80	90.72

Total of averages = $93.84 + 114.93 + 100.94 + 90.85 = 400.56$

Since the total of average is more than 400 an adjustment is made by multiplying each average by $400/400.56$.

Ratio to moving average method is a more logical procedure for measuring seasonal index. This method eliminates both trend and cyclical fluctuations from the time series. But if cycles are not regular and are of different intensity seasonal index calculated by above method would contain some effect of cyclical variation.

Ratio to moving average method

This method is also known as percentage of moving average method. The computation of seasonal indices by this method is similar to that of ratio to trend method except that in place of least square trend, moving average trend is used. The various steps involved in the computation of seasonal indices by the ratio to moving average method are as follows:

- (1) Obtain 12 monthly (4 quarterly) centered moving averages for the given series

- (2) Express each original value on the time series as a percentage of the trend value
- (3) Arrange these percentages season-wise for all the years and average it. These would be preliminary seasonal indices.
- (4) If the sum of the indices is not 1200 (for monthly) or 400 (for quarterly) figures multiply them by the correction factor (c.f.) $1200/\text{Sum of monthly indices}$ or $400/\text{sum of quarterly indices}$ as the case may be. These are the ratio to moving average seasonal indices.

Illustration: Calculate the seasonal indices by the ratio to moving average method from the following data

Year	1st quarter	2nd quarter	3rd quarter	4th quarter
1993	70	64	63	65
1994	67	60	68	63
1995	70	65	65	69

Solution:

Year	Quarter	Given figure	4 figure moving total	2 figure moving total	4 figure moving average	Given figure % to moving average
1993	I	70				
	II	64	→ 262			
	III	63	→ 259	521	65.125	96.73
	IV	65	→ 255	514	64.250	101.67
1994	I	67	→ 260	515	64.375	104.08
	II	60	→ 258	518	64.750	92.66
	III	68	→ 261	519	64.875	104.82

	IV	63	→	527	65.875	95.64
		→	266			
1995	I	70	→	529	66.125	105.86
		→	263			
	II	65	→	532	66.500	97.74
		→	269			
	III	65				
	IV	69				

Calculation of seasonal index (Trend eliminated values)

Year	1st quarter	2nd quarter	3rd quarter	4th quarter
1993	—	—	96.73	101.67
1994	104.08	92.66	104.82	95.64
1995	105.86	97.74	—	—
Total	209.94	190.40	201.55	197.31
Average	104.97	95.20	100.76	98.66
SI adjusted	105.08	95.30	100.86	98.76

$$\text{Arithmetic average of averages} = \frac{104.97 + 95.20 + 100.76 + 98.66}{4}$$

$$= \frac{399.59}{4} = 99.90$$

By expressing each quarterly average as percentage the average of average is 99.90.

This method of measuring seasonal variation is considered to be most satisfactory as in practice most widely used. This index does not fluctuate so much as the index based on straight line method. However seasonal index cannot be obtained for each month if 12 month moving average is taken six month in the beginning and six month at the end are left out for which we cannot calculate seasonal indices.

Link relative method

The construction of indices of seasonal variation by link relatives method is also known as Pearson's method, involves the following steps.

- (1) Convert the original data into link relatives by the formula

$$\text{Link relative} = \frac{\text{Current season's figure}}{\text{Previous season's figure}} \times 100$$

- (2) Average these link relatives for each month (or quarter or other time period) while calculating the average mean or median may be used. However median would be a better average as it is not influenced by extreme item.

- (3) Convert the average link relatives (LR) into chain relatives (CR) on the base of the first season by the formula

$$\text{CR of any month} = \frac{\text{LR of that month} \times \text{CR of preceding month}}{100}$$

- (4) Earlier it was assumed the CR of first month or quarter as 100. The new CR of January based on CR of December would not necessarily be 100. The difference between two CRs of January. Therefore needs correction.

- (5) The correction is done by subtracting a correction factor from each chain relative. The correction factor is

$$\text{Cf} = \frac{\text{New chain relative} - \text{Old chain relative or 100}}{12}$$

If figures are given quarterly the correction factor would be

$$\text{Cf} = \frac{\text{CR (New)} - \text{CR (Old)}}{4}$$

The correction factors for February would be $2 \times \text{Cf}$ for march $3 \times \text{Cf}$. Likewise Cf for second quarter would be $2 \times \text{Cf}$ for third quarter $3 \times \text{Cf}$ and fourth quarter $4 \times \text{Cf}$. These are corrected chain relatives.

- (6) Express the corrected chain relatives as percentages of their averages. These provide the required seasonal indices by the method of link relatives.

Illustration: Calculate seasonal indices by the link relative method for the following data:

Year	1st quarter	2nd quarter	3rd quarter	4th quarter
1991	70	64	63	65
1992	67	60	58	63
1993	70	65	65	69
1994	72	61	58	64
1995	62	57	53	58

Solution:**Link Relatives**

Year	1st quarter	2nd quarter	3rd quarter	4th quarter
1991		91.43	98.44	103.17
1992	103.08	85.71	96.67	108.62
1993	111.11	92.86	100.00	106.15
1994	104.35	84.72	95.08	110.34
1995	96.87	91.93	92.98	109.43
Total	415.41	446.65	481.17	537.71
Mean LR	103.85	89.33	96.23	107.54
CR	100.00	$\frac{100 \times 89}{100}$ = 89	$\frac{89 \times 96.23}{100}$ = 85.64	$\frac{85.64 \times 107.57}{100}$ = 92.09
Adjusted CR	100	89(1.09) = 90.09	85.64-(-2.18) = 87.82	92.09(-3.27) = 95.36
Seasonal indices	100×1.071 = 107.2	90.09×1.071 = 96.49	87.82×1.071 = 94.06	95.36×1.071 = 102.13

$$\text{The second CR of first quarter} = \frac{92.09 \times 103.85}{100} = 95.64$$

The difference between the old and new chain relatives of first quarter is $95.64 - 100.00 = -4.36$. Thus difference in one quarter $-4.36/4 = -1.09$.

Subtract (1×1.09) , (2×1.09) and (3×1.09) from 2nd, 3rd and 4th quarter respectively. The total of adjusted CRs is $100 + 90.09 + 87.82 + 95.36 = 373.27$.

Thus seasonal indices will be obtained by multiplying adjusted chain relative by a correction factor $400/373.3 = 1.071$.

Measurement of cyclical variation

Measurement of cyclical fluctuations is a difficult proposition. This is because successive cycles vary so widely in timing amplitude and pattern and because the cyclical rhythm is very closely intertwined with irregular factors. Because of this reason it is impossible to construct meaningful typical index.

- (1) Residual method
- (2) Referenic cycle analysis method
- (3) Direct method
- (4) Harmonic analysis method

Residual method: This is the most common method for estimating the cyclical movement of time series. This method consists of eliminating seasonal variation and trend thus obtaining the cyclical irregular movements. Symbolically

$$\frac{T \times S \times C \times I}{S} = T \times C \times I \text{ and } \frac{T \times C \times I}{T} = C \times I$$

Next the data are usually smoothed in order to obtain the cyclical relatives, since they are expressed in percentage. Thus under residual method the steps involved are -

- (1) Obtain trend value (T) and seasonal indices
- (2) Divide the original data by T to get SCI then divide SCI by S to get CI
- (3) Take the moving total of CI value for monthly series generally 3 monthly moving total is taken. The moving total is generally taken by giving weights of 1,2,1 to the three months. After this the weighted moving total is divided by 4 to get the moving average. If 5 monthly

moving average is taken the weights are 1, 2, 4, 2, 1. These moving averages are the cyclical variation of the series.

Referenic cycle analysis: Under this method the index of variation in each series is calculated with reference to a given referenic year which may be a peak or trough year of the cycle. Obviously from peak year of reference the economic series will show downward falling index till the cycle takes a turn or vice-versa will be true if the reference year is the peak year.

The steps involved in this method are as follows:

1. Select reference data which are the data of peaks and troughs of a business cycle. Here business cycle refers to general business movements.
2. Obtain a cyclical pattern for each series for the period between each two successive troughs. Each period is same for all series so that comparison between the series may be possible.
3. To obtain a cyclical pattern for a series, the data are adjusted for seasonal variation. These are then divided into reference cycle segments for each segment the monthly value are expressed as percentage of average of all values in the segment.
4. Each reference cycle segment is broken into nine stages to correspond to the same 9 stages in the business cycle and the reference cycle relative calculated in stage 3 are averaged for each of the 9 stages.

The nine stages are

1. The 3 months centred on initial trough.
2. The first third of expansion period.
3. The second third of expansion period.
4. The last third of expansion period.
5. The 3 months centred on the peak.
6. The first third of contraction period.
7. The second third of contraction period.

8. The last third of contraction period.

9. The three month centred on the terminal trough.

The nine state averages for each reference cycle segment helps to reduce the erratic movement in a series and thus give a reference cycle pattern for a particular series under consideration.

