+3 Commerce (Pass/Hons.)
B.B.A. / IMBA

BUSINESS STATISTICS

Author

Mrs. Sunita Mall,
Lecturer, ICFAI National College, Cuttack

Editor

Dr. R. K. Bal
Professor P.G. Deptt. of Commerce, Utkal University
&
Mr. S.K. Acharya
Asst. Course Co-ordinator,
Management Programme, DDCE, Utkal University

General Editor

Prof. S.P. Pani
Director, DDCE

UTKAL UNIVERSITY

D.D.C.E.
Education For All

DIRECTORAT OF DISTANCE & CONTINUING EDUCATION
UTKAL UNIVERSITY, BHUBANESWAR-751007

USINESS STATISTICS

Author :

Mrs. Sunita Mall
Lecturer, ICFAI National College, Cuttack

Published by :

DIRECTOR,
DIRECTORATE OF DISTANCE & CONTINUING EDUCATION
UTKAL UNIVERSITY, VANIVIHAR, BHUBANESWAR-751007
Phone No.: 0674-2376700

Published : 2015

Copies : 600 nos.

Printed at :
inteCAD
442, Saheed Nagar, Bhubaneswar - 751 007
Tel. : 0674 - 2544 631

UTKAL UNIVERSITY

D.D.C.E.
Education For All

DIRECTORATE OF DISTANCE & CONTINUING EDUCATION
UTKAL UNIVERSITY : VANI VIHAR
BHUBANESWAR:-751007

From the Director's Desk

The Directorate of Distance & Continuing Education, originally established as the University Evening College way back in 1962 has travelled a long way in the last 52 years. **'EDUCATION FOR ALL'** is our motto. Increasingly the Open and Distance Learning institutions are aspiring to provide education for anyone, anytime and anywhere. DDCE, Utkal University has been constantly striving to rise up to the challenges of Open Distance Learning system. Nearly ninety thousand students have passed through the portals of this great temple of learning. We may not have numerous great tales of outstanding academic achievements but we have great tales of success in life, of recovering lost opportunities, tremendous satisfaction in life, turning points in career and those who feel that without us they would not be where they are today. There are also flashes when our students figure in best ten in their honours subjects. In 2014 we have as many as fifteen students within top ten of honours merit list of Education, Sanskrit, English and Public Administration, Accounting and Management Honours. Our students must be free from despair and negative attitude. They must be enthusiastic, full of energy and confident of their future. To meet the needs of quality enhancement and to address the quality concerns of our stake holders over the years, we are switching over to self instructional material printed courseware. Now we have entered into public private partnership to bring out quality SIM pattern courseware. Leading publishers have come forward to share their expertise with us. A number of reputed authors have now prepared the course ware. Self Instructional Material in printed book format continues to be the core learning material for distance learners. We are sure that students would go beyond the course ware provided by us. We are aware that most of you are working and have also family responsibility. Please remember that only a busy person has time for everything and a lazy person has none. We are sure you will be able to chalk out a well planned programme to study the courseware. By choosing to pursue a course in distance mode, you have made a commitment for self improvement and acquiring higher educational qualification. You should rise up to your commitment. Every student must go beyond the standard books and self instructional course material. You should read number of books and use ICT learning resources like the internet, television and radio programmes etc. As only limited number of classes will be held, a student should come to the personal contact programme well prepared. The PCP should be used for clarification of doubt and counseling. This can only happen if you read the course material before PCP. You can always mail your feedback on the course ware to us. It is very important that you discuss the contents of the course materials with other fellow learners.

We wish you happy reading.

(S.P. Pani)
DIRECTOR

CONTENT

BUSINESS STATISTICS

UNIT – I

LESSON – 1

NATURE OF STATISTICS

INTRODUCTION

In your day-to-day life you must have heard statements like "India's population is increasing at the rate of 2.2% per annum." "The profits of A.B.C., Co.Ltd. increased from Rs.5 lakhs in 1977 to Rs.6.Lakhs in 1978", etc. Such statements are very common in class room lecturers, daily newspapers, speeches or on radio. They contain facts and figures. They are very convenient forms of communications as well as very clear, precise and meaningful. Analysis of such statements helps one in arriving at certain conclusions. For example, 146 students have passed B.Com. with 1st class Honours from colleges under Utkal University and there is provision for 80 student to be admitted to M.Com class in 1979. It is evident from the statement that the rest boys are deprived of higher education.

What is it ? – Nature of Statistics

Facts and figures regarding any phenomenon (having inter connection) whether they relate to population, production, national income, profits, sales etc. are called statistics'. In this respect the term statistics is synonymous (equal) with facts and figures and is used in plural sense. They relate to numerical data and take the form or counts and measurements.

All numerical statements do not constitute statistics. Statements may convey numerical information, but if not connected with each other do not constitute statistics. To constitute statistics two characteristics must be satisfied : - (1) The number should represent a phenomena rather than an isolated condition of a subject. A single incident does not make a phenomenon, or a series of numerical facts which have no interconnection cannot be said to make a phenomenon. (2) Such numbers are capable of analysis regarding causes and effects, behaviour pattern, changes in magnitude, trend of movement etc. Statistics counts of numbers of various kinds, as males and females, married and unmarried, or postgraduates and undergraduates, They may also be numbers computed on the basis of these measurements or counts, e.g. the proportion of female students their average heights or average weight etc. In this sense numerical facts should be more correctly called 'data' rather that 'statistics'.

In addition to meaning numerical facts, 'Statistics' also refers to a subject. Just as Mathematics refers to a subject as well as to symbols, formulae and theorems and Accounting refers to principles and methods as well as to accounts, balance sheets and income statements statistics refers to a body of methods of obtaining and analyzing data in order to base decision on them. It is a branch of scientific method used in dealing with phenomena (like income agricultural production, marks secured by students etc.) that can be described numerically either by counts or by measurements. Thus the world statistics refers either to quantitative information or to a method of dealing with quantitative information.

The methods by which statistical data are analysed are called statistical methods, although the term is sometimes used more loosely to cover the subjects 'Statistics' as a whole. Mathematical theory which is the basis of these methods is called the theory of statistics or mathematical statistics. Statistical methods are applicable to a very large number of fields like economics, sociology, anthropology, business, agriculture, psychology, medicines, education etc.

Statistics is usually not studied for its own sake, rather it is widely used as a tool in the analysis of problems in natural, physical and social sciences, Statistical methods are used by governmental bodies, private business firms, and research agencies as an indispensable aid in (1) forecasting (2) controlling and (30 exploring. Such statistical methods range form the most elementary descriptive devices to complicated mathematical procedures which can only be understood by persons receiving special training.

Definitions

There are many definitions of the term 'Statistics'. Some authors have defined statistics 'as statistical data' as used in plural sense whereas other as statistical methods' used in singular sense. A few well recognized definition are analysed below : -

A. As Statistical Data : -

Webster defined statistics as "the classified facts representing the condition of the people in a state – especially those facts which can be stated in numbers or in tables of numbers or in any tabular or classified arrangement'. If we analyse this definition, it cannot be regarded as the one most commonly used bècause it is very narrow. It confines the scope of statistics to only such facts and figures which relate to the conditions of the people in a state, In fact, statistics at present relate to all aspects of human activities. This definition does not suit modern conditions.

Value and Kendall defined statistics as, "By Statistics we mean quantitative data affected to a marked extent by multiplicity of causes". The definition also is very narrow because it does not include many important features of statistics as used in statistical data.

Harac Secrilsts defined statistics as "By Statistics we mean aggregates of facts effected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of 'accuracy, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner, for a pre-determined purpose and placed in relation to each other".

This definition clearly points out certain characteristics which numerical data must posses in order that they will be called statistics.

These are as follows : -

1. **Statistics are Aggregates of Facts –**

This means that statistics represents a number of facts. Single and isolated figures are not statistics for the simple reason that such figures are unselected and cannot be compared. If the income of an officer is stated as Rs.1,800 per month this does not constitute statistics, although it is a numerical statements of facts. Similarly a single figure relating to production, sale, birth & death, employment, purchase, accident etc, cannot be regarded as statistics. But aggregates of such figures over a period of time or figures relating to some countries over the same period of time will be called as statistics because they are comparable and some relationship between them can be found out.

2. **Statistics are Affected to a Marked Extent by Multiplicity of Causes –**

Usually facts and figures are affected to a considerable extent by a number of forces operating together. Statistics are aggregates of such facts only which grow out of a 'variety of circumstances' – when their size, shape or form at any particular moment is the result of the action and interaction of number of forces, differing amongst themselves and it is not possible to say as to how much of it is due to any one particular cause. For example statistics of production of rice are affected by the rain fall, quality of soil, seeds and manure, method of cultivation etc. It is very difficult to study separately the effect of each of these forces on the production of the rice. The same is true for statistics of prices, imports exports, sales and profits etc.

3. **Statistics are Numerically Expressed –**

All statistics are numerical statements of facts, i.e. expressed in numbers. Qualitative statements, like 'Population of India is rapidly increasing or India is a poor country, do not constitute statistics, Such statement like'. The population of India increased by 25% in 1975 against 2.2% in 1974 is a statistical statement.

4. **Statistics are Enumerated or Estimated According to Reasonable Standards of Accuracy –**

This means that if aggregates of numerical facts are to be called statistics they must be reasonably accurate. This is necessary because statistical data are to serve as basis for statistical investigations. Facts and figures about any phenomenon can be derived in two ways- by actual counting and measurement or by estimate. Estimates cannot be as precise and accurate as actual counts or measurements. Where the scope of statistical enquiry is very wide or where the numbers very large, actual counting or measurement (enumeration0 is usually out of question and figures can only be estimated by experts. The degree of accuracy expected in such figures depends to a large extent on the purpose for which statistics are collected and also on the nature of particular problem about which data are collected. Whatever standard of accuracy is one adopted it should be uniformly maintained throughout the enquiry. Reasonable standards of accuracy must be obtained, otherwise data may be allogether misleading.

5. Statistics are Collected in a Systematic Manner –

Before collecting statistical data in respect of any phenomenon, a suitable plan of data collection must be prepared and the work must be carried out in a systematic manner otherwise it would lead to wrong conclusions. Such data will not conform to reasonable standards of accuracy.

6. Statistics are Collected for a Pre-determined Purpose –

The purpose of collecting data must be decided in advance. Moreover, such purpose should be well defined and specific. A general statement of purpose is not enough. For example, if the objective is to collect data on prices one must have to decide in advance whether they will be wholesale prices or retail prices, how frequently they will be collected and in respect of what commodities they will he collected.

7. Statistics Should be Placed in Relation to Each Other –

Numerical data are collected mostly for the purpose of comparison. If the collected figures are not capable of being compared with each other they loose a very large part of their value. Comparison can be made only if data are homogenous, i.e. they relate to the same phenomenon or subject and only likes are compared with like. Statistical data are often compared period-wise or region wise. Data of heterogeneous character are not comparable as they cannot be placed in relation to each other. The per capital income of an Indian may be compared over different years, or production of wheat of different countries during the same period may be compared. But it will be meaningless to compare the height of human beings with their incomes.

Some definition of statistics :

As statistical methods or science of statistics many authors have defined statistics in different ways. A few important definitions are analysed below :

Seligman defines it as "Statistics is the science which deals with the method of collecting, classifying, presenting, comprising and interpreting numerical data collected to throw some light on any sphere of enquiry. "Though this definition is very short it is quite simple as well as comprehensive.

King defines it as, "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 is not very exhaustive and it limits the scope of the science of statistics.