Direct Method: It is a method based on calculating variation each month or quarter with respect to previous year same month and quarter to see upward changes in case of rising cycle and downwards in case of declining cycle. This roughly results in eliminating seasonal variation and trend.

Harmonic Analysis: If the cyclical variations are of same duration and the amplitude of its various phases is constant a suitable curve may be fitted. Curve is fitted after the irregular movements have been smoothed. The advantage of it is that it will have a prediction value.

Measurement of irregular variation: Irregular movements are those which are left after the cyclical irregular movements have been smoothened. By the very nature of these movements which are erratic no special formula can be suggested to isolate or identify irregular fluctuation. However from a time series if trend, cyclical and seasonal variations are taken out the residual is nothing but the irregular variation.

In multiplicative model $\frac{Y}{TSC}$ or $\frac{TSCI}{TSC} = I$

(where S and C are in fractional form not percentage)

In additive model the irregular variations are

$Y - (T + S + C)$ or $T + S + C + I - (T+S+C)$

However in actual practice only trend and seasonal variations are isolated and the cyclical and irregular fluctuations are kept together because cycles differ in period and amplitude and irregular fluctuations are so mixed up with cyclical fluctuations that it is impracticable to separate them in a meaningful manner.

LESSON - 3

BUSINESS FORECASTING

- ▣ INTRODUCTION
- ▣ DEFINITIONS
- ▣ STEPS IN FORECASTING
- ▣ METHODS OF FORECASTING
- ▣ LIMITATIONS OF FORECASTING

INTRODUCTION

Forecasting is the art of predicting the likelihood of an economic activity or any part thereof for some future period. Forecasting as a subject matter of statistics is based on the use of statistical methods viz. collection of data, sampling and significance tests for estimation and interpretation, correlation regression for model building, time series analysis for extrapolation etc.

In a world where the future is not known with certainty virtually every business and economic decision rests upon the future forecast. Thus forecasting aims at reducing the area of uncertainty. If the future were known with certainty forecasting would be unnecessary. It should be realised at the outset that the object of forecasting is not to determine a curve or series of figures that will tell what exactly will happen in future in advance, but is to make analysis based on definite statistical data which will enable the executive to take advantage of future conditions. Thus business forecasting depends on the analysis of past and present conditions which are indicative of the nature of future condition.

DEFINITION

Some important definitions of the term business forecasting is given below which will help to understand the concept more precisely. "Business forecasting refers to the statistical analysis of the past and current

movements in a given time series, so as to obtain a clue about the future pattern of the movements" – Neter and Wasserman.

"Forecasting refers to the use of knowledge we have at one moment of time to estimate what will happen at another moment of time. The forecasting problem is created by the interval of time between the moments" – Frederik Ekeblad.

"In statistics the term (forecasting) refers to the extending or projecting time series into the future based on past behaviour of the quantitative data".

"Business forecasting is not so much the estimation of certain figures of sales, production, profits, etc. as the analysis of known data, internal and external, in a manner which will enable policy to be determined to meet probable future conditions to the best advantage".

The above definitions show that forecasting can be for a very near future or a distant future and also can pertain to either a single or complex event. It is clear from the above discussions that business forecasting is based on the analysis of internal and external data. The results obtained by such analysis are projected into future to have an idea of the likely course of events. It is also clear from the definitions that forecasting is done on the basis of analysis of qualitative data combined with the knowledge of such qualitative factors which are relevant to the problem under study.

Steps in Business forecasting

The forecasting of business situation involves the following steps:

1. Analysis of past condition

The historical analysis of the problem would reveal the course that the business has followed in the past. For the analysis of past condition we have already discussed the various type of factors like trend, cyclical, seasonal, and irregular factors affecting the time series relating the business. The general trend of the series will give an idea of the direction

in which the series was moving in the past and its future probable course over long period. The cyclical fluctuation would reveal period of boom or a period of depression. It also would reveal the period of the trade cycle. Seasonal variation would indicate the events in the immediate future. Thus observation and analysis of the past behaviour is one of the most vital parts of forecasting. However it should be remembered that though future may be some sort of extension of past, it may not be the exact replica. Hence in making forecasts we should not assume that history repeats itself. However we should believe that there are certain regularities in the past behaviour for reducing the uncertainties of the future.

2. Analysis of present condition

The analysis of present condition would reveal those factors which are influencing the phenomena under study on a particular direction. The analysis of present conditions is done with a view to have an idea about the future course of events.

3. Selecting and compiling data to be used as measuring devices

Very often it happens that analysis of a problem is done by using inappropriate data for the purpose leading to undependable inferences and unreliable forecasts. Great care has to be taken to see that the data which are analysed are appropriate for the purpose of forecasting.

4. Analysing the data

In the last step, the data are analysed in the light of one's understanding of the reasons why changes occur. For example if it is reasoned that a certain combination of forces will result in a given change, the statistical part of the problem is to measure these forces, from the data available and to draw conclusion on the future course of action.

METHODS OF FORECASTING

There are various methods of making business forecasts and no method can be suggested for universal applicability. While forecasting care should be taken to select the correct technique for a particular situation.

Also before applying a method of forecasting the following questions should be answered.

1. What is the purpose of the forecast – how is it to be used.
2. What are the dynamics and components of the system for which the forecast will be made?
3. How important is the past in estimating the future?

The important methods of business forecasting are:

1. Business Barometers
2. Time series analysis
3. Extrapolation
4. Regression analysis
5. Econometric models
6. Opinion polls
7. Causal models
8. Exponential smoothing
9. Survey method

1. Business Barometers: These consist of a set of statistical time series which when used in conjugation with one another and combined with one or more, provide indications as to the direction in which the economy is heading. The index numbers relating to business conditions are called business barometers. These facilitate various forms of business forecasting. The following are some of the important series which aid businessmen in forecasting:

Gross national product, employment, wholesale price index, consumer price index, industrial production, volume of bank deposits, dispensable personal income, stock prices, bond yields.

Index numbers relating to different activities in the field of production, trade, finance etc. may also be combined into a general index of business activity. With the help of index numbers, it becomes comparatively easy to forecast the future course of events. The indicators are grouped into three

broad categories – lead indicators, coincident indicators, lag indicators. The indicators signal in advance a change in the economic activity as a whole and are called lead indicators like call money rates, new capital issue business inventory, new company formation, pending order for consumer goods etc. Most of these series indicate the future course of general economic activity with a lead of 2 to 10 months. The coincident indicators move approximately with the economic activity and constitute as a measure of current economic activity – Indicators like unemployment rate, personal income, index of industrial production, wholesale price index, business profits, level of inventory etc. The coincident indicators amongst the other things help in confirming the signals given by the lead indicators. The lag indicators comprise those time series which trail behind the economic activity.

2. Time series analysis: As has already been discussed, a historical series can be decomposed into various components like trend, seasonal variation, cyclical variations and random variations. By doing so it is possible to study the general trend of the series and to have an idea about the effect of seasonal factors. Eventhough the effects of the cyclical and random factors are difficult to be isolated, yet the trend values and seasonal variations help a lot in understanding the behaviour of a time series. The analysis of time series serves two purposes.

1. It provides the initial approximation forecast that takes into account those empirical regularities which may with reasonable assurance be expected to persist.
2. After the trend and seasonal effects have been identified the original data may be adjusted.

It should however be remembered that time series analysis should be used as a basis of forecasting only when the data are available for a long period of time and the tendency disclosed by the trend and seasonal factors are fairly clear and stable.

3. Extrapolation: Extrapolation is the most widely used as well as the simplest of the forecasting methods. As extrapolation relies on relative

consistency in the pattern of past movements of the time series it gives the most reasonable forecast. Though this device does not isolate the effects of various factors influencing the time series it takes into account the totality of the factors influencing the problem.

4. Regression analysis: The regression approach contributes a lot to the solution of the forecasting problem. Regression studies are meant to disclose the relative movements of two or more interrelated series. The changes in one variable as a result of specified change in the other variable are estimated by the help of regression analysis. Eventhough a number of factors affect the business phenomena, regression analysis provides a handy tool for business forecasting. With the help of regression studies it is possible to study a problem both at macro level as well as at micro level. It should be noted that all the factors affecting the business situation may not be able to be accounted for in the multiple regressions and consequently the forecast made by regression analysis cannot be infalliable. The results of regression analysis should be suitably modified in the light of personal experience and the influence of qualitative factors do not enter a regression analysis.

5. Econometric models: The term econometrics refers to application of mathematical economic theory and statistical procedures to economic data in order to verify economic theorems and establish quantitative results in economics. The econometric models take the form of a set of simultaneous equations. The values of constants in such equations are supplied by a study of statistical time series, and a large number of equations may be necessary to produce an adequate model. The advent of computers have made the use of this method relatively easy. However econometric methods need a large quantum of data and in the absence of adequate data the econometric model formed would not give a dependable forecast. At the same time econometric models are expensive, technical and complicated.