Boddington has defined statistics as "the science of estimates and probabilities. This definition is also not acceptable because estimates and probabilities are only a part of statistical methods.

Objective of Statistics

In the words of A.L. Boddington. "The ultimate and of statistical research is to enable comparison to be made between past and present results, with a view to ascertaining the reasons for changes which have taken place and the effect of such changes in the future".

From the above statement it is very clear that facts and figures are collected about any phenomenon relating to the past and the present and from their analysis some conclusion is arrived in regard to its future. A systematic comparison is the main object of statistics. To achieve this, data relating to past and present are collected and presented in the shape of time-series from which valuable conclusions are drawn. The conclusions are used fo' the purpose of for-casting the future trend of different problems.

It helps in assessing the results of past achievements of human activities and it is also useful for making for-casts about he future course of events.

Cause of the Recent Growth of Statistics :

The following are two main factors, which are responsible for the development of statistics in modern times : -

1. Increased Demand for Statistics –

In the modern time considerable development has taken place in the field of business and commerce, governmental activities and science. Statistics helps in formulating suitable policies, and as such its need has been increasingly felt in all these spheres.

The magnitude of business has considerably increased and business affairs have become very much complicated. This has resulted in an increased demand for statistics, the complexity in business is on account of numerous government regulations, labour disputes, ever increasing taxes and technological revolution, which the business world has witnessed in recent years. Statistics have to be collected on all these problems.

The activities of the government have also increased in size and complexity. Modern states are welfare states and there is hardly any sphere in which the government has not entered. With the enlargement of the functions of government, the demand for statistics has also increased.

In the field of science many types of researches are going on. Research workers make extensive use of statistical data as their tool.

2. **Decreasing Cost of Statistics**

The time and cost of collecting data are very important limiting factors in the use of statistics But with the development of electronic machines, such as calculators, computers etc, the cost of analyzing data has considerably gone down. This has led to the increasing use of statistics in solving various problems.

Moreover with the development of statistical theory the cost of collecting and processing data has gone down. For example, sampling techniques enable us to know the characteristics of the population by studying only a part of it. Again, since 1935, a branch of statistics known as design of experiments has made rapid progress and it is now possible to collect analyses and compile statistics more promptly and economically.

Many scholars have contributed to the science of statistics but Sir Ronald Fisher must be credited with at least half of the essential and important developments in the theory as it stands now.

The theory of statistics is not complete and final. Many more problems are to be solved by research scholars in future so that the theory can be still better used.

Importance of Statistics

Statistical methods have become useful tools in the world of affairs. Economy and a high degree of flexibility are the important qualities of statistical methods that render them especially useful to businessmen, economists, scientists, government, research scholars etc.

1. **Importance in Economic and Social Studies -**

Prof. Alfred Marshall observed "Statistics are the straw out of which I like every other economist, have to make bricks". This proves the significance of statistics in economics. In the field of economics it is almost impossible to find problem, which does not require an extensive use of statistical data. Economics is concerned with he production and distribution of wealth as well as consumption, saving and investment income. Statistical data and statistical methods are of immense help in proper understanding of the economic problems and in the formulations of economic problems and in the formulation of economic policies. That to produce, how to produce and for whom to produce – these are the questions that need a lot of statistical data in the absence of which it is not possible to arrive at correct conclusion. Statistics of production help in adjusting the supply to demand. Statistics of consumption enable us to find out the way in which people of different sections of society spend their income. Such statistics are very helpful to know the standard of living and taxable capacity of the people. In reducing disparities in the distribution of income and wealth statistics are of immense help. Similarly in solving problems of rising prices, growing population, unemployment, poverty etc., one has to depend upon statistics. Statistical methods help not only in formulating appropriate economic policies but also in evaluating their effect. For example in order to check the over-growing population if emphasis has been placed on the family planning methods, one can ascertain statistically the efficiency of such methods in attaining the desired goal.

In recent years economics as which comprised the application in statistical methods to the theoretical economic methods is widely used in economic research. Statistical methods of sampling are useful for collecting the basic data of economic studies. The demand function, production function, cost functions and consumption functions present many difficult problems in the analysis of which statistical tools are of immense use. Economists now utilize statistical data in building a sound factual foundation for their reasoning specially in the following fields :-

(i) In Measuring Gross National Product and input-output analysis ;

(ii) Utilization of financial statistics in the fields of money and banking, short-term credit consumer finance and Public finance.

(iii) In studying facts in the study of competition, oligopoly and monopoly by comparing market prices, cost and profits of individual firms.

(iv) In studying the theories of prices, pricing policy and price trends and their relationship to the general problem of inflation,

(v) In operational studies of public utilities which require statistics and legal tools of analysis.

(vi) In analyzing population and-economics and economic geography and in many other fields.

Mathematics and its main offspring, statistics and accounting are the powerful instruments which the modern economist has at his disposal and of which business through the development o research agencies and methods is making constantly great use.

A sociologist may attempt to demonstrate with the help of statistical data the relationshi between sales of liquor and crime or between suicide and poverty etc.

2. Statistics and the State –

Since times the ruling kings and chiefs have relied much on statistics in framing suitabl military and fiscal policies. Most of the statistics such as that of crimes, military strength, populatio taxes etc., which were collected by them were the byproduct of administrative activity. In rece years the functions of the state have increased tremendously. The concept of a state has changed from that of the simply maintaining law and order to that of a welfare state. Statistics and statistical data are indispensable in these days for a clearer appreciation of any problem affecting the welfare of mankind. Problems relating to poverty, unemployment, food shortage, protective tariff, uneconomic agricultural holdings etc. cannot be fully weighed without statistical data.

Statistics to-day are not exclusively a by-product of administrative activity. The State collects statistics on several problems. Such statistics help in framing suitable policies and studying their effects. All ministries and departments of the government whether they be finance, transport, defence, railways, food, commerce post and telegraph or agriculture depend much on factual data for their efficient functioning. For example, the transport department cannot solve the problem of transport in Bhubaneswar, Cuttack or Rourkela unless it knows how many buses are operating at present, what is the total requirement in these places and many additional buses are to be added to the existing ones. Not only during peace times, but during days of war also statistics are indispensable. It is impossible to fight a war successfully in the absence of factual data a.·out the strength of the enemy.

Statistics are so significant to the state that the government in most countries is the biggest collector and user of statistical data. Such data are of immense help to many institutions who further process them and arise at useful conclusions, which help in decision-making. Official statistics occupy a very important place in almost all countries.

3. Need in Planning –

Economic activities are more closely directed to the production of such goods and the provision of such services, as the government may decide from time to time to be most urgently required. Our future is being planned and such planning to be successful must be soundly based on the correct analysis of complex statistical data. Planning cannot be imagined without statistics. If you study the economic plans implemented in various countries in recent times you will find that all of them are statistical study of the economic resources of the respective countries, and they suggest possible ways and means of utilizing these resources in the best possible manner. Various plans that have been prepared for the economic development of India have also made use of the statistical data available about various economic problems. One of the re·sons of failure of Indian plans is what we have insufficient or incorrect data. National Sample Survey Scheme was primarily started to collect data for the use in planning in India. The success failure of plans and the progress report bear witness to it. Thus one finds that in the field of economic planning the sue of statistical data and methods is indispensable.

4. Statistics Discloses Causal Relationship Between Related Facts–

Food prices are low when new harvest arrives in the market and they rise in the off-season. A-clever buyer buys his annual requirements at the harvest time. Such causal studies are at the bottom of all sound human endeavour. Statistics is the light bearer that enlightens the way to life's adventure.

5. Usefulness in Business and Commerce –

With the growing size and ever increasing competition, the problems of business enterprises are becoming complex and they are using more and more statistics in decision-making. Many business problems are now solved, through statistical data and method used in interpretation. An owner of a small firm, in the past, acted as the storekeeper, manager, accountant, salesman, purchaser etc. It was possible for him to make personal contacts with the customers and know exactly what they wanted. With the growth in size, business problems have become very complex, and it is impossible for the owners to have personal contacts with thousands of customers, Management has become a specialized job and a manager has to plan, organize, supervise and control the operations of the business house. Most of the production these days is in anticipation of demand and therefore unless a very careful study of the market product and consumer is made the firm may not be able to make profits. Trial and error method of taking a decision no more holds good. To-day the businessman succeeds or fails accordingly as his forecasts prove to be accurate or wrong. Business now runs on estimates and probabilities. If the estimates or forecasting of the business man is correct, there will be higher degree of accuracy and it will result in profits to the business, otherwise there may be a loss, In recent years statistics and statistical methods have provided the businessman with one of his most valuable tools for decision-making.

Business activities can broadly be grouped under the following 1- Production, 2- Sale, 3- Purchase, 4-Finance 5-Personnel, 6-Accounting 7-Market and Product Research and 8-Quality Control. For better or worse the modern business executive is largely dependent on Statistical data and method of analysis for essential information. Statistical information is needed from the time the business is launched till the time it comes to an end. At the time of floatation of the concern, facts are required for the purposes of drawing up the financial plan of the proposed unit, With the help of statistical methods in respect of any of the above categories, much quantitative information can be obtained which can be of immense use in formulating suitable policies. For example, a manufacturer must know in advance how much is to be produced, how many workers and how many raw material will be needed to be produce that estimated quantity, and what quantity type, size, colour or grade of the product is to be manufactured. He must, therefore, have a sound production plan which cannot be prepared without Statistical analysis of the past and present data.

Statistical methods of analysis are helpful in the marketing function of an enterprise through its help in market research, advertisement campaigns in comparing the sales performances. Statistics also directs attention towards the effective use of advertising funds. Similarly, in setting labour disputes; in having a sound financial structure; in purchasing quality materials or products with the competitive price etc. statistical data and methods are valuable in business and commerce. The theory and technique of sampling can be used in connection with the various business surveys with a considerable saving in time and money. These techniques are now being extensively used in test checking of accounts. Statistical quality control is being used in industry for establishing quality standards for products for maintaining the requisite quality, and for assuring that the individual lots sold are of a given standard of acceptance.

The scientific management movement in India has emphasized the need for collecting facts and interpreting them carefully so that it brings benefits to commerce. Operations research, the main offspring of scientific management, has been statistical in its approach and is very helpful in the filed of business and commerce. Thus you see that statistics are the life-blood of successful commerce.

Though statistical methods are extremely useful in taking decisions they are not perfect substitute for common sense. A practitioner of business statistics must, therefore, combine the knowledge of he 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.

6. Utility to Bankers, brokers, Insurance Companies, Underwriters, Investors in Insecurities, Railways etc. –

Bankers, brokers of stock exchanges, investors in insurance companies or corporations and public utility concerns make extensive use of statistical data. A banker has to make a statistical study of business cycles to forecast a probably boom or a depression and has to study in detail the seasonal variations in the demand for call money form its depositors. On the basis of such demand, bankers usually keep a reserve.

Stock-exchange brokers, speculators and investors use statistical data in for-casting the demand for shares, stocks etc. They collect information about interest rates fluctuation of investment market to have a definite idea about the situation to gain maximum profit. They also collect information about money rates at different center and study their future trends.

Insurance corporations and companies have to depend upon mortality tables, vita statistics and they have to collect facts lie and incidents, which had taken place in the past. Theory of probability works out fully in the field of insurance. Premium rates in life insurance have to be fixed on the basis of mortality rates in different age groups. Sickness or unemployment insurance also depends on statistical data.

Public utility concerns like railways, electric supply companies, water supply companies, gas supply companies etc. also make extensive use of statistics. A railway operating over a wide area has many sources of possible wasteful expenditure. Statistical data and methods help the railways and other public utility concern to avoid wastages of all kind and work efficiently and economically.

7. **Desirability In Research –**

Statistics is indispensable in any research work. Most of the advancement in knowledge has taken place because of experiments conducted with the help of statistical methods. For example experiments about crop yields and different types of fertilizers and soils or the growth of animals under different diets and environments are mostly designed and analysed with the help statistical methods. Statistical methods also affect research in medicine and public health. In market research, quantity control etc, statistical data are not useful but necessary. In fact, there is hardly any research work to-day where research worker does not make extensive use of statistical data. It is also impossible to understand the meaning and implications of most of the findings of research in different fields without statistical facts and methods.

Limitations of the Science of Statistics –

From the discussions on importance and indispensability of statistics you might by thinking statistics are like magical devices, which always provide the correct solution to problems. But actually it is not so. Unless the data are properly collected and critically interpreted there is every likelihood of drawing wrong conclusions. Statistical techniques cannot be made to answer all our queries. Therefore, It is necessary to know the limitations and the science of statistics : -

1. **Does not Study Qualitative Phenomena –**

Statistics can be applied only to hose phenomena and he problems, which are capable of quantitative expressions. Such characteristics, which cannot be expressed in figures have very little use of statistical methods. In a study of honesty, statistical method cannot be of much help. Health, culture, character, friendship skill, poverty, cruelty, pessimism etc. cannot be quantitatively expressed. In studying any such phenomenon the statistical aspect is assumed to be subsidiary to other considerations. These subjective concepts can be related in an indirect fashion to numerical data. For example, we may study the intelligence o f students on the basis of the makes obtained by them in an examination. Honesty itself may not be capable of quantitative analysis but many factors, which are related to this phenomenon, are capable of being expressed in figures and as such can throw some fight on the study of the problem. A study of the number of thefts or cases of cheating or swindling can indirectly tell us something of the problem under study, Similarly, health of the people is judged by a study of its death rate, longevity of life and the prevalence of any disease. Again, the crime in different countries can be measured indirectly in terms of men and women who go to prison and if the number of such persons is decreasing, we can say that there is a better enforcement of the law and crime in the country. All these figures only indirectly relate to the real problem. They are subsidiary to other information like the manner in which people in the two countries live, the value they attach to principles of right conduct, the type of work they perform, the food they take and so on.

2. **Does not Reveal the Entire Story of a Problem –**

Since many problems are affected by such factors, which are incapable of statistical analysis, it is not always possible to examine a problem in all its manifestations only by a statistical approach. Statistical method is one of the several methods of studying a problem. It helps us in studying the trends, in framing an idea of the probabilities, in knowing how a given phenomenon has been behaving generally. It does not lead us to unquestionable conclusions. Many problems have to be examined in the background of a country's culture, philosophy or religion. Statistical results should not always be treated as the sole determinants of the value of a group. Such results are more in the nature of estimates than exact statements. These may be approximately correct. If a statistical conclusion or inference is confirmed by other available evidences and methods of study it can be taken as fairly reliable.

3. **Statistical Laws are True only on Average or in the Longrun –**

Statistics as a science is not as accurate as many other sciences are and statistical methods are not very precise and correct. They are not like the exact laws of physical sciences, which are said to hold true in every individual case that is subject to them. Laws of statistics are not universally true like the laws of Physics or astronomy. Statistics deals with such phenomena, which are affected by a multiplicity of causes and it is not possible to study the effects of each of these factors separately as is done under experimental methods. Due to this limitation in the statistical methods the conclusions arrived at are not perfectly accurate but they show approximate tendencies. Hence the same conclusions cannot be arrived at under similar conditions at all times.

4. **Does not Study Individuals –**

Statistical methods have no place for an individual item of a series. Statistics deals with aggregates or mass phenomena and as such, throws light on the characteristics of the whole of a given group. Of course, for purposes of analysis these aggregate are very often reduced to single figures. A statistical series is condensed into an average for purpose of comparison (this will made clear to you in another lesson) though an individual item of the series has no specific recognition. If only 100 persons die of starvation in India and if the percentage of these deaths to the total population of the country is negligible, statistically we will be justified in ignoring it. But this fact does not reduce the torture of death for these 100 people. From this point view is something very important and material but in statistical analysis such a problem does not occupy any significant place. This type of apathy to individual items of a series is a serious limitation or defect in many investigations.

You can understand this aspect still better with the following example. If the per capita income (on an average) has been rising for the last twenty years, it will be presumed that income of people has risen. It may so happen that, only rich persons have become richer, and very poor persons, the lower middle class people, these individuals in no way have more incomes. In fact, the per capita income will not reveal the extent of their miseries.

5. **Statistics is Liable to be Misused and Abused –**

The greatest limitation of statistics is that only one who has an expert knowledge of statistical methods can scientifically handle statistical data. Statistics, like medicines in the hands of quacks, are capable of being easily misused by the in experts. One might interpret statistics according to his convenience and make the worse appear to be the better case. Many people, therefore, look at statistics with an eye of suspicion. Statistics is a delicate tool and should be used with caution. Misuses are as common as valid uses of statistical uses of statistics. The misuse of statistics may arise due to several reasons. For example, if statistical conclusions are based on incomplete information, one may arrive at fallacious conclusions. The statement that the incidents of death among sick persons is higher in hospitals than at home due to lack of proper treatment and care. But this conclusion turns out to be completely erroneous if it si borne in mind that in India only seriously ailing persons are hospitalized. Similarly, due to change in definition, inaccurate measurement and classification, inappropriate comparison, defective method in selecting cases etc. statistical conclusions are likely to be misused and abused. Statistics are like clay and they can be moulded in any manner so as to establish right or wrong conclusions. In this context, W.I.King pointed out "One of the shortcoming of statistics is that they do not bear on their face the label of their quality". It requires experience and skill to draw sensible conclusions from the data; otherwise, there is every likelihood of wrong interpretations. The very fact that statistics may lead to wrong conclusions in the hands on in experts and inexperienced people limits the possibility of mass popularity of such a useful science. Statistics also cannot be used to full advantage in the absence of proper understanding of the subject to which it is applied.

COLELCTION OF DATA

The collection of data begins only after proper planning of the statistical enquiry as discussed above. The collection of data refers to a purpose of gathering of information relevant to the subject matter of enquiry. The method of collection of data depends mainly upon the nature, object and the scope of enquiry on one hand and the availability of resources and the time on the other. Statistical data may be classified in to primary and secondary depending upon the nature of data and mode of collection.

Primary and Secondary Data :

The data, which are collected by an investigator or agency for the first time and are thus original in character, are termed as primary data. Or which have already been collected and processed by some

The distinction between primary and secondary data is a matt only. In general, the data are primary to the source who collects and processes time and are secondary for all other sources who later use such data. Primary data are shape of raw materials to which statistical methods arc applied for the purpose of analysis and interpretation. Secondary data are usually in the shapo of finished products since they have been treated statistically in some form or the other. After statistical treatment the primary data lose their original shape and become secondary data.

Choice Between Primary and Secondary Data :

At the outset, an investigator has to decide whether he will use primary data or secondary data. The following factors must be taken into consideration for making the choice between the two methods of collection of data :

(i) Nature, object and scope of enquiry.

(ii) Availability of finance.

(iii) Availability of time.

(iv) Degree of accuracy desired.

Now-a-days, in a large number of statistical enquiries, secondary data re generally used because reliable published data on a large number of diverse fields are now available in the publications of the govemment, private organizations, research institutions, international agencies, periodicals and magazines etc. Primary data re generally collected only if their do not exist any secondary data suited to the investigation under study. In some of the investigations, both primary and secondary data may be used.

Methods of Collecting Primary Data :

The methods commonly used for the collection of Primary data are as follows :

(i) Direct personal investigation.

(ii) Indirect oral interviews.

(iii) Information form local sources and correspondents.

(iv) Mailed questionnaire method.

(v) Schedules sent through enumerators.

original in character, are termed as primary data. On the other hand, secondary data are those, which have already been collected and processed by some agency or person.

The distinction between primary and secondary data is a matter of degree or relativity only. In general, the data are primary to the source who collects and processes them for the first time and are secondary for all other sources who later use such data. Primary data are in the shape of raw materials to which statistical methods are applied for the purpose of analysis and interpretation. Secondary data are usually in the shape of finished products since they have been treated statistically in some form or the other. After statistical treatment the primary data lose their original shape and become secondary data.

Choice Between Primary and Secondary Data :

At the outset, an investigator has to decide whether he will use primary data or secondary data. The following factors must be taken into consideration for making the choice between the two methods of collection of data :

(i) Nature, object and scope of enquiry.

(ii) Availability of finance.

(iii) Availability of time.

(iv) Degree of accuracy desired.

Now-a-days, in a large number of statistical enquiries, secondary data re generally used because reliable published data on a large number of diverse fields are now available in the publications of the government, private organizations, research institutions, international agencies, periodicals and magazines etc. Primary data re generally collected only if their do not exist any secondary data suited to the investigation under study. In some of the investigations, both primary and secondary data may be used.

Methods of Collecting Primary Data :

The methods commonly used for the collection of Primary data are as follows :

(i) Direct personal investigation.

(ii) Indirect oral interviews.

(iii) Information form local sources and correspondents.

(iv) Mailed questionnaire method.

(v) Schedules sent through enumerators.

The above methods have been discussed briefly.

(i) Direct Personal Investigation :

In this method the investigator has to collect the information personally from the sources concerned. The investigator establishes personal contact with the information's and conducts on the spot enquiry. He interviews personally everyone who is in a position to supply the information he requires.

This method is suitable when the field of inquiry is limited or nature of inquiry is confidential and when maximum degree of accuracy is desired. It is suitable for intensive investigation. More skill and tact is required for collection data by this method.

Merits :

The advantages of director personal interviews are :

(a) Information is more reliable and accurate.

(b) Response is more encouraging as the investigator approaches personally.

(c) The investigator can extract proper information from the respondents by talking to them all their educational level and if need be, ask them questions in their language of communication.

(d) The investigator can handle delicate subjects effectively by his skill, intelligence and insight to collect proper information.

Demerits :

The following are the limitations of this method.

(a) This type of investigation is not suitable for extensive enquiries.

(b) It may be very costly if the respondents are spread over a wide area.

(c) More time is required for collecting information.

(d) The chances of personal prejudice and bias are greater under this method.

(e) Eh success of this method largely depends upon the intelligence, skill, tact, insight and diplomacy of the investigator.

(ii) Indirect Oral Interview :

When the direct personal investigation is not practicable either because of the unwillingness

Demerits :

(a) The data collected by this may not be reliable because it is subjected to the biasness of the local agents.

(b) This method can be applied if high degree of accuracy is not desired.

(v) **Mailed Questionnaire Method :**

In this method, a questionnaire is prepared, which contains a list of questions relating to the field of enquiry and providing space for the answers to be filled by the respondent. This questionnaire is mailed to the respondents with a request for quick response, which the specified time. A very polite covering note is attached to the questionnaire within contains the aims and objectives of collecting the information and also the definitions of various terms and concepts used in the questionnaire. An assurance is usually given to the respondents that the answers given by them will be kept strictly confidential.

The success of this method depends upon the skill with the questionnaire is drafted and the degree of response from the informants. This methods is usually followed by private individuals researches and some times the Government also.

Merits :

(a) Under this method, it is possible to cover vast areas where informants are spread over.

(b) This methods is by far the most economical method in terms of money, time and manpower provided the respondent supply the information in time.

(c) Under this method, original data are collected, which is free from the biasness of the investigator.

Demerits :

(a) This method cannot be used amongst illiterate population.