6. Opinion polls: Opinion poll is the survey of opinion of knowledgeable persons or experts in the field whose views carry a lot of weight. For

example opinion survey of sales representatives and marketing experts may be helpful in formulating demand forecast.

7. Causal model: A causal method is the most sophisticated kind of forecasting method. It expresses mathematically the causal relationship. The causal model takes into account everything known of the dynamics of the flow system and utilises predictions of related events. These models need constant revision as more and more knowledge about the system becomes available. This model though sophisticated are beyond the reach of the individuals and industries, as like econometric models they are highly technical, complicated and expensive.

8. Exponential smoothing: This method is an outgrowth of the recent attempts to maintain the smoothing function of moving averages without their corresponding drawbacks and limitations. Exponential smoothing is a special kind of weighted average and is found extremely useful in short term forecasting of inventories and sales.

9. Survey method: The survey method is widely used as a tool of forecasting for the existing and new products. Forecasts are made on the basis of field surveys which give the necessary information both quantitative and qualitative. However the important limitation of survey method is that the chances of getting biased information are high.

LIMITATIONS OF FORECASTING

It should be kept in mind that business forecasting is not a sure road to success. The assumptions under which business forecasts are made may not always be satisfied and when it is so the forecasts are likely to be misleading. It also should not be forgotten that though history repeats itself it does not repeat itself so precisely; rather some new factor also may come up. Business forecasting only discloses what is likely to happen in future under certain conditions and nothing else.

UNIT - VIII

PROBABILITY

- INTRODUCTION
- TERMINOLOGY
- CONCEPT OF PROBABILITY
- IMPORTANCE OF PROBABILITY
- TYPES OF PROBABILITY
- LAWS OF PROBABILITY

INTRODUCTION

The word 'probability' or 'chance' is very commonly used in day-to-day conversation. The terms like possible, probable, likely, perhaps, seems etc. convey the same sense i.e., it speaks of some uncertainty in the happening of a particular event. Thus ordinarily we can interpret that probability connotes uncertainty about happening of an event. Thus in common parlance the term probability refers to the chance of happening or not happening of an event. The moment we use the word chance we indicate an element of uncertainty. However instead of talking subjectively when the chances or uncertainty is tried to be explained numerically we can make sensible numerical statements on uncertainty. Thus if we could assign numerical value to the statements of chances of happening it becomes more precise. Thus the theory of probability provides a numerical measure of the element of uncertainty. It enables us to take decision under conditions of uncertainty with a calculated risk.

The theory of probability was originated from the game of chances such as throwing a dice, tossing a coin, drawing a card from a pack of cards. Jeram Cardan (1501-76) was the first man to use the numerical expression of probability in his book "Book on Games of Chances". However it was the French mathematician Blaise Pascal (1623-62) and Pierre de Fermat (1601-65) who laid the solid stone of foundation of the mathematical

theory of probability. The contribution of James Bernoulli to the theory of mathematical probability is immense. The other personalities who have made notable contribution to the subject of probability are Thomas Bayes, De Moivre, Pierre-Simon de Laplace. The contribution made by Russian mathematician to the modern theory of probability is very significant. The modern theory of probability was developed by eminent Russian mathematicians like Chebychev, Markov, Kolmogorov.

Starting with the game of chance, probability today has become one of the basic tools of statistics. In fact statistics and probability are so fundamentally interrelated that it is difficult to discuss statistics without an understanding of the meaning of probability. The statistical results are well interpreted with the knowledge of probability. As many statistical procedures involve conclusions based on samples which are subject to random variation, the probability theory helps us to express numerically the uncertainties in the resulting conclusion.

Today the concept of probability has assumed great importance and the mathematical theory of probability has become the basis for statistical application in both social and decision making research.

TERMINOLOGY

Before we discuss the term probability from different view points we need to familiarise ourselves with certain terms that are used in this context.

1. Random experiment - It is an experiment which if conducted repeatedly under homogeneous condition does not give the same result. The result may be any one of the possible outcomes. For example a coin is tossed and we get head it does not mean if we toss it again we get a head. That is the outcome will be any one of the possible outcomes (head or tail)
2. Trial and Event - The performance of a random experiment is called as a trial and the outcome is an event; for example if a coin is tossed repeatedly the result is not unique. We may get any of the two faces

head or tail. Thus tossing a coin is a random experiment and getting a head or tail is an event.

Event is called simple if it corresponds to a single possible outcome of the experiment or trial, otherwise it is compound or composite event. Thus in tossing a single die the event of getting 5 is a simple event but the event of getting an even number is a composite event.

Exhaustive cases: All possible outcomes of an event are called exhaustive cases for the experiment. Thus in a toss of a single coin, we get head (H) or tail (T). Hence the exhaustive number of cases are two. In a throw of single dice the exhaustive cases are 6 as the dice has six faces each marked with different numbers. Similarly if 2 dice are thrown the exhaustive cases would be 36 (6×6).

Favourable cases: The number of outcomes of a random experiment which result in happening of a desired event are called favourable cases. For example in drawing a card from a pack of cards the cases favourable to getting a spade is 13 and to getting an ace of diamond is only 1.

Mutually exclusive cases: Two or more events are said to be mutually exclusive if the happening of any one of them prevents the happening of all other events in a single experiment. In this case one and only one event can happen preventing all others to happen. The events are said to be mutually exclusive if no two or more of them can happen simultaneously; for example in the toss of a coin the events head or tail are mutually exclusive, as, if head comes tail cannot come up or if tail comes up, head cannot come.

Equally likely cases: The outcomes are said to be equally likely or equally probable if the chances of their happening are equal or there is no preference of any event over the other. Thus in tossing of a coin all the outcomes (H,T) and in rolling a dice all outcomes (1, 2, 3, 4, 5, 6) are equally likely if the coin or the dice are unbiased.

Independent events: An event is said to be independent if its happening is not affected by and does not affect the happening of any one of the others. For example in the toss of a coin repeatedly getting head on first throw is independent of getting head in second, third or subsequent throws. However in drawing cards from a pack of cards the result of second draw will depend upon the card drawn in the first draw if the card is not replaced. If the card drawn in the first draw is replaced before the second draw of a card is done then the result of first draw will be independent of the first draw.

Similarly drawing of balls from an urn gives independent events if the draws are made with replacement. If the ball drawn in the earlier draw is not replaced, the resulting draw will not be independent.

Permutation and Combinations: The word permutation refers to arrangements and the word combination refers to groups.

- (i) The permutation of n dissimilar things taken all at a time is n or $n!$

Thus if there are 4 letters A, B, C and D the total numbers of ways in which they can be arranged is ABCD, ABDC, ACBD, ADBC, ADCB, BACD, BADC, BCDA, BCAD, BDCA, BDAC, CDBA, CDAB, CBAD, DCBA, CBDA, CABD, CADB, DCAB, DBAC, DABC, DBCA, DACB i.e., 24 different ways which can be calculated as $4!$ (pronounced as four factorial) which is nothing but multiplication of all number upto four i.e. $1 \times 2 \times 3 \times 4 = 24$. Similarly if we have 3 letter ABC the total number of ways in which they can be arranged is ABC, ACB, BAC, BCA, CAB, CBA or it is equal to $3!$ or $3 \times 2 \times 1 = 6$.

- (ii) The permutation of n dissimilar things taken r at a time is n_{pr} (pronounced as n permutation r). Thus if we have to arrange two letters out of four letters ABCD we can arrange them as AB, BA, AC, CA, AD, DA, DC, CD, BC, CB, BD, DB in 12 different ways. If two letters are arranged at a time out of 3 letters ABC the arrangements can be done as AB, BA, AC, CA, BC, CB or in 6 different ways. The number of ways the arrangements can be done can easily be calculated by the formula $n!/(n-r)!$ where n is the total number of

items and 'r' is the number of items taken at a time for arrangement.

Thus the arrangement of 2 out of 4 will be $4!/(4-2)!$. $4 \times 3 \times 2 \times 1 / 2 \times 1 = 24/2 = 12$ and 2 out 3 will be $3!/(3-2)! = 6/1 = 6$ different ways respectively.

- (iii) The number of permutations of n things where n of them are of one kind and are of another kind the arrangements will be in $n!/(n_1!n_2!)$ different ways. Thus if we have to find out the permutations of the letters of the word FARIDABAD (where A occurs 3 times and D occurs 2 times) the answer would be

$$\frac{9!}{3!2!} = \frac{9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1 \times 2 \times 1} = 32,240$$

- (iv) The fundamental rule of counting is that if an operation can be performed in 'm' ways and having been performed in any one of this ways a second operation can be performed in 'n' ways. The total number of ways of performing the two operations together is $m \times n$.