(b) The informants may not return the questionnaire at all or may not return in time, and even if they return, some returns may be improperly and inaccurately filled.

(c) The method is not flexible. In case of any inadequate or incomplete answers it is difficult to obtain supplementary or correcting information. It is also difficult to verify the accuracy of the answers given.

To achieve fairly good result from this method, the following points should be taken into consideration.

(a) The questionnaire should not be unduly long otherwise the in formats will simply ignore it.

(b) The questions should be short and free from ambiguity.

(c) The number of respondents selected should be fairly large to guard against possible non responses.

(d) Prepaid postage stamps for the return post should be affixed.

(e) This methods should be preferred in enquiries where the respondents are under legal obligation to fill in the questionnaire and return the same to the investigating authority.

(v) **Schedules Sent Through Enumerators :**

In this method the enumerators go to the informants along with the schedules or questionnaires and help them in recording their answers. The enumerators explain the aims and objects of the investigation to the informants and also emphasizes the necessity and usefulness of correct answers. In this method, the selection of enumerators is a very important task and should be carefully done. They must have clear idea about the purpose and subject of enquiry. Specific training and instructions are given to the enumerators so that they can tackle persons of different nature and can extract correct information from them.

This method is the most common method being employed by all research organizations. Information received under this method is highly reliable.

Merits :

(a) This method can be adopted in those cases where informants are illiterate.

(b) The data collected by this method is more accurate and reliable. Accuracy of the answers can be checked by supplementary questions wherever necessary.

(c) The problem of non-response on the part of the informants in minimized.

Demerits :

(a) This method is very costly as enumerators are generally paid persons.

(b) The success of this method depends upon the personal qualities, unbiaseu attitude, courtesy, and tact of the enumerators.

(c) The quality of data collected depends to a great extent on the efficiency and wisdom with which the schedule is prepared or drafted.

COLLECTION OF SECONDARY DATA

We know that secondary data are those which have already breed collected and analysed by some person or agency. So the problems associated with the original collection of data do not arise here.

The main sources of secondary data may be broadly classified into the following two categories :

(i) Published sources, and

(ii) Unpublished sources.

Published Sources

Governmental and non-governmental organizations publish statistics on different subjects. Such data are very useful and reliable for research purposes. The chief sources of published statistics are :

(a) Governmental Publications

Various ministries and departments of the Union and the State Governments publish statistics on a variety of subjects like population, national income, agriculture, import and export, trade and industries, etc. Some of the main publications are : Statistical Abstract of India (annual), Monthly Abstract of Statistics, agricultural Statistics of India, Annual survey of Industries, Indian Trade Journal, Labour Gazette etc. Besides these, the reports of many special enquiry committees set up by government are also published.

(b) Publications and Reports of Trade Associations, Chambers of Commerce and Financial Institutions :

Various commercial and trade associations like Sugar Mills Association, Indian Cotton Mills Federation, Bombay Mill Owners Association, Jute Mills Association, sugar Mills Association and Federation of Indian Chambers of Commerce etc., publish statistics regarding trade and commerce. Stock exchanges, trade unions, banks and other financial institutions also collect and publish statistics regularly.

(c) Journal and Newspapers :

Statistics of number of important current socio-economic problems can be obtained from newspapers and periodicals. Newspapers like 'Economic Times', 'Business Standard' and 'Financial Express' and many periodicals like 'Commerce', 'Capital', 'Eastern Economist', 'Journal of Industry and Trade' etc give considerable amount of statistics in their issues.

(d) Research Work Done by Scholars in the Universities and Institutions :

Individual research scholars the different departments in the various universities of India and various research organizations and institutes like Indian Statistical Institute, Calcutta and

Delhi, Indian Council of Agricultural Research, New Delhi, N.C.E.R.T., New Delhi, National Council of Applied Economic Research, New Delhi etc. publish the findings of their research programmes in the form of research papers, monographs or journals. These are a constant source of secondary data on the subjects concerned.

(e) Publications of International Bodies :

Official publications of various International Organisations such, as U.N.O, I.M.F., World Bank, I.L.O., W.H.O etc. provide valuable statistical information on a variety of important economic and current topics.

Unpublished Sources :

The sources of unpublished data are varied and such statistics may be found with scholars and research workers, trade associations, chambers of commerce, labour bureaus etc. The records maintained by private firms and by various departments and offices of the government are also the sources of secondary data.

Advantages of Secondary Data :

(a) It is much cheaper to use Secondary Data.

(b) The time required for collection of secondary data is much less as compared to primary data.

(c) In some subjects, the collection of primary data is not possible on the part of a person or agency. For example, census data cannot be collected by an individual or organization, it can only be obtained from Government publications.

Limitations of Secondary Data :

(a) It is difficult to find out suitable and adequate secondary data for the purpose of investigation.

(b) The other problem is finding secondary data, which are sufficiently accurate and reliable.

After the collection of data from primary or secondary source, the next step in a statistical enquiry is to edit the data, i.e. to scrutinize the data. The main purpose of editing is to detect possible errors and irregularities. Editing requires great care and attention.

Editing Primary Data :

While editing primary data, the following points should be given due attention.

(i) The editor should see that each questionnaire and schedule is complete in all respect. If some questions have not been answered and those questions are of vital importance the informants may be contacted again to get information.

(ii) The Editor should verify that the answers to questions are not contradictory in nature. This is called editing the data for consistency.

(iii) The editor should see that the information provided by the informant is correct in all respect. This is one of the most difficult tasks of editor.

(iv) The editor should check that the information supplied by the various people is homogenours and uniform.

Editing and Scrutiny of Secondary Data :

Secondary data must be used with caution, as it is very difficult to verify such data and to edit them to find out inconsistencies, errors and omissions. Secondary data should be scrutinize to verify its suitability, accuracy and adequacy. Hence before using such data, the investigator should consider the following aspect.

(a) The Reliability of Data :

To establish the reliability of secondary data, we should satisfy ourselves about :

(i) The methods used for the collecting organization.

(ii) The reliability of source of information, and

(iii) The methods used for the collection and analysis of the data.

It should be verified that the data relates to normal period free from abnormal happenings and the collecting agency was unbiased. If the data were collected on the basis to sample, it should be ascertained that the sample was representative and the sampling procedure was unbiased.

(b) The Suitability of Data :

Before using secondary data, the investigator must satisfy himself that the data available are suitable for the purpose of investigation. The suitability of data can be judged by comparing the objectives, nature and scope of the present enquiry with the original investigation. For example, if the purpose of the present enquiry is to study the trend in wholesale prices, and data provide only retail prices, such data will be unsuitable.

(c) Adequacy of Data :

Adequacy of data is to be judged in the light of the geographical area covered by the original enquiry and the expected coverage of the present enquiry. If the original data refer to an area, which is wider or narrower than the area of the present enquiry, the should not be used.

SELF-TEST – 3

1. What do you mean by a Statistical enquiry ? Explain the process of planning of statistical enquiry.

2. Distinguish between primary and secondary data. What factors are taken into consideration for making the choice between primary and secondary data ?

3. Discuss briefly the various methods that are used in the collection of primary data. Point out their merits and demerits.

4. Discuss the 'Questionnaire' method of collection of primary data.

5. What are the main sources for collection of secondary data ? Explain the advantages and disadvantages of secondary data.

6. What is editing of data ? What factors are given due attention by the auditor while editing primary and secondary data?

7. Write short notes on :

(i) Statistical Unit.

(ii) Primary Data.

(iii) Secondary Data.

(iv) Mailed Questionnaire Method.

(v) Published sources of Secondary Data.

(vi) Scrutiny of Secondary Data.

CENSUS AND SAMPLE INVESTIGATION

Objectives of the Lesson :

When you complete this lesson you will be able to know the following.

- Meaning of Universe or Population.
- Merits and Demerits of Census Investigation.
- Details of Sample Investigation.
- Various Random Sampling Methods.
- Different non-random Sampling Methods.
- Errors in Sampling.

INTRODUCTION

Collection of statistical data is required for the purpose of analysis and interpretation. There are two ways in which the required information may be collected.

(i) Complete enumeration or census method in which every item of the population is studied, and.

(ii) Partial enumeration or sample method in which a part of the universe is studied.

"The universe or population consists of the total collection of items and **elements** that fall within the scope of a statistical investigation". It is the collectivity or totality of objects under consideration. A population containing a finite number of objects or items is known as finite population e.g., students of Correspondence Courses of Utkal University, population of state, production of a factory during a day etc. On the other hand a population having an infinite number of objects is termed as an **infinite population**.

Census Investigation

In census investigation, information is collected from every unit of the universe relating to the problem under investigation. Population census, which is carried on once in every ten years, is an example of census study. Money, material, time, personnel and effort required for carrying out a complete enumeration are bound to be very large.

Merits

The following are the merits of 'Census Investigation'.

(i) The results are more accurate and reliable since the information is collected from each item of the population.

(ii) In a census investigation intensive information is obtained from each and every item, thus many facts relating to the problem are brought to light. For example, population census gathers information not only about the magnitude of population; it also collects information on various other points, such as, age, marital status, education etc.

(iii) It is an appropriate method, where units of diverse characteristics constitute the universe.

Demerits

(i) Census method of investigation is a costly affair. More labour and time is also required in this method.

(ii) It is not possible to undertake census enquiry where the universe is infinite or complex. It is difficult to contact every item of the universe. This method of enquiry can not be adopted.

Sample Investigation

A few units selected from the universe in accordance with some specified procedure constitute a sample. Under sample investigation information is collected from some representative units (sample), which are selected from the universe. In our day-to-day life we also use the sampling technique frequently. In the market, we examine a sample of wheat or rice and form an idea about the quality of the total lot and then decide whether the quality is acceptable or not.

In many instances, sampling, which is the process of learning about the population on the basis of a sample drawn from it, constitute the only possible and practicable method to obtain desired information. If the universe is infinite or very large then only a sample study is possible.

Merits

(i) In sample survey, the cost involved is much loss as compared to census enquiry.

(ii) Sample enumeration requires much less time because the colume of data collected and processes is less. If the results of the survey are urgently required, this method is most appropriate.

(iii) With small data, it is possible to have a detailed enquiry.

(iv) Sample survey is the only appropriate method in many circumstances. If the universe is infinite or too large over a large geographical area, it is difficult to collect information from each unit. This method is most suitable in opinion surveys, quality control, etc.

Demerits

(i) If the sample is not representative of the universe, the results will not be accurate, If the size of the sample is inadequate it may tail to indicate the true characteristics of he population.

(ii) If the items of the sample are not selected without any bias, the conclusions derived from the study may not be correct.

(iii) Sample investigation is a special technique and requires specialized knowledge and skill. If the investigator does not possess that he may draw wrong conclusions.

Methods of Sampling

The purpose of sampling is to study a part of the universe to draw inference about the whole universe. So the sampling process involves the selection of the sample, collecting the information from the sample and making an inference about he population.

There are various methods of sampling that may be used singly or along with others. The choice of the methods will be determined by the purpose for which sampling is sought and the nature of the population. Some of the important methods that are popularly used in practice are given below :

1. *Random Sampling Methods*

(a) Unrestricted Random Sampling.

(b) Restricted Random Sampling.

(c) Systematic Sampling.

(d) Stratified Sampling.

(e) Multi-stage Sampling

2. *Non-Random Sampling Methods*

(a) Judgment Sampling.

(b) Convenience Sampling.

(c) Quota Sampling.

Random Sampling Methods

(a) Unrestricted or Simple Random Sampling :

A random sample is a sample selected in such a way that every item in the population has an equal chance of being included in the sample. The selection of the item is free from personal bias of the investigator. Simple random sampling refers the sampling technique in which each and every item of the population has an equal chance of being included in the sample. This random selection process is not unsystematic or haphazard process.