Thus if a journey between Pondicherry and Madras can be performed in 3 ways and return journey can be performed in 3 ways, the total number of performing the journey is $3 \times 3 = 9$. However, if the return journey is not to be performed by the same train by which one went to Madras, then the number of ways of performing the return journey is only 2 and the total number of ways of performing both the journey would be $3 \times 2 = 6$.

- (v) The number of combination of 'n' different things taken 'r' at a time is n_{cr} (pronounced as n combination r) or $n!/(r!(n-r)!)$. Thus if we have to pick up two alphabets out of three A, B, C we can pick AB or AC or BC or 3_{c2} or $3!/(2!(3-2)!)$ or $3 \times 2 \times 1 / 2 \times 1 \times 1 = 3$. We had seen that number of permutations in this case was 6 because each combination can be arranged in two ways like (AB, BA), (AC, CA), (BC, CB). Thus number of permutation is equal to number of combination multiplied by r. In other words

$$n_{pr} = n_{cr} \times r \text{ or } n_{cr} = n_{pr}/r.$$

CONCEPT OF PROBABILITY

The probability of a given event is an expression of likelihood or chance of occurrence of an event. A probability is a number which ranges between zero to one (zero for an event which cannot occur and one for an event certain to occur). The number assigned for the probability would depend upon the interpretation of the term probability. The general rule of the happening of an event is that if the event can happen in 'm' ways and fail to happen in 'n' ways, the probability of the happening of the event is $P = m/m+n$ i.e., number of cases favourable to the event divided by number of exhaustive cases. How the number of cases favourable to the event and exhaustive cases should be computed is a matter on which opinion differs. However, broadly speaking there are four different schools of thought on the concept of probability.

(i) Classical or a priori probability: The classical approach to probability is the oldest and simplest. It was pronounced by Laplace and he had defined probability as the ratio of 'favourable' cases to the total number of equally likely cases. The term 'equally likely' conveys the notion that each outcome of an experiment has the same chances of appearing as any other. Thus as per this definition if the probability of occurrence of an event is $P(A)$ then we have

$$P(A) = \frac{\text{Number of favourable cases}}{\text{Total number of equally likely cases}}$$

Probability of non occurrence of an event $P(A')$ is

$$\begin{aligned} P(A') &= \frac{\text{Number of unfavourable cases}}{\text{Total number of equally likely cases}} \\ &= 1 - P(A) \end{aligned}$$

Classical probability is often called a priori probability because if we keep using orderly examples of unbiased dice, fair coin, etc., we can state the answer in advance (a priori) without rolling a dice or tossing a coin etc.

This theory of probability is suitable for calculating the probabilities of various events in game of chance where various events are equally likely to happen. Thus while tossing a dice all the six sides have equal likelihood of coming up. Thus probability of getting 5 on a single throw of a dice would be $1/6$ and probability of not getting 5 in a single throw is $5/6$ or

$$P = 1/6 \text{ and } Q = 5/6.$$

It means that $P + Q = 1$ or $1 - Q = P$ or $1 - P = Q$.

However this definition of probability suffers from certain weaknesses like

1. This definition cannot be applied if one is not able to find out list of the cases which can be considered equally likely. It may also happen we may be able to enumerate the different outcomes but they may not be equally likely. For example if we take the rain, how does it apply to probability of rain? The possible cases are rain or no rain. But at any given time it will not usually be agreed that they are equally likely.
2. The classical approach may be impossible to apply in various real life situations. If one wants to know the probability of a male living after the age of 60 or the probability of a bulb burning less than 2000 hours, classical probability fails to be apply here. Thus in real life situations are disorderly and they often make it difficult and at times impossible to apply classical probability.

(ii) Relative frequency probability: Classical probability fails to express the probability as soon as we deviate from the field of coin, dice, cards and other simple games of chances. Also the classical approach does not explain actual results in certain cases. For example if a coin is tossed 10 times we may get 7 heads and 3 tails. The probability would have been 0.7 and 0.3 respectively but according to priori probability it would be 0.5 and 0.5.

In such situations the probability of the happening of an event is determined on the basis of past experience or on the basis of relative

success in the past. Thus if 70% of the male live more than 60 years age the relative frequency of a male living more than 60 years would be 7. However it should be noted that the probability should be estimated on the basis of a large number of readings of the past. The larger the number of readings, the greater would be the accuracy of the result. This approach to probability is called posterior probability as it is based on past experiences. It should be understood that priori probability is applicable to games of chances and posterior probability is applicable mostly in real life situations and various economic and social phenomena. Posterior probabilities are based on experience of the past. They are empirical in nature. The posterior probability or empirical probability of an event is thus

$$P = \frac{\text{Relative frequency}}{\text{Number of trials}}$$

Thus out of 10000 items produced by a machine in past if 500 are found to be defective the probability of defective article to be produced would be $500/10000 = 0.05$. The relative frequency approach though useful in practice has difficulties from a mathematical point of view. Quite often this approach is used without evaluating a sufficient number of outcomes.

(iii) Subjective Probability: This probability as the name says is subjective in its approach. Thus it is defined as the probability assigned to an event by an individual based on whatever evidence is available. Hence such probabilities are based on the belief or intuition of the person. This theory is also known as personalistic theory of probability as it presumes that any decision reflects the personality of the decision maker and subjective elements are important in assigning a probability to an event. This is very true in actual life. Suppose we have data relating to price of a share for the last two years and further supposing that out of 1000 quotations relating to share prices there was a price rise on 400 occasions, then the empirical probability of a price rise in this share is $400/1000$ or 0.4 . However with these data some people would buy the shares and others would sell them for reasons that their subjective estimate coupled with

relative frequency approach may give them different ideas about the price of these shares in future. Here the personality of the decision maker is reflected in the ultimate decision. The decisions under this theory is taken on the basis of the available data plus the effects of other factors, many of which are subjective in nature. The personalistic approach to probability is very broad and highly flexible in nature. It permits probability assignment to events for which there may be no objective data. However it needs much care and consistency; otherwise the decision made may be misleading. If used with care this concept is extremely useful in the context of business decision making.

(iv) Axiomatic approach to probability: This approach was first suggested by Kolmogorov. This is entirely mathematical in its approach and is based on set theory. When this approach is followed, no precise definition of probability is given, rather we give certain axioms or postulates on which probability calculations are based. To start with some concepts are laid down and certain properties or postulates commonly known as axioms are defined. From these axioms alone the entire theory of probability is derived by deductive logic. The axiomatic approach includes the concept of both classical as well as empirical definition of probability. This approach to probability is based on the following axioms:

1. The probability of an event ranges from 0 to 1
2. The probability of the entire sample space is 1, and
3. If for mutually exclusive events the probability of happening of either A or B denoted by $P(A \cup B)$ read as A union B shall be,

$$[P(A \cup B) = P(A) + P(B)]$$

For events of simultaneous occurring probability of both A and B denoted by $P(A \cap B)$ read as A intersection B shall be $P(A \cap B) = P(A)P(B)$. It may be noted that out of four interpretations of the concept of probability each has got its own merits and demerits and one may use whichever approach is convenient and appropriate for the problem under consideration.

IMPORTANCE OF PROBABILITY

Though probability theory started with the game of chances, i.e., from gambling, it has assumed a great importance in almost all fields of study. It has been employed to treat many critical and weighty problems. It is the foundation of the classical decision procedures of estimation and testing. In fact it has become an indispensable tool for all types of formal studies that involve uncertainty. It could be noted that the concept of probability is employed not only for various types of scientific investigations but also for many problems in every day life. It has become a part of our every day lives. Highlighting the importance of probability theory Ya Lun Chou has beautifully pointed out that, "statistics as a method of decision making under uncertainty is founded on probability theory", since probability is at once the language and measure of uncertainty and the risks associated with it. Before learning statistical decision procedure the reader must acquire an understanding of probability theory.

TYPES OF PROBABILITIES

Basically there are three types of probabilities:

1. Marginal
2. Joint
3. Conditional

Marginal Probability: Marginal probabilities are also otherwise called as conditional probabilities. Marginal probabilities are the probabilities of single event which is expressed symbolically as $P(A)$, the probability of the event of happening. Thus it is the simple probability of occurrence of an event.

Joint Probability: The probability of two or more independent events occurring together or in succession is the product of their marginal probabilities. Symbolically joint probability is denoted as $P(AB)$ and mathematically the joint probability is stated

$$P(AB) = P(A) \times P(B)$$

where $P(AB)$ – probability of events A and B occurring together or in succession, known as joint probability

$P(A)$ – marginal probability of event A occurring

$P(B)$ – marginal probability of event B occurring.

Conditional Probability: Conditional probability is the probability that a second event (B) will occur if a first event (A) has already occurred. If the events are independent i.e., happening of A does not affect happening of B the conditional probability denoted by $P(A/B)$ will be $P(A)$.

$$P(A/B) = P(A)$$

When the events are dependent

$$P(A/B) = P(AB) / P(B)$$

LAWS OF PROBABILITY

There are two important laws or theorems of probability. They are

1. Additions Theorem
2. Multiplication Theorem

(1) Additions Theorem

The probability that an event will occur in one of the second possible ways is calculated as the sum of probabilities of occurrence of several different possible ways. It is assumed that all these events are mutually exclusive and equally likely. Thus this theorem states that if two events A and B are mutually exclusive (so that if one happens the other cannot happen) the probability that any one of them would happen is the sum of the probabilities of the happening of A and B. Symbolically

$$P(A \text{ or } B) = P(A) + P(B)$$

The general rule of addition can be stated as under: "If an event can happen in different ways which are mutually exclusive the probability that it will happen is the sum of the probabilities of its happening in these different ways".

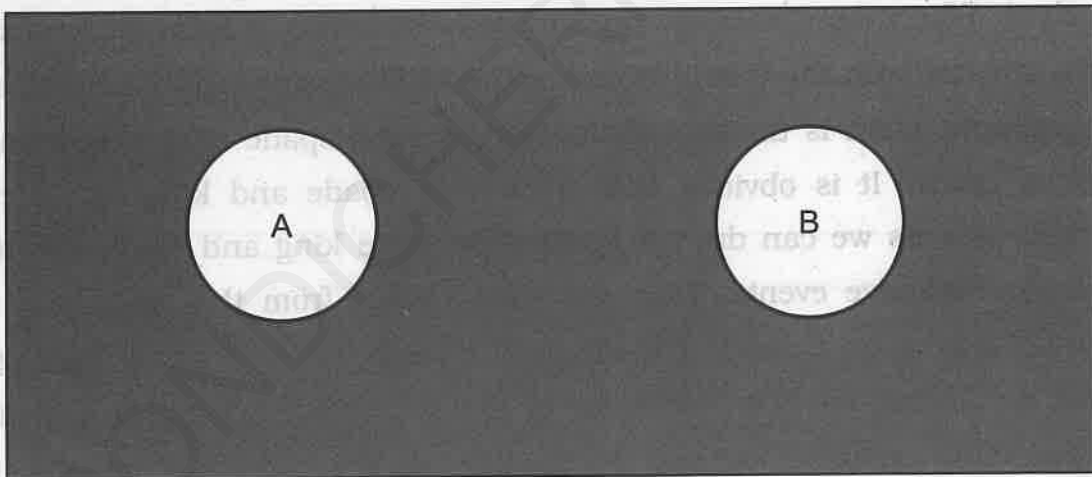
Thus if the probabilities of r mutually exclusive events are $P_1, P_2, P_3, \dots, P_n$ then the probabilities that someone of the event would happen would be $P_1 + P_2 + P_3 + \dots + P_n$.

Proof of the theorem: If an event A can happen in a_1 ways and B in a_2 ways then the number of ways in which either events can happen is $a_1 + a_2$. If the total number of probabilities is n , then the probability of either the first or second event happening is

$$\frac{a_1 + a_2}{n} = \frac{a_1}{n} + \frac{a_2}{n} \text{ where } a_1/n = P(A) \text{ and } a_2/n = P(B)$$

Hence $P(A \text{ or } B) = P(A) + P(B)$

This will be illustrated through the following diagram.



Mutually Exclusive Events

FIG. 8.1

In the above diagram the entire sample space is enclosed between the rectangle which means it is the total number of ways in which the event can take place. The two circles indicate the number of ways favourable for A and B events respectively. The probability of either A or B would happen would be the sum of their probabilities of happening.

Illustration: Suppose from a pack of 52 cards one card is drawn at random, what is the probability that it is either a king or a queen?

Solution: Since the events are mutually exclusive (i.e., if the card drawn is a king it cannot be a queen) the probability of drawing a king is $4/52$ and similarly probability of a queen is $4/52$ (as there are 4 kings and 4 queens in the pack) the probability that the card is either a king or queen is.

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(B) \\ &= 4/52 + 4/52 = 2/13 \end{aligned}$$

when the events are not mutually exclusive.

The above discussed addition theorem will not be applicable when the events are not mutually exclusive (when two or more events can take place together). When the events are not mutually exclusive or in other words it is possible for both the events to occur, the addition rule must be modified. For example, what is the probability of drawing a spade or a king from a pack of cards? It is obvious that the events spade and king can occur together also as we can draw a spade king since king and spade are not mutually exclusive events. Thus we must reduce from the probability of drawing a king or a spade the chances of both of them being together. Hence for finding out the probability of one or more of two events that are not mutually exclusive, we use the modified form of addition theorem.

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

$P(A \text{ or } B)$ = probability of A or B happening when A or B are not mutually exclusive.

$P(A)$ = probability of A happening

$P(B)$ = probability of B happening

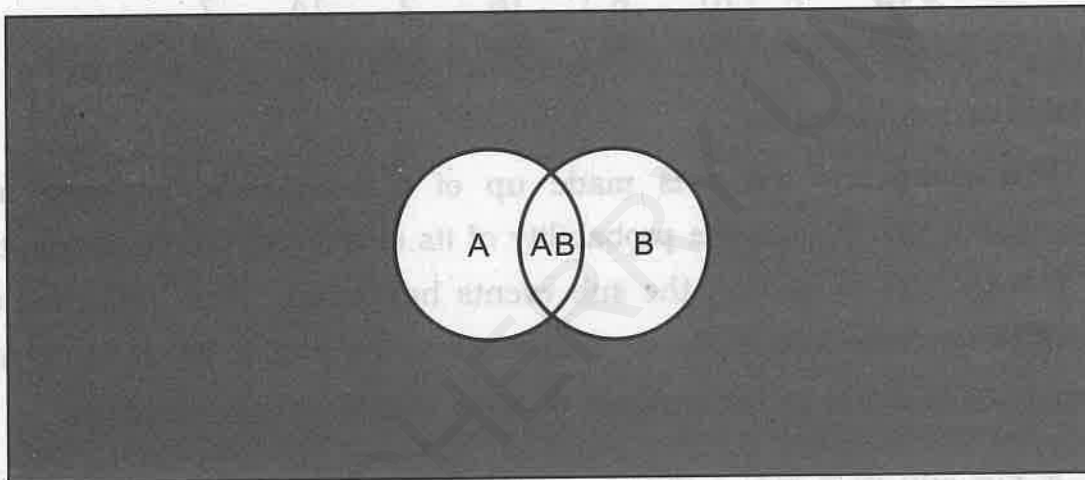
$P(AB)$ = probability of A and B happening together

In the example taken, the probability of drawing a spade or a king shall be

$$P(\text{spade or king}) = P(\text{spade}) + P(\text{king}) - P(\text{spade and king})$$

$$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}$$

The following diagram also explains how the probability works when events are not mutually exclusive or are overlapping.



Mutually Exclusive Events

FIG. 8.2

In the above diagram events A and B overlap each other. In such a case if we find out the probability of either A or B happening and if we add the probability of happening of (A) and (B) we will be counting (AB) twice. Therefore the formula of addition has to be modified as

$$P(A \text{ or } B) = P(A) + P(B) - P(AB)$$

Illustration: One ticket is drawn at random from a bag containing 30 tickets numbered 1 to 30. Find the probability that it is a multiple of 3 or 5.

Solution: One ticket can be drawn in 30_{C_1} ways or 30 ways. This is the total number of ways a ticket can be drawn. The probability of having a multiple of 3 or 5 would be

Multiples of 3 are 3 6 9 12 15 18 21 24 27 30

Multiples of 5 are 5 10 15 20 25 30

Thus there are 10 multiples of 3 and 6 multiples of 5. However two digits 15 and 30 are common to both the sets. Therefore the events are not mutually exclusive. The required probability would be

$$\frac{10}{30} + \frac{6}{30} - \left(\frac{10}{30} \times \frac{6}{30} \right) = \frac{16}{30} - \frac{2}{30} = \frac{14}{30} = \frac{7}{15}$$

Multiplication Theorem

If a compound event is made up of a number of separate and independent sub events, the probability of its occurrence is the product of the probabilities of each of the sub events happening. Thus this theorem says if two events (A) and (B) are independent, the probability of their joint occurrence is equal to the product of their individual probabilities.

$$\text{ie } P(A \text{ and } B) = P(A) \times P(B)$$

The theorem can be extended to three or more events

$$P(A, B, C) = P(A) \times P(B) \times P(C)$$

Proof of the theorem: If an event A can happen in n_1 ways of which a_1 are successful and event B can happen in n_2 ways of which a_2 are successful we can find each successful event of $2n$ first with each successful event in the second case. Thus the total number of successful happening in both cases $a_1 \times a_2$ similarly total number of possible cases is $n_1 \times n_2$

Then by definition the probability of the occurrence of both events are

$$\frac{a_1 \times a_2}{n_1 \times n_2} = \frac{a_1}{n_1} \times \frac{a_2}{n_2} \quad \text{But } a_1/n_1 = P(A) \text{ and } a_2/n_2 = P(B)$$

Thus $P(A \text{ and } B) = P(A) \times P(B)$

Illustration: A bag contains 4 white and 3 black balls. A ball is drawn out of it and replaced in the bag. Then a ball is drawn again. What is the probability that (i) both the balls drawn were black (ii) both were white (iii) the first ball is black and second white (iv) the first ball was white and second black?