In practice it not easy to ensure true randomness in the selection fo items of a sample. However, to ensure randomness, the statisticians make use of certain methods like lottery method or consult table of random numbers.

Lottery Method :

This is the simplest and the most popular method of obtaining a random sample. In this method, all items of the universe are numbered on small and identical slops of paper, which are folded and mixed together in a container or drum. A blindfold selection in the made of the number of slips required to constitute the desired size of sample.

For example, if we want to select 20 candidates out of 200, randomly we assign the numbers 1 to 200, one number to each candidate, write these numbers on separate slips of paper, fold these slips, mix them thoroughly and then make a blindfold selection of 20 slips.

Table of Random Numbers :

When the size of the population increases the lottery method becomes time-consuming and cumbersome to use. The table of random numbers is an alternative method of random selection. "The random numbers are generally obtained by some mechanism, which, when repeated a large number of times ensures approximately equal frequencies for the numbers from 0 to 9 and also proper frequencies for various combinations of numbers (such as 00... 99, 000, 001; .....999 etc.) that could be expected in a random sequence of the digits 0 to 9. Several standard tables of random numbers are available, like Tippet' Random Number Table, Fisher and Yates table of random numbers, Kendall and B. Smith table of random numbers, Rand Corporation table of random numbers, etc. Any page fo the random number table is selected at random and the numbers in any row, column or diagonal is picked up at random. For example if we have to select 20 items out of 2,000, the procedure is to number all the 2,000 items from 1 to 2000. A page from the random number table maybe selected tat random and the first 20 numbers upto 2,000 are noted down. Items bearing those numbers will be included in the sample.

Merits of Unrestricted Random Sampling

(i) It is a scientific method of taking out a sample from the universe. There is little possibility of personal bias.

(ii) As the size of the sample increase, it becomes more representative of the population.

(iii) This method provides the reliable information at the least possible cost. Thus it saves time, money and labour in investigating a problem.

Demerits

(i) If the sample size is small, it may not be representative of the population.

(ii) When only a portion of the population data re accessible, random sampling is not possible.

(iii) From the point of view of field survey if the units of the universe are spread over a large area, the units selected at random tend to be widely dispersed. So the time and cost of collecting data becomes too large.

(c) Restricted Random Sampling

Restricted random sampling may assume the following forms.

(i) Systematic Random Sampling :

Systematic or quasi-random sampling method is generally used in those cases where a complete list of the population from which the sample is to be drawn is available.

Under this method, the process of selecting the sample is to pick up every Kth item of the population, where 'K' refers to the sampling interval.

$$K = \text{Sampling Interval} = \frac{\text{Size of the universe}}{\text{Size of the sample}}$$

If we want to select 200 units from 2000 units, then we have to pick up every 10th item. The process is the first item between the first and the Kth is selected at random. Suppose in our example if it is seven, we shall go on adding 10 and obtain numbers of desired sample. So the units of the sample are : 7th, 17th, 27th, 37th, ...etc.

Merits

(i) The systematic random sampling is simple and convenient to adopt.

(ii) The time and effort involved in this method are relatively smaller.

(iii) It is suitable and gives satisfactory result if there is no unique variation in the universe.

Demerits

(i) If there is some periodic features and variations present in the population, the sample collected by adopting this method becomes less representative.

(ii) This method is not necessarily a very scientific method.

(ii) Stratified Random Sampling

This method of sampling is generally used when a population is heterogeneous and is composed of different segments or strata. In this method, the population is classified into a certain number of groups called strata and then selecting random samples independently from each group or stratum. The divisiōn fo the population into strata or groups is done according to some relevant characteristics, so that there is greater homogeneity within each stratum. In practice geographical, sociological and economic characteristics are often used for stratification of the universe.

Stratified random sampling is widely used in market research and opinion pools, for it is fairly easy to classify people into occupational, economic, social, religious and other strata. The number of units selected from each stratum should be proportional to the number of units in that stratum in the population.

Merits

(i) The sample selected under this method is more representative of the population as it ensures a desired representation to various strata in the population.

(ii) Stratified sampling ensures greater accuracy as units are selected at random from each containing homogeneous units.

(iii) Stratification is appropriate when the original universe is badly skewed.

Demerits

(i) It is a very difficult task to divide the population into various homogeneous strata. Stratified sampling is not possible unless some information concerning the population and its strata is available.

(ii) If different strata of a population overlap, such a sample would not be representative.

(iii) Multistoried Sampling or Cluster Sampling

This method refers to a sampling procedure, which is carried out at several stages. At first the first stage units are selected randomly by some suitable method. Then a sample of second stage units is selected from each of the selected first stage units by some suitable methods. Further stages may be added as required.

For example, if it is decided to select a sample of 2000 households from the State of Orissa to study their income and consumption pattern, the first step is to select a few districts of random. At the second stage, each district selected may be sub-divided into a number of villages and a sample of villages may be taken at random. At the third stage, a number of households may be selected from each of the villages selected at the second stage.

Merits

(i) This method of sampling in very helpful in many large-scale surveys where the preparation of the list of all the units in the universe is difficult.

(ii) It permits the field work to be concentrated and yet large area to be covered.

Demerits

(i) A multi-stage sample is generally less accurate than a sample containing the same number of final stage units, which have been selected by some suitable single stage process.

Non-Random Sampling Methods

(a) Judgment Sampling

Judgment sampling refers to the selection of items for the sample by the investigator exercising his judgment. Thus the choice of the sample items depends exclusively on the discretion of the investigator. Judgment sampling is also called "purposive sampling" and "deliberate sampling". Such a method requires a deep and thorough knowledge of the universe and a good deal of experience of the part of the investigator.

The judgment sampling is suitable where the number of items in the universe is small and also the size of the sample o be drawn is small. It si suitable when some known characteristics of the universe are to be intensively studies.

Merits

(i) If the person handling the operations have full knowledge of the composition of the universe, Judgment sampling can ensure proper representation of various sections of the universe.

(ii) This method is often used in solving many economic and business problems when only a small number of sampling units is in the universe.

Demerits

(i) This method is not scientific because the personal bias of the investigator may affect the results.

(ii) In selecting items, inclination becomes more important than judgment. The investigator selects those items for the sample, which conform to his preconceived notions.

(iii) The results obtained by this method cannot be compared to the results of other sampling studies. Since an element of subjectiveness is possible, this method cannot be recommended for general use.

(b) Convenience Sampling

This is an easy way of selecting items. A fraction of the population is selected at convenience. A sample obtained from readily available lists such as stock exchange directory, telephone directory, automobile registration records etc. is a convenience sample. If a researcher has to study the financial performance of public sector and he takes only the units, which are close to his town, he is following convenience-sampling method.

This method is suitable when the universe is not clearly defined and a complete source of list is not available.

The results obtained by this method may not be representative of the population. They are generally biased and unsatisfactory.

(c) Quota Sampling

In this method interviewers are given quotas of persons in different age groups, different regions, different social classes etc.; and are then instructed to obtain the required number of interviewers to ill each quota. The quotas ensure that the total sample included approximately the right proportion of persons of various categories. The interviewers complete the quota assigned to them by supplementing new respondents in place of those not available or are in co-operative.

This method is commonly used in making surveys of public opinion. If the interviewers are trained and they follow the instructions closely, this method provides satisfactory results.

This method is subject to various of errors and bias. The results obtained by this method cannot be checked for accuracy.

Errors in Sampling

The error arising due to drawing inferences on the basis of sampling is termed as sampling error. Sampling errors are of two types :

(a) Biased errors

(b) Unbiased errors

Biased Errors :

These errors arise due to biasness in selection of units from the population and estimating by applying different statistical methods. Thus these errors creep in because of the faulty selection of the sample and faulty collection of data. In addition to these, faulty methods of analysis may also introduce bias.

Unbiased Errors :

These errors arise due to chance. If the error is due to chance, increasing the sample size will usually lead to more accurate results. These errors are also called 'random errors'.

Reducing Sampling Errors :

Sampling errors should be reduced to the minimum so as to attain the desired accuracy. Personal bias can be avoided to a great extent by selecting the sample units randomly. The sampling error usually decreases with increase in sample size. Editing of data is done to eliminate some obvious errors reflecting the bias of the respondents. Appropriate methods of analysis should be adopted to avoid bias in interpretation.

SELF-TEST – 4

1. Distinguish between a census and sample investigation. Explain the conditions under which each of these methods will be suitable.

2. Discuss census enquiry with its merits and limitations.

3. What is sampling ? Explain the reasons for increasing popularity of sampling methods?

4. What is random sampling ? Discuss briefly the methods of random sampling.

5. Distinguish between restricted random sampling and unrestricted random sampling. What are their merits and demerits ?

6. Discuss about the non-random sampling methods. State the circumstances when these methods are suitable.

7. What are the sampling errors ? How can they be reduced ?

8. Write short notes on :

(a) Stratified Sampling.

(b) Cluster Sampling.

(c) Quota Sampling.

(d) Sampling Errors.

CLASSIFICATIO OF DATA

Need and Meaning :

The data, which are collected, are quite large in quantity and these are not as such comparable. They are not fit for analysis and interpretation unless they are condensed. Mostly these mass data are without any form or structure. Such data might have been collected relating to some common features or people. But still there may be many types of dissimilarities even within the same group of persons. Fort he purpose of analysis and interpretation, data have to be divided into homogenous groups. The need for classification arises to remove two defects – (1) Volume and (2) Heterogeneity of data, and to bring them to some form or structure so that some analysis out of it becomes possible. This is one of the objectives of the science of statistics.

Before the data are tabulated it is necessary to arrange them in similar groups, so that at the time of tabulating them there will be no difficulty. As a rule the first step in the analysis is to classify and tabulate the data collected. If published data have been utilized, one has to rearrange them into new group and tabulate the new arrangement so that data will serve the purpose of enquiry.

Classification defined. "The process of arranging data in groups or classes according to resemblances and similarities is technically called classification.

Classification is the groupings of related facts into class. Facts in one class differ from those of another class with regard to some characteristic. On the basis of such facts, specific data are classified. Without reorganizing the data collected there can be no classification. In order to make the data easily understandable, the first task of the statistician is to condense and simplify them in such a manner that irrelevant details are eliminated and their significant features stand out prominently.

Even after classification, statistical data are not fit for comparison and interpretation but this is certainly the first step towards tabulation of data. Classification is, therefore a preliminary to tabulation and it prepares the ground for proper presentation of statistical data. Consider the following example. Students after passing matriculation or equivalent examination apply for admission to college. Such applications relate to I.Sc., I.Com, and I.A. Classes. There may be first divisioners, second divisioners an third divisioners among the applicants in each, some of them may belong to scheduled caste, scheduled tribe, refugees etc. If you want to analyse these particulars systematically in order to have some idea, you have to classify them class-wise, sex-wise (male and female), division-wise (1st division, 2nd division etc), and applicants belonging to S.C. and S.T. etc. The process through, which this information in a summary form is obtained, is called the classification of data.

Characteristics of and Ideal Classification :

There is no hard and fast rule, which can be laid down as the characteristics of an ideal classification. The classification of data in each investigation has to be decided after taking into a. count the nature, scope and purpose of enquiry. But an ideal classification should possess the following characteristics.

(a) **It should be Unambiguous –** Classification aims at removing irrelevant material and simplifying the analysis. Each class or group should, therefore, be so defined that there will be no confusion. But some time it is not possible to define each class vividly. If population is divided into two parts – i.e. literate and illiterate., it is difficult to define who is literate. Anybody who can read and write a simple letter is said to be literate.

(b) If it will change enquiry to enquiry, the data will be fit for comparison. The meaning of simple letter should be used in the some sense from time to time for differentiating literate from illiterate.

(c) **It should be Flexible –** This characteristics is contradictory to the second one as it will appear initially. But it is not so, a good classification should have the capacity of adjustment to new situations and circumstances, otherwise its practical utility will not be there. That which is stable is not necessarily rigid; it is used in a relative sense. With changes in time new classifications have to be included and those, which are not of practical utility, are to be dropped. Major classification usually does not change; only minor classifications are adjusted to situations.

Objectives of Classification –

The main objectives of classifying data are :

1. To condense the mass of data and to put them separately so that the like and unlikes are not mixed together. This enables easy understandability. Millions of figures can thus be arranged in a few classes having common features.

2. To facilitate comparison.

3. To pin point the most significant features of the data at a glance.

4. To give prominence to the important information gathered and to eliminate the unnecessary elements.

5. To enable a statistical treatment of the data collected and complied.

Bases or Types of Classification –

Statistical data are classified on the basis of the characteristics possessed by the different groups of units of a universe. These characteristics may be either descriptive or numerical. Literacy, unemployment, blindness, occupation, sex, religion, colour etc. are examples of descriptive characteristics. These things cannot be expressed in quantitative form or counts. Age, income, height, weight, children, students etc. are examples of numerical characteristics.

Mainly there are two bases or types of classification. When data are classified on the basis of qualities or attributes which are incapable of quantitative measurement, the classification is said to be according to attributes, and when the data are classified on the basis of quantitative measurement the classification is said to be according to class-intervals

1. Classification According to Attributes -

Some authors feel that classification of this nature be classed as qualitative classification. Under this type, data are classified on the basis of some attributes or quality. All those units in which a particular characteristic or quality or attribute is present, are placed in one group and those in which it is absent are placed in another group. If the attribute 'blindness' is studied we shall have two classes those who are blind and those who are not blind. The point to note in this type of classification is that this attribute cannot be measured. It is not possible to measure the degree of blindness in each case. This type of classification in which only one attribute is studied and the data are divided in two parts is called simple classification or classification according to dichotomy. (Dichotomy means two).

Similarly, we may classify population on the basis of sex, i.e. into males and females, or literacy, i.e. into literates and illiterates, or religion i.e. into Hindu, Non-Hindu. Here we divide the data on the basis of different attributes into different classes. If in addition to 'blindness' the above three attributes are added, a person can be either blind or not blind, a person may either a male or a female, may be either literate or illiterate may be either a Hindu, or a Non-Hindu. Now each of the two classes is capable of further subdivision. Such a classification in which more than one attribute is considered is called manifold classification. An example of manifold classification is given below.

Population

Males Females

Literates Illiterates Literates Illiterates

Hindu Non-Hindu Hindu Non-Hindu Hindu Non-Hindu Hindu Non-Hindu

Emp. Unemp Emp. Unemp Emp. Unemp. Emp Unemp Emp. Unemp.

Emp. Unemp Emp Unemp Emp Unemp

(Please, note that 'emp' indicates Employed and 'unemp' indicates unemployed)

When we classify data on the basis of attributes, sometimes we have to divide data arbitrarily – rather than putting any definite demarcation among them. The differences between the classes are not always natural or very well defined. If for example, population is divided into two groups – tall and short, who are tall and who are short is a matter of opinion and arbitrarily this has to be decided. If we say that any body whose height is more than 160 centimeteres shall be regarded as tall and below it short, such a classification is more definite and precise. But this is not always possible. For example, the difference between a literate and an illiterate is always a matter of opinion. Here one attribute gradually changes into another. It is better to define them and follow them consistently.

2. Quantitative Classification or Classification According to Class Intervals-

This type of classification is applicable to such cases where direct quantitative measurement of data is possible. Data relating to height, weight, income age, children, production, consumption etc. come under this category. In such cases data are classified on the basis of values or quantities. If for example, the heights of students of a college are expressed, they may be as follows – 150 centimeters to 155 centimeters,

155 – 160 to 165 centimeters to 170 centimeters;

160 centimeters 170 to 175 centimeters etc. If in respect of the above five groups there are 150, 115=95.120 and 92 students, the data can be classified as follows :

Here the data are divided into a number of classes, each of which 150 to 155 centimeters is called a class interval. The span of a class, i.e. the difference between the upper limit and lower limit is known as class interval. The size of the class-interval is determined by the number of classes and the total range in the data. The limits within, which a class-intervals lies are called class-intervals. The class-limits are the lowest and the highest values that can be included in the class. 150 is the lower limit, and 155 centimeter is the upper limit of the first class-interval in the example mentioned above. The difference between two class limits is termed as class magnitude or magnitude of the class-interval. (5 centimeters in the example). The numbers of items, which fall in any class interval, are called class-frequency. T he class-frequency of the class-interval 150-155 centimeter is 150. This means that 150 students have a height between 150 to 155 centimeters. This is the number of observations corresponding to that particular class interval.

Variable heights in centimeters	No.of students (Known as frequency)
150 -155 centimeters	150
155 -160	115
160 -165	96
165 -170	120
170 -175	92
Total	572

The quantitative characteristics relate to variables such as height, weight, exports, imports, wages etc., which can be numerically expressed. These are called 'quantitative variables' or simply 'variables'. By a variable it is meant 'a quantity, which assumes different values that may be measured in some approximate unit. The term variable's refers to the characteristic that varies in amount or magnitude in a frequency distribution. A variable may be (i) Continuous or (ii) Discrete. If a variable can take any numerical value within a certain range it is called continuous variable. A variable is said to be continuous when it may pass from one value to the next by infinitely small gradation. For example, in measuring the height of colleges students it is possible to come across any measure between 165 to 170 centimeters, there may be a student whose height is 165.2034cm. The point to note here is that any conceivable height may occur within this range.

Variables, which take only discrete or exact values, are called discrete variables. Members of family, children, scores in cricket match, rooms in houses etc. are discrete variables. They cannot be in fractions. Thus a variable is said to be discrete when there are gaps between one value and the next. If in a class, there were 48 students and one student is admitted the count leaps from 48 to 49 without passing through fractional value. It is not slightly above 48 or a little under 49. The count is exactly 49. The number of rooms in a house can only take certain values such as 2,3 or 4 etc. Similarly the number of ex-employees and number of machines in an establishment are disc.ete variables. Generally speaking continuous data are obtained through measurements, while discrete or discontinuous data are derived by counting.

Classification according to class intervals involves three basic problems. They are :-

(a) Number of classes and their magnitude.

(b) Choice of class limits.

(c) Counting the number in each class.

(a) Number of Classes and their Magnitude :

The number of classes should preferably between 5 and 20. It should not contain more than 20 to 25 classes, depending upon the total number of items in the series. One of the important rules to be followed is that all class intervals should have a fairly good frequency. Even though there is no rigidity, the number of classes should not be less than five because in that case, the classification may not reveal the essential characteristics. If the number of items is less, the number of class should also be less because otherwise there will be no frequency in some classes.

A balance should be struck between too many and too small number of classes. An ideal number of classes for any frequency distribution would be that, which gives that maximum information in the clearest fashion. However, the choice of number of classes basically depends upon : -

(i) The number of figures to be classified.

(ii) The magnitude of the figures.

(iii) The details required, and

(iv) Ease of calculation of further statistical work.

Magnitude of Intervals :

The magnitude of class intervals depends on the range (difference between the maximum and minimum values) of the data and the number of classes. If you have to measure the heights of 100 persons, and you desire to have 10 classes the magnitude shall be 10 centimeters. The magnitude should be such that it does not eliminate or distort the important characteristics of data. The magnitude of the class interval should be 2,5,10,25,50 etc. (multiples of 5) rather than odd figures like 1,3,7,11, 23 etc. in common. The multiples of 2,5 and 10 are use and human mind considers them almost as natural magnitudes.

In general the class intervals should be of equal magnitude. If the size of the class interval is unequal it may give a misleading impression and in such cases, comparison of one class with the other may not be possible.

The starting point, i.e. the lower limit of the first class should either be zero or multiple of 5. For example, if the lowest value of the data is 64 or 63 and you have taken a class-interval or 10 the first class should be 60-70 in place of from 64 to 74 or 63 to 73. Similarly, if the lowest value of the data is 16 and the class internal is 5 then the first class should be 15 to 20 and not 16 to 21.

(b) Class Limits :

The most important thing that should be kept in mind while choosing the class limits is that these should be chosen in such a nanner that the mid-point of a class interval and the actual average of items of that class interval should be as close to each other as possible. The class limits must be such that midpoint of class intervals are familiar and common figures ending with 0,2,5,10 etc. These are capable of easy and simple analysis. The starting point, i.e. the lower limit of the first class should be determined in a manner that frequencies of each class get concentrated near the middle of the class-interval. This is necessary because in the interpretation of a frequency table and in subsequent calculations based upon it, the mid point of each class is taken to represent the value of all items included in the frequency of that class.

Class limits should be definite and clearly stated. It should not be indeterminate in nature. If a class is stated as below 10 or 70 and above one of the two limits in each one is not definite. Such classes are called 'open-end' classes and should avoided.

The class limits may be written in any of the following ways :

I	II	III	IV	V
0-10	0 and under 10	0-9	Below 40	5
10-20	10 - 20	10-19	Below 30	15
20-30	20-30	20-29	Below 20	25
30-40	30-40	30-39	Below 10	35

In the first method, items whose values are just 10 or 20 can be classified either 0-10 group and 10-20 respectively or in 10-20 and 20-30 classes respectively. Usually the item is classified in the next higher class so that the item whose value is exactly 10 would come in 10-20 group.

In the second method this point is clear. Items whose values are less than 10 would be in 0-10 class interval. This is the exclusive method of classification. When he items equal to the size of either the lower limit or the upper limit are excluded from the frequency of that class it is known as 'exclusive' method. Usually in exclusive method the items whose values are equal to the upper limit of a class are grouped in the next higher class. Items with values less than the upper limit are taken into account in that class. Exclusive method ensures community of data.

The third method is inclusive method. When items having values equal to the lower and the upper limits of a class are included in the frequency of that very class, the classes are known as inclusive. In practice this methods is like the second method as 0-9 means 0 and under '10'. To emphasize this point sometimes the class interval is written as 0-9.99.

The fourth method indicates that below 40 means 0-39 or 0-39.99 below; 20 means 0-19 or 0-19.99 etc. Such type of classification arises in a series of cumulative frequency, which you will learn later.

The fifth method of classification indicates the mid-values only. It is the class mid-point, It is the value lying half-way between the lower and upper class limits of a class interval. Mid-point of a class is ascertained as follows :

$$\text{Mid point of a class} = \frac{\text{Upper limit of the class + lower limit of the class}}{2}$$

For the purpose of further calculations the mid-point of each class is taken to represent that class.

All the types of class limits can be conveniently grouped under the exclusive or the inclusive method of classification. One of the fundamental principles to be followed while adopting the specific method of classification is that in case of continuous variables the upper limit exclusive method must be used because this method ensures continuity. The inclusive method should in general be used in case of discrete variables.

If the class intervals do not have continuity i.e. inclusive method of classification is followed, it is necessary to make an adjustment do determine the correct class interval and to have continuity. Please note the following carefully.