Solution: The events are independent and capable of simultaneous occurrence. The rule of multiplication would be applied.

The probability that

$$(i) \text{ both are black} = \frac{3}{7} \times \frac{3}{7} = \frac{9}{49}$$

$$(ii) \text{ both are white} = \frac{4}{7} \times \frac{4}{7} = \frac{16}{49}$$

(iii) first is black and second is white

$$= \frac{3}{7} \times \frac{4}{7} = \frac{12}{49}$$

(iv) the first one is white and second black

$$= \frac{4}{7} \times \frac{3}{7} = \frac{12}{49}$$

Illustration: The students A, B, C are given a problem in statistics. The probability of their solving the problem are $\frac{3}{4}$, $\frac{1}{4}$ and $\frac{2}{4}$ respectively. What is the probability that if all of them try the problem would be solved?

Solution: 1st Method

Probability that A solves and B and C fail

$$\frac{3}{4} \times \frac{3}{4} \times \frac{2}{4} = \frac{18}{64}$$

Probability that B solves and C fails

$$\frac{1}{4} \times \frac{1}{4} \times \frac{2}{4} = \frac{2}{64}$$

Probability that C solves A and B fail

$$\frac{1}{4} \times \frac{3}{4} \times \frac{2}{4} = \frac{6}{64}$$

Probability that A and B solve and C fails

$$\frac{3}{4} \times \frac{1}{4} \times \frac{2}{4} = \frac{6}{64}$$

Probability that A and C solve B fails

$$\frac{3}{4} \times \frac{3}{4} \times \frac{2}{4} = \frac{18}{64}$$

Probability that B and C solve and A fails

$$\frac{1}{4} \times \frac{1}{4} \times \frac{2}{4} = \frac{2}{64}$$

Probability that A,B,C all solve

$$\frac{3}{4} \times \frac{1}{4} \times \frac{2}{4} = \frac{6}{64}$$

The probability that any one of the above situations happen as:

$$\frac{18}{64} + \frac{2}{64} + \frac{6}{64} + \frac{6}{64} + \frac{18}{64} + \frac{2}{64} + \frac{6}{64} = \frac{58}{64}$$

Second Method

The probabilities that A,B,C would not be able to solve the problem are $\left(1 - \frac{3}{4}\right)$, $\left(1 - \frac{1}{4}\right)$ and $\left(1 - \frac{2}{4}\right)$ respectively. "Thus the values are

$\frac{1}{4}$, $\frac{3}{4}$ and $\frac{2}{4}$ has the probability of no one can solve the problem is

$$q_1 \times q_2 \times q_3 = \frac{1}{4} \times \frac{3}{4} \times \frac{2}{4} = \frac{6}{64}$$

Therefore probability of the problem being solved is $1 - \frac{6}{64} = \frac{58}{64}$

The probability of the happening of an event in one trial being known the probability of its happening r times in n trials is

$$n_{cr} p^r q^{n-r}$$

This is a very convenient rule as would be illustrated by the following example.

Illustration: What is the probability of getting exactly 2 heads in a single throw of three coins?

Solution: Three coins when thrown at a time may fall in any one of the following manners

H H H	- three heads	H T T	} 1 head two tails
H H T	} 2 head one tail	T H T	
H T H		T T H	
H T T		T T T	

Since we have to find out the probability of exactly 2 heads (which means 2 heads and 1 tail) the probability would be $3/8$ as there are 3 cases where there are only two heads. The total number of outcomes being 8.

This problem can be solved by the formula $P = n_{cr} p^r q^{n-r}$ easily.

Where we want the event to happen r times out of n :

In this problem probability of a head in a single throw of a coin or $P = 1/2$ and not getting a head or $q = 1/2$. Thus when we want 2 heads out of the possible 3 the formula would be

$${}^3C_2 p^2 q^{3-2} \text{ or } {}^3C_2 p^2 q^1$$

$$\text{or } \frac{3!}{2! (3-2)!} \times \left(\frac{1}{2}\right)^2 \times \left(\frac{1}{2}\right) = 3 \times \frac{1}{4} \times \frac{1}{2} = \frac{3}{8}$$

Conditional Probability

In the multiplication theorem explained above we assumed that there is a simultaneous occurrence of two independent events. Thus if the events are not independent the above theorem would not be applicable. Two events A and B are said to be dependent when B can occur only when A is known to have occurred and vice versa. The probability attached to such an event is called conditional probability and is denoted by $P(A/B)$

$P(A/B)$ would mean probability of A given that B has occurred.

$P(B/A)$ would mean probability of B given that A has occurred.

If two events A and B are dependent then the conditional probability of B given A is

$$P(B/A) = \frac{P(AB)}{P(A)} \text{ and } P(A/B) = \frac{P(AB)}{P(B)}$$

The general rule of multiplication in its identified form for conditional probability would be

$$P(AB) = P(B) \times P(A/B) \text{ or } P(A) P(B/A)$$

Proof: Suppose a_1 is the number of cases for the simultaneous happening of (A and B) out of $a_1 + a_2$ cases in which A can happen with or without B.

$$P(B/A) = \frac{a_1}{a_2 + a_1} = \frac{a_1/n}{a_2 + a_1/n} = \frac{P(AB)}{P(A)}$$

Illustration: A bag contains 5 red and 3 white balls. Two draws are made without replacement. What is the probability that both the balls are red?

Solution: Probability of drawing a red ball in the first draw is

$$P(A) = 5/8$$

Probability of drawing a red ball on the second draw given that first ball is red of $P(B/A) = 4/7$ (since only 7 balls are left and only 4 of them are red).

The combined probability of the two events is

$$\begin{aligned} P(AB) &= P(A) \times P(B/A) \\ &= 5/8 \times 4/7 = 20/56 \end{aligned}$$

Illustration: A bag contains 4 red and 6 green balls. Two draws of one ball each are made without replacement. What is the probability that one is red and the other is green?

Solution: Probability of drawing a red ball $P(A) = 4/10$

Probability of drawing a green ball in the second draw given that the first draw has given a red ball $P(B/A) = 6/9$ (since only 9 balls are left out of which 6 are green)

Probability of the combined event

$$P(AB) = P(A) \times P(B/A) = 4/10 \times 6/9 = 24/90$$

But it could also happen the first ball may be green and second ball is red.

Probability of drawing a green first $P(B) = 6/10$ and red next (given green has been drawn) $P(A/B) = 4/9$.

$$P(AB) = P(B) \times P(A/B) = 6/10 \times 4/9 = 24/90.$$

Now when any one of the two situations (red and green or green and red) can happen and both of them are mutually exclusive the required probability will be

$$\frac{24}{90} + \frac{24}{90} = \frac{48}{90} = \frac{8}{15}$$

Bayes' Theorem

One of the most interesting applications of the results of probability theory involves estimating unknown probabilities and making decisions on the basis of new information.

The concept of conditional probability discussed above takes into account information about the occurrence of one event to predict the probability of another event. This concept can be extended to revise probabilities based on new information and to determine the probability that a particular effect was due to a specific cause. The Bayes' theorem is based on the proposition that probabilities should be revised when new information is available. The need to revise probabilities arises from a need to make better use of available information. The probabilities before the revision are called 'prior probabilities' and those after revision 'posterior probability'.

Imagine a situation where two uncertain events (A) and (not A) are possible. Suppose we know the probability of A's happening. Not A's happening is $1 - P(A)$'s happening. These probabilities are prior probabilities because they are determined before the sample information is taken into account. Suppose the investigation is conducted and some further information is available. In the light of sample observation probability of A undergoes revision. This is called posterior probability or revised probability as they are obtained by revising the prior probability in the light of the additional information gained from the sample observation (via Bayes' rule). Thus prior probability which is unconditional probability becomes a posterior probability which is a conditional probability by using Bayes' rule.

Bayes' Rule provides a powerful method in improving the quantity of probability for aiding the management in decision making under uncertainty for a problem with two events A and B. Bayes' theorem is stated as

$$P(A/B)$$

$$= \frac{P(AB)}{P(B)}$$

$$= \frac{P(A) P(B/A)}{P(B)}$$

$$= \frac{P(A) P(B/A)}{P(A) P(B/A) + P(\text{Not } A) P(B/\text{Not } A)}$$

In a similar way Bayes' theorem can be extended to n events also. Application of Bayes' theorem is a powerful method of evaluating new information in order to revise prior probabilities. When correctly used Bayes' theorem can be of tremendous aid in decision making.