Weekly Wages in Rs.	No. of Workers.
10-19	5
20-29	10
30-39	15
40-49	22
50-59	7

To adjust the class limits (to bring continuity). We take here the difference between 20 and 19, which is 1. By dividing it by two we get .5. Tis .5 is called the correction factor. We deduct.5 from the lower limits of all classes and add.5 to upper limits of all classes. Then the adjusted class intervals will be as follows :

Weekly Wages in Rs.	No. of Workers.
9.5-19.5	5
19.5-29.5	10
29.5-39.5	15
39.5-49.4	22
49.5-59.5	7

Now please mark the formula for adjustment.

$$\text{Correction factor} = \frac{\text{Lower limit of the } 2^{nd} \text{ class} - \text{Upper limit of the } 1^{st} \text{ class}}{2}$$

(.5 in the above example)

One interesting thing is to be noted –

Before adjustment the class interval was 9 but after adjustment it is 10.

(c) **Counting the Number of Items in Each Class -**

The next thing is to count the number of items falling under each class. This has to be done in discrete as well as continuous variable. The method that is adopted for the purpose is known as constructing a frequency table. Whenever a variable has been measured in a group of items, the classification following manner.

Array – This means the data are to be arranged in ascending (increase in value) or descending (decrease in value) order. This will help us in knowing the range of data. Then data are to be condensed. A first step in such soncensation would be achieved by representing the repetitions of the items by 'tallies' instead of requiting the items. The number of tallies corresponding to any given item is the frequency of that item. 'Frequency' usually represented by 'f' thus means the number of times a certain value of the variables is represented in the given data.

The number of items or 'frequency' in each class can be counted by the following methods.

(i) **By Tally Sheets -** Under this method, the class intervals are writter on a sheet a paper (called Tally Sheet) and for each item a stock (1) is marked against the class interval against which it falls. Usually after every strokes (or vertical bars) in a class the fifth items is indicated by drawing a horizontal or diagonal line over or through the strokes. To facilitate counting blocks of five bars are prepared and some space is left between each block. Finally the number of bars corresponding to each value of the variables are counted and placed in a column known as 'frequency' (f) The process shall be clear from the following example.

Number of marks obtained by 50 students

Marks	Tally Sheet or Tallies	F(frequency or Total)
0-10	IIII...	4
10-20	I	11
20-30	III	3
30-40		5
40-50	I	16
50-60	III	8
60-70	III	3
Total		50

This methods of classification helps in condensing data only where values are largely repeated, otherwise it does not serve the purpose of condensation. In actual practice, this method is rarely used.

(i) **By Mechanical Aids –** Various types of machines are now used for sorting and listing data. Some of these machines are hand operated while others are operated by electricity. Needle sorting (had operated) has become very popular these days. Large number of items can be sorted with it under any number of heading and sub-headings. Cards of convenient size and shape with a series of holes are used in this method. Each hole stands for a value and when cards are stacked, a needle passes through particular hole representing a particular value. These cards are later, separated and counted. In this way frequency of various classes can be found out by the repetition of this technique.

This technique of punched cards is also equally popular. In this method the data are recorded on special cards by punched holes made by means of a special key punch which can be operated either by hand or electricity. Hollerity and I.T.C. sorting machines sort the cares at a speed of about 24,000 per hour. Thus you see that mechanical aids have made the work of classification very easy, quick and accurate.

3. Geographical Classification :

This method of classification is also known as area-wise classification. Even though this classification relates to quantitative phenomena, some authors prefer to have it as a separate type of classification.

In this type of classification, data are classified on the basis of geographical or area or locational differences between the various items. For example, the production of wheat or rice in India may be presented state-wise, or the same relating to Orissa may be presented district wise, or drought situation in Orissa my be presented district, sub-division and Panchanyatwise.

Such classification are usually listed in alphabetical order for easy reference. It may also be presented by size or on the basis of the ranks or importance. For example, in wheat production of India U.P. will be shown first as its production is the highest. Normally in reference table the alphabetical order or arrangement is followed whereas in summary tables arrangeme is made by the size or the ranks, (About the types of Table you will known in the next lesson).

4. Chronological Classification :

When a data are arranged over a period of time the type of classification is known as chronological classification. You may present the phenomena like population, production, import, export, sales etc. on the basis of time (i.e. year or month). Time series in which one variable is always time (year, month) are usually presented in chronological order and usually it starts with the earliest period. If emphasis has to be given on the recent events, a reverse time order may be used, like 1978, 1977 and so on.

TABULATION OF DATA

In lesson no.5 it has already been mentioned that data are condensed and put forth in some arrangement to enable comparison and to draw conclusion. Classification and Tabulation of Statistical data are the two stages or steps through which mass data are condensed. Classification proceeds the stage of Tabulation though some authors discuss these two stages in one Chapter. For a clear understanding they are discussed in two separate lessons here.

Definition and Meaning :

The first thing to be considered in tabulation is what is it ? or what does it signify ? Different authors have defined Tabulation of statistical data in different ways. A few definitions are given below :

"A statistical table is the logical listing of related quantitative data in vertical columns and horizontal rows of numbers with sufficient explanatory and qualifying words, phrases and statements in the form of titles headings and notes to make clear the full meaning of data and their origin" – Tuttle.

The definition given by Tuttle is very exhaustive because it contains the features of a good table. In fact it is the most satisfactory and accepted definition on tabulation.

Karmal, another statistician, has defined it as "A table summarizes the data by using columns and raws and entering figures in the body of table". This definition is not exhaustive because it does not include many important features and objects of tabulation.

O.P. Bajpai has defined it as "Tabulation is a process or orderly arrangement of data into series of boxes made up of rows and columns where they can be read in two dimensions". From all these definitions, in the broadest sense, it can be defined as orderly arrangement of data in columns and rows".

Tabulation involves a systematic presentation of data to elucidate the problem under investigation. It is a process between the collection of data and their final analysis. In fact tabulation is meant for properly arranging the answers relating to the questioned raised in any inquiry or investigation, and is very helpful in analysis of the collected data and to draw conclusions from them. Tabulation is the final stage in collection and compilation of data. The main object of a statistical table is to arrange the physical presentation of numerical facts that the attention of the readers is automatically directed to relevant information. By simplifying complex data much time is saved and confusion of mass data is avoided. It facilitates comparison by bringing related items of information close to each other. For this purpose the table is divided and sub-divided into parts and for each part there are totals and sub-totals.

Croxton and Cowden, two important authorities in statistics, have put it as one of the methods of presentation of data. There are five ways of presenting facts – (1) descriptive or textual, (2) semi-textual, (3) tabular statement, (4) diagramtic presentation and (5) graphic presentation of data.

Importance

Tables make it possible for the enumerator to present a huge mass of data in a detailed orderly manner within a minimum of space. Because of this, tabular presentation is the corner-stone of statistical reporting. The role of tabulation or its importance or advantages, in whatever name we may put it, can be described as follows : -

1. It simplifies complex data – Unnecessary details and repetition are avoided when data are tabulated systematically. Anybody reading the table gets a clear idea of the data presented.

2. It facilitates comparison – Tabulation facilitates comparison. Because data are presented in rows and columns. Since a table is divided into various parts and for each part there are totals and sub-totals, the relationship between different parts of data can be studied more easily with the help of a table that without it.

3. It gives identity to the data – Data are arranged in a table with specific title and number This type of arrangement can easily identity the variables under study. This type of identification will become a source of interpreting a problem.

4. It reveals the trend and pattern of data – tabulation of data depicts the trend present in the variable under consideration. It also shows the pattern of the figures which cannot be studied in a descriptive way.

5. It facilitates Statistical Analysis – Tabulation of data in a systematic manner helps in determining statistical measures like averages, dispersion, correlation etc. These statistical measures will be useful for analyzing and interpreting the data under discussion

6. It economises space – Economy of space is achieved by tabulations; because all unnecessary details and repetitions are avoided without sacrificing quality and utility of the data.

Parts of a Table :

A table contain the following important parts :

(1) Table number

(2) Title of the table.

(3) Caption.

(4) Stub.

(5) Body of the table.

(6) Headnote.

(7) Footnote.

(1) Table Number :

Each and every table should be numbered for easy identification and future reference. The table number may be given on the top of the table or it may be give on the left hand side along with the title of the table. Sometimes table number is given at the bottom of the centre.

(2) Title of the Table :

Each table must have a suitable title. It should be given at the top of the table in the centre. The name of the table can show the nature of the data, the specific period of the data and the geographical, political or physical distribution of the data. Title should be written in prominent letters and sub-titles are generally written in small letters.

(3) Caption :

Caption refers to the column headings (arranged vertically). It explains what the column represents. It may consists of one or more column headings. Under a column heading there may be sub-heads. There will be sub-caption headings in that case. Caption should be clearly defined and placed at the middle of the column. If different columns are expressed in different units, the units should be mentioned with the captions. As compared with the main part of the table the caption should be shown in small letters, this helps in saving space. Where there are many columns, it is also desirable to number each column for easy reference.

(4) Stub :

Stubs refer to the headings of horizontal rows. They are at the extreme left and they explain the figures appearing against each stub. The box over the stub on the left of the table gives description of the stub contents and each stub labels the data found in its row of the table. The stubs are usually wider than columns headings but should be kept as narrow as possible without sacrificing precision and clarity of statements.

(5) Body :

This body of the table contains the numerical information. This is the most vital part of the table. Data presented in the body arranged according to descriptions or classifications of the captions and stubs.

(6) Head Note :

This may also be known as prefatory note. It is a brief explanatory statement applying to all or a major part of the material in the table. It is placed just below the title and in smaller or less prominent type. Usually it is used to explain certain points relating to the whole table that have not been included in the title, captions or stubs. For example, the unit of measurement is frequently written as a head note, such as "in thousands", or "in million tones" or "in crores" etc.

(7) Footnotes :

Any thing in a table which may not be clearly understood by the user form the title, captions ar d stubs should be explained in the footnotes. Such footnotes, if used, are placed below the body of the table.

GRAPHIC AND DIAGRAMATIC REPRESENTATION OF STATISTICAL DATA

Introduction

Graphs and diagrams are the most common appealing ways of presenting statistical data. They are very useful for studying the relationship between the variables. By the help of graphs and diagrams, the quantitative data can be presented in a very simple and clear manner.

Advantages of Graphic and Diagramatic Representation

(i) Graphs and diagrams are the simplest method of presenting data. They are easily understood. No mathematical knowledge is required to understand the graphs and diagrams.

(ii) Graphs and diagrams are attractive to the eye, interesting and impressive.

(iii) Graphs and diagrams represent the characteristics of the entire data. They also show clearly the relationships between the variables.

(iv) Graphs and diagrams facilitate comparison between two or more sets of data.

Construction of a Graph

Graphs are constructed on graph papers. Each graph paper has thick lines for each division of an inch or centimeter measure, and thin lines for smaller parts of an inch or centimeter. I order to construct a graph, two simple lines are first drawn which intersect each other at right angles. The point of intersection is known as the point of origin. By the intersection of two liens four quadrants are formed.

The vertical line YY' is known as Y-axis or the 'ordinate' and the horizontal line XX' is known as X-axis or the 'abscissa'. Both positive and negative values can be shown on the graph. One variable, say X is taken on X-axis and other variable, say Y, is taken on Y-axis. In quadrant I, both the values of X and Y are positive, in quadrant II, X is negative but Y is positive, in quadrant.

III, both the values of X and Y are negative, in quadrant IV, X is positive but Y is negative.

In order to plot a point in the graph, one value of variable X and another value of variable Y are required. Generally the independent variable is taken on the X-axis and the dependent variable on the Y-axis.

1.4 General Rules for Graphic Presentation

While constructing graphs of statistical data, the following points should be remembered.