The application of Bayes' theorem is being illustrated in the following examples.

Illustration: Suppose probability of A's winning a prize in a lottery is $1/5$. Suppose C who speaks 3 times out of 5 informs A that he has won a prize. What is the probability that A has actually won a prize after his information.

Solution: Let A stand for winning the prize and B for truth in his information.

The prior probability of A = $1/5$

The prior probability of Not A = $4/5$

Now probability of $P(B/A)$ i.e., probability A is speaking truth when the probability of A's winning a prize is known to be $1/5$ is $3/5$.

So $P(B/A) = 3/5$, thus $P(B/\text{Not } A) = 2/5$.

Now probability of A winning a prize after the statement of C would be

$$P(A/B) = \frac{P(A) P(B/A)}{P(A) P(B/A) + P(\text{Not } A) P(B/\text{Not } A)}$$

$$= \frac{\frac{1}{5} \times \frac{3}{5}}{\frac{1}{5} \times \frac{3}{5} + \frac{4}{5} \times \frac{2}{5}} = \frac{\frac{3}{25}}{\frac{3}{25} + \frac{8}{25}} = \frac{\frac{3}{25}}{\frac{11}{25}} = \frac{3}{11}$$

Thus the posterior probability of A (3/11) is more than the prior probability (1/5).

This problem can also be solved in a simpler way without using the notations of Baye's theorem. Thus the posterior probability of A's winning the prize two types of situation (1) C makes a right statement (2) C makes a wrong statement.

Event	Prior Probability	Conditional Probability	Joint Probability
A	1/5	3/5	$3/5 \times 1/5$
Not A	4/5	2/5	$4/5 \times 2/5$

A will win the prize if event (A) occurs and A will not win it if Not A occurs.

We want to calculate probability of A's winning after C has given the information that A has won - conditional probability of A's winning = $1/5 \times 3/5$.

Probability of all types of happening i.e., either A wins or does not win is marginal probability A wins or A does not win that is nothing but $(3/5 \times 1/5) + (4/5 \times 2/5)$.

Thus

$$\text{Probability of A's winning condition C's information} = \frac{\text{Chances of A's winning condition C's information} + \text{Chances of A losing condition C's information}}{\text{Chances of A's winning condition C's information} + \text{Chances of A losing condition C's information}}$$

$$\begin{aligned}
 &= \frac{\frac{1}{5} \times \frac{3}{5}}{\left(\frac{3}{5} \times \frac{1}{5}\right) + \left(\frac{4}{5} \times \frac{2}{5}\right)} \\
 &= \frac{\frac{3}{25}}{\frac{3}{25} + \frac{8}{25}} = \frac{\frac{3}{25}}{\frac{11}{25}} = \frac{3}{11}
 \end{aligned}$$

Illustration: Three major political parties Congress, BJP and Janata are contesting in general election. The chances of their forming government are $1/5$, $1/2$, $3/10$ respectively. Furthermore the chances of their continuing the liberalisation policy (if elected to power) are 0.7, 0.5 and 0.6 respectively. If it is known that the liberalisation policy continues what is the probability that congress has been elected to power.

Solution: Let the events be defined.

C party congress is elected to power

B party BJP is elected to power

J party Janata is elected to power

we are given $P(C) = 1/5$

$$P(B) = 1/2$$

$$P(J) = 3/10$$

If continuation of liberalisation policy is denoted by L,

We have $P(L/C) = 0.7$

$$P(L/B) = 0.5$$

$$P(L/J) = 0.6$$

The necessary computation are conveniently carried out in the following table.

Event	P(Event)	P(L/Event)	Joint Probabilities	
C	1/5	0.7	$1/5 \times 0.7 = 0.2 \times 0.7$	= 0.14
B	1/2	0.5	$1/2 \times 0.5 = 0.5 \times 0.5$	= 0.25
J	3/10	0.6	$3/10 \times 0.6 = 0.3 \times 0.6$	= 0.18
P(L)				= <u>0.57</u>

We are required to find out the probability that party C has been elected to power and continuing the liberalisation ie. we need $P(C/L)$

Applying Baye's theorem

$$P(C/L) = \frac{P(C) P(L/C)}{P(C) P(L/C) + P(B) P(L/B) + P(J) P(L/J)}$$

$$\text{From the table} = 0.14 / 0.57 = 0.24$$

REVIEW QUESTIONS

UNIT - I

STATISTICS - A CONCEPTUAL FRAMEWORK

1. Define Statistics and discuss its functions and limitations.
2. Explain how Statistics plays an important role in managerial planning and decision-making.
3. Explain the uses of Statistics in business.
4. Distinguish between Descriptive and Inductive Statistics.
5. Write an essay on the scope of Statistics with special reference to modern business and industry.
6. Discuss the relationship of Statistics with other sciences.
7. What are the various divisions of Statistics and what are their natures?
8. "Statistics should not be used as a blind man does a lamp post for support instead of for illumination" - Illustrate.
9. "The proper function of Statistics is to enlarge individual experience" - Comment.
10. Discuss the role of Statistical methods in economic planning with special reference to India.

UNIT - II

STATISTICAL ENQUIRY AND METHODS OF SAMPLING

1. Define a Statistical unit. State in brief the precaution you would take in the selection of a statistical unit for conducting an enquiry.
2. Define 'Primary Data'. Explain the various methods that are used in the collection of primary data, pointing out their merit and demerits.
3. Define 'Secondary Data'. State their chief sources and point out the dangers involved in their use. What precaution are necessary before using such data.

4. Define a random sample and show how you would achieve randomness. How do you select a random sample from a finite population?
5. What is Sampling? What precaution would you take in choosing a sample? Describe briefly the various types of sampling methods used in a statistical enquiry.
6. What do you mean by classification and tabulation? Explain the purpose and importance of classification and tabulation of statistical data.
7. Explain the different types of graphs and diagrams used for representing a frequency distribution with a suitable illustration.
8. Prepare a frequency table taking class intervals 20-24, 25-29, 30-34 and so on from the following data:

21	20	55	39	48	46	36	54	42	30
29	42	32	40	34	31	35	37	52	44
39	45	37	33	51	51	52	46	43	47
41	26	52	48	25	25	37	33	36	27
54	36	41	33	23	39	28	44	45	38

9. Draw a suitable diagram to represent the following data:

Year	Production (m.tonnes)	Year	Production (m.tonnes)
1979-80	43.07	1983-84	48.74
1980-81	39.58	1984-85	41.94
1981-82	44.05	1985-86	52.68
1982-83	39.58		

10. Draw a Pie diagram for the following data of plan outlays:

Agriculture and Rural Development	-	12.9%
Irrigation	-	12.5%
Energy	-	27.2%
Industries and Minerals	-	15.4%

Transport, communication etc.	-	15.9%
Social Services	-	61.1%
Total		<u>100.00%</u>

UNIT - III

MEASURES OF CENTRAL TENDENCY

1. What is Central Tendency? State giving illustrations, the circumstances when 'Median' may be more suitable measure of central tendency than the arithmetic mean.
2. State the empirical relationship between mean, median and mode after defining each.
3. Discuss the merits and demerits of Geometric mean. Explain its utility and algebraic characteristics.
4. Define weighted average and mention its applications.
5. (a) Calculate the the geometric mean of the following price relative.

Commodity	Price relative	Commodity	Price relative
Wheat	207	Sugar	124
Rice	198	Salt	107
Pulses	156	Oils	196

(b) Calculate the geometric mean from the following data:

6.5, 169.5, 11.0, 112.5, 14.2, 75.5, 35.5, 215.0.

6. The following table gives the weekly wages in rupees in a certain commercial organisation.

Weekly wages(Rs.)	30	32	34	36	38	40	42	44	46	48-50
Frequency	2	9	25	30	49	62	39	20	99	3

- (i) Calculate the median and the third quartile wages
- (ii) The number of wage earners receiving between Rs. 37 and Rs. 46 per week.

7. Calculate the Mean and Mode of the following distribution.

Size of the item	15	20	25	30	35	40	45	50
No. of items	4	12	30	60	80	90	95	97

8. The following are the prices of shares of a company from Monday to Saturday.

Days	Price (Rs.)	Days	Prices (Rs.)
Monday	200	Thursday	160
Tuesday	210	Friday	220
Wednesday	208	Saturday	250

Compute the average value by short-cut method.

9. Calculate the geometric and harmonic mean from the following data:
0.0031; 0.0051; 0.0351; 0.5691; 2.8254; 10.5; 1054.3; 1005.4
10. Compare and Contrast arithmetic mean, geometric mean and harmonic mean. Which one is least affected by extreme items?

UNIT - IV

MEASURES OF DISPERSION

1. What do you understand by Dispersion? What are the principal measures of dispersion? Discuss their merits and demerits.
2. In what ways measures of variation supplement measures of central tendency?
3. Define (i) mean deviation and standard deviation and (ii) distinguish between them.
4. An Analysis of the monthly wages paid to workers in two firms A and B belonging to the same industry, gives the following results.

	Firm A	Firm B
No. of workers	160	150
Average wage	560	575
Variance of wage distribution	400	625

Find out:

- (a) Which firm pays larger amount as monthly wages?
 (b) In which firm is there greater variability in individual wages?
5. Calculate the mean deviation from the mean for the following data:

Size	2	4	6	8	10	12	14	16
Frequency	2	2	4	5	3	2	1	1

6. In two factories A and B engaged in the industry, the average weekly wage and standard deviation are as follows:

Factory	Average monthly wage(Rs.)	S.D. of wages (Rs.)	No. of wage earners
A	460	50	100
B	490	40	80

- (i) Which factory A or B pays larger amount as weekly wages?
 (ii) Which factory shows greater variability in the distribution of wages?
 (iii) What is the mean and standard deviation of all the workers in two factories taken together?
7. Calculate the quartile deviation for the following data and verify it graphically.

Class interval	8-12	12-16	16-20	20-24	24-28
Frequency	5	12	20	10	3

8. Find out the standard deviation from the following data:

X:	10	11	12	13	14
f:	3	12	18	12	3

9. Find the coefficient of standard deviation from the following data:

x:	10-19	20-29	30-39	40-49	50-59	60-69
f:	167	127	99	87	189	78
x:	70-79	80-89	90-99			
f:	93	127	69			

10. What is skewness? How does it differ from dispersion? Describe the various measures of skewness.

11. Show by means of sketches the relative position of mean, median and mode for frequency curves which are skewed for the right and left respectively.
12. Differentiate moments and kurtosis with suitable example.

UNIT - V

CORRELATION

1. Explain what is meant by Correlation between two variables. What are the methods of finding existence of Correlation? How can it be measured?
2. What is Rank Correlation? State the merits and demerits of Spearman's Rank Correlation.
3. Distinguish the following with suitable example.
 - (i) Positive and negative correlation
 - (ii) Linear and non-linear correlation
 - (iii) Simple, partial and multiple correlation
4. (a) Define the Pearson coefficient of correlation. Interpret r , when $r = 1, -1$ and 0 .
(b) Mention the uses of Correlation.
5. Apply Spearman's Rank difference method and calculate coefficient of Correlation between X and Y from the data given below.

X:	22	28	31	23	29	31	27	22	31	18
Y:	38	25	25	37	31	35	31	29	28	30

6. From the following data calculate Karl Pearson's Coefficient of Correlation.

Year	1977	1978	1979	1980	1981	1982	1983
Exports (in crores)	115	118	122	120	126	127	125
Imports (in crores)	140	138	142	146	145	148	146

7. Calculate Karl Pearson's coefficient of Correlation and interpret its value given the following:

	X series	Y series
Sum of deviation from assumed mean	-14	18
Sum of squares of deviations from assumed mean	4304	6308
Sum of the products of deviations from their respective assumed mean values	1510	
No. of pairs of observations	12	

Calculate Karl Pearson's coefficient of Correlation between X and Y.

8. In order to find out the correlation coefficient between two variables X and Y from 12 pairs of observations, the following calculations were made:

$$\Sigma x = 30, \Sigma y = 5, \Sigma x^2 = 670, \Sigma y^2 = 285, \Sigma xy = 334$$

On subsequent verification it was found that the pair ($x = 11, y = 4$) was copied wrongly, the correct value being $x=10, y=14$. Find the current value of r .

9. The rank of the same student in a test in Accountancy and Statistics are given below. Calculate the correlation coefficient and comment the value.

Rank in Accountancy 1 2 3 4 5 6 7 8 9 10

Rank in Statistics 1 3 5 6 7 4 8 10 9 2

10. The corresponding values of two variables are given below:

X: 2 3 5 8 9

Y: 4 6 19 16 18

Karl Pearson's r between X and Y is -1, 0, +1, None of these.

UNIT - VI

REGRESSION ANALYSIS

1. What is Regression Analysis? Indicate its uses in business.
2. Distinguish clearly between Correlation and Regression Analysis.
3. Explain what are Regression lines. Why are there two such lines? Also derive their equations.
4. Obtain the lines of regression for the following data.

Sales Revenue (Rs. crores)	Advertisement Expenditure (Rs. Lakhs)				Total
	5-15	15-25	25-35	35-45	
75-125	4	1	-	-	5
125-175	7	6	2	1	16
175-225	1	3	4	2	10
225-275	1	1	3	4	9
Total	13	11	9	7	40

Estimate the sales for an advertisement expenditure of Rs.80 lakhs.

5. What is meant by simple linear model? Under what assumptions are the parameters of the model estimated?
6. Given $X = 68$, $Y = 150$, $\sigma_x = 2.5$, $\sigma_y = 20$ and Correlation coefficient between X and Y as $+0.6$, estimate the value of Y , when X is 60.
7. Calculate the regression equation of X on Y and Y on X from the following data and estimate X when Y is 20:
 $X:$ 10 12 13 17 18
 $Y:$ 5 6 7 9 13
8. In a regression analysis: $n = 25$, $\Sigma x = 75$, $\Sigma y = 50$, $\Sigma x^2 = 623$, $\Sigma xy = 30$ and $\Sigma y^2 = 228$. Find the estimated regression equation and the estimated variance for the slope parameter.
9. Write down the two regression equations that may be associated with the following pairs of values:

X: 152 114 138 154 144 153 141 117 136 154

Y: 193 300 414 594 676 549 320 483 481 659

10. A study of price levels of a commodity at two places revealed the following:

	Place A Rs.	Place B Rs.
Average price per unit	124.60	135.90
Standard deviation	13.50	17.10
Correlation between A and B	0.72	

(i) Obtain the two regression equations.

(ii) Find the expected price level at place A when it is Rs. 140 at place B, and

(iii) Find the expected price level at place B when it is Rs. 120 at place A.

UNIT - VII

INTERPOLATION AND EXTRAPOLATION

1. Explain briefly the usefulness of Interpolation and Extrapolation in statistical studies.
2. Give the meaning of Interpolation and Extrapolation. Mention the assumption underlying Interpolation and Extrapolation.
3. State Newton's formula for Interpolation and discuss some its uses.
4. Explain Lagrange's method of Interpolation. Point out its usefulness.
5. Using the binomial expansion method estimate the missing figure for 1981-82 from the data given below:

Year	Average value (Rs. crores)	Year	Average value (Rs. crores)
1977-78	89.56	1980-81	27.85
1978-79	52.54	1981-82	?
1979-80	37.36	1982-83	23.36

6. Extrapolate the business done in 1984 from the following data:

Year	1979	1980	1981	1982	1983
Business done (Rs. lakhs)	150	235	365	525	780

7. Explain clearly the concept of Time Series analysis. Indicate the importance of such analysis in business.
8. Explain briefly the various methods of determining the trend in Time Series. Explain the merits and demerits of each method.
9. Fit a line trend by the method of least square to the following data:

Year	1975	1976	1977	1978	1979	1980
Production (Rs. crores)	7	10	12	14	17	24

UNIT - VIII

PROBABILITY

1. Define Probability and explain the importance of this concept in statistics.
2. Explain with examples the concept of independence and mutually exclusive events in Probability.
3. State and prove the addition and multiplication theorems of Probability for two mutually exclusive events.
4. What is Bayes' theorem? Explain with the help of a suitable example.
5. Define mathematical Probability of an event A. If A and B are two events, can they be mutually exclusive as well as independent?
6. Two cards are drawn from a well shuffled pack of 52 cards. Find the Probability that they are both aces if the first is (i) unplaced and (ii) not replaced.
7. Three perfect coins are tossed together. What is the Probability of getting at least one head?

8. An investor buys six different types of shares. The Probability for prices of these shares to increase are respectively 0.3, 0.4, 0.7, 0.5, 0.8 and 0.1. Find the Probability that prices of all six types of shares will increase.
9. A factory finds that on an average 20% of the bolts produced by a given machine will be defective for certain specified requirement. If 10 bolts are selected at random from the day's production of this machine, find the Probability that (i) exactly 2 will be defective, (ii) 2 or more will be defective, (iii) more than five will be defective.
10. The Probability that there is atleast one error in accounts statement prepared by A is 0.2 and for B and C they are 0.25 and 0.4 respectively. A, B and C prepared 10, 16 and 20 statements respectively. Find the expected number of correct statements in all.

**Directorate of Distance Education,
Pondicherry University,**

R.V.Nagar, Kalapet, Puducherry - 605 014.

Phone : 0413-2654439, 0413-2655256

Fax : 0413-2655258

E-mail : director.dde@pondiuni.edu.in