(i) Title : every graph must have a comprehensive title so that the facts presented in the graph may be clearly known from the title.

(ii) Structure : In the graph, the independent variable should be measured along the X-axis and dependent variable along the Y-axis. The scale along the Y-axis should begin from zero as origin.

(iii) Choice of Scale : the choice of the scale should be such that it can accommodate the whole data. There is no strict rule for proper choice of scale.

(iv) Use of False Base Line : If the scale entries on the vertical axis require so much space that the Y-axis is unnecessarily elongated, then a portion of the scale may be omitted. So instead of showing the entire scale from zero to highest value involved, only the necessary portion is shown. The portion which lies between zero and the lowest value of the variable is left out.

(v) Foot Notes : Explanatory notes describe the important points of the graph should be given at the bottom of the graph.

(vi) Source : the source note should indicate the source of information from which the graph has been constructed.

1.5 Representation of Statistical Data

There are various types of charts and diagrams which are used for the representation of statistical data. The following are some of the common types of charts and diagrams :

(i) Line chart

(ii) Bar chart

(iii) Pie chart

(iv) Histogram

(v) Pictogram

(vi) Cartogram

1.6 Line Chart

Line chart is the most commonly used method of presenting statistical data. Line chart is a graph where the data are represented in the form of lines. This chart has the following characteristics :

(a) It show the relationship between two variables i.e. one independent variable and another dependent variable.

(b) It also shows the trend and tendency of the values of the variable.

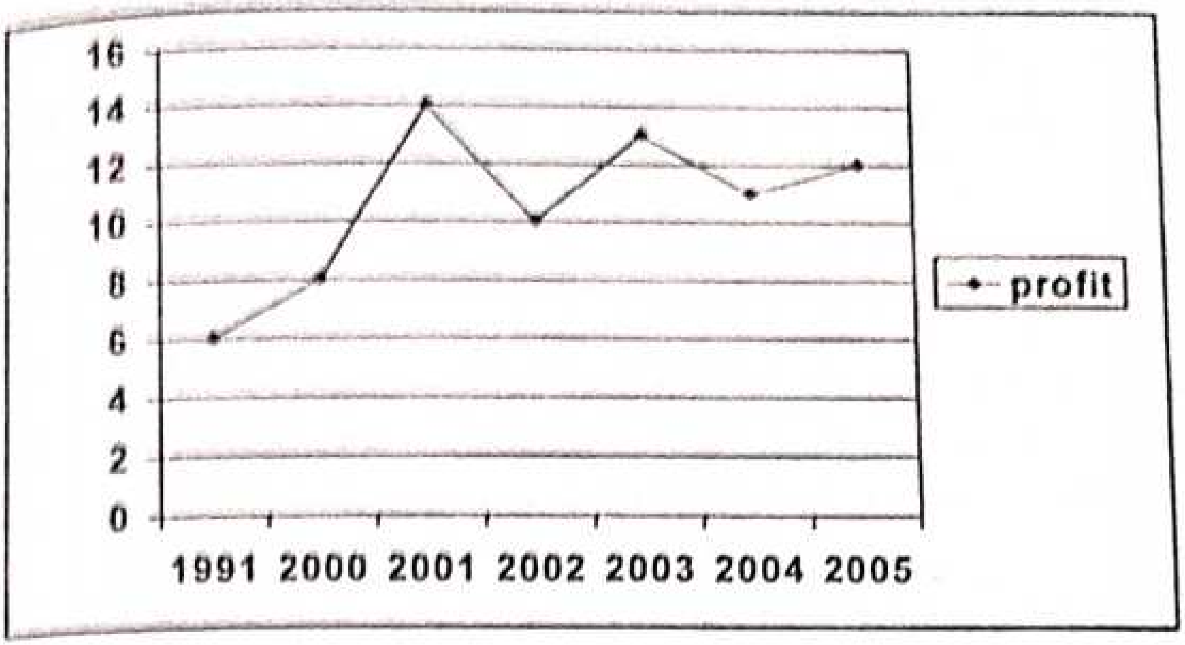
In order to draw a line chart, the independent variable, viz. time, year, months etc. should be taken on X-axis and the dependent variable viz, profits, costs, production etc. is taken on the Y-axis. Corresponding to the time factor, the values of the dependent variable are plotted. These points are joined by straight lines. This chart is known as line chart.

Illustrative Examples

Ex. 1 Draw a line chart from the following information regarding profit earned by Kalinga Ltd, during last seven years.

Year	1991	2000	2001	2002	2003	2004	2005
Profit (Rs. in lakhs)	6	8	14	10	13	11	12

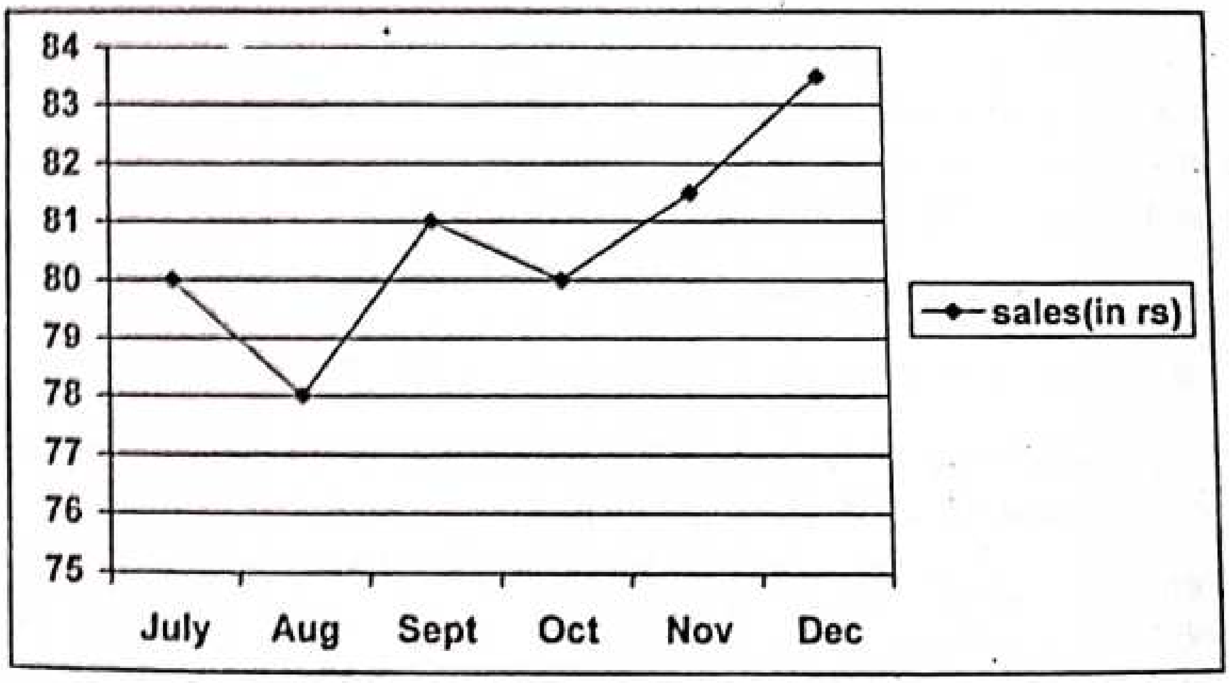
Solution : Profit earned by Kalinga Ltd from 1999 to 2005



In the above chart, years have been taken on the X-axis and the profits earned on Y-axis. This line chart shows the trend of the profit earned during different years.

Ex.2 The following data shows the sales of Utkal Ltd., from July to Decembor. Prepare a line chart to show the trend of sales.

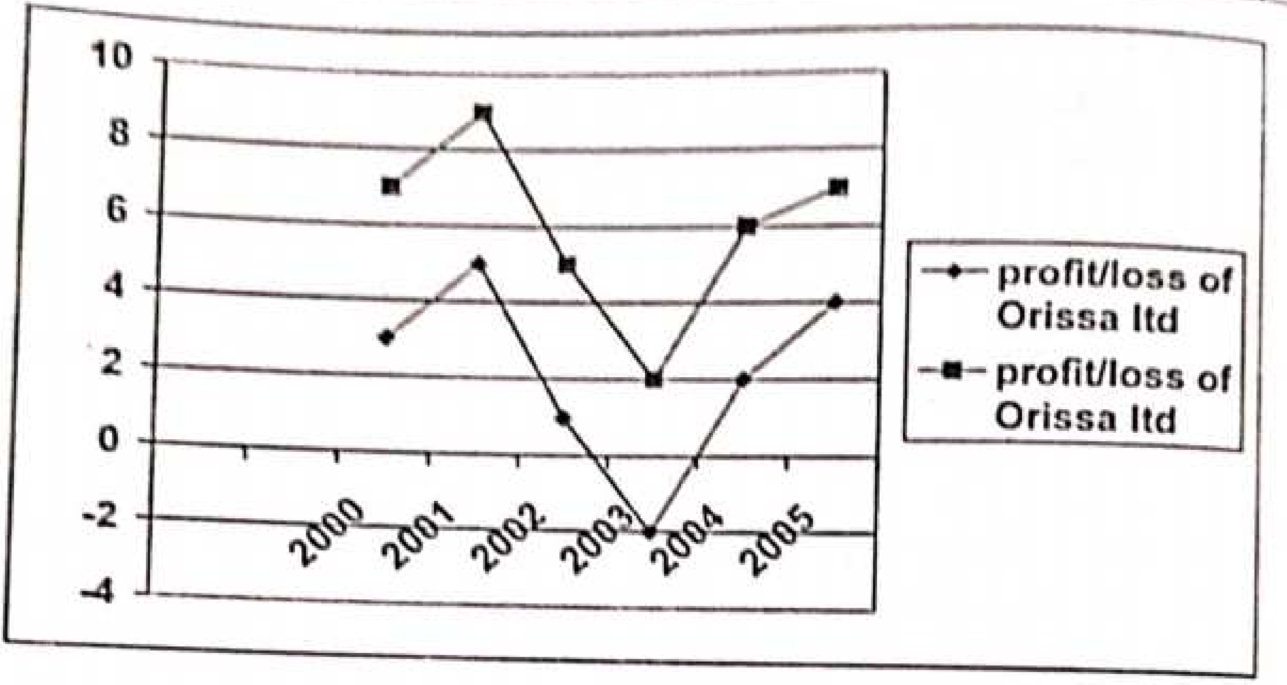
Month of 2005	July	August	Sept.	Oct.	Nov.	Dec.
Sale (Rs. in lakhs)	80	78	81	80	81.5	83.5



On the above line chart, false base line has been taken. Otherwise the Y axis will be unnecessarily elongated.

Ex.3 The following figures show the profit-loss of Orissa Ltd., and Utkal Ltd., for the last six years. Prepare a line chrat.

Year	2000	2001	2002	2003	2004	2005
Profit/loss of Orissa Ltd. (Rs in lakhs)	+3	+5	+1	-2	+2	+4
Profit/loss of Utkal Ltd. (Rs. in lakhs)	+7	+9	+5	+2	+6	+7



1.7 BAR CHART

A simple bar chart or bar diagram represents the magnitude of only one variable. A bar chart consists fo a group of equidistant rectangular bars. One bar reprints one figure. So the number of bars will be equal to the number of figures. Simple bar chart can be drawn wither on horizontal or vertical base. But commonly the bars on horizontal base are drawn. The following points should be noted while preparing a bar chart :

(a) All bars should be of same width.

(b) The intervening space between the bars must be equal.

(c) The scale is determined on the basis of highest value of the series and the length of bars should be proportional to the size of the value.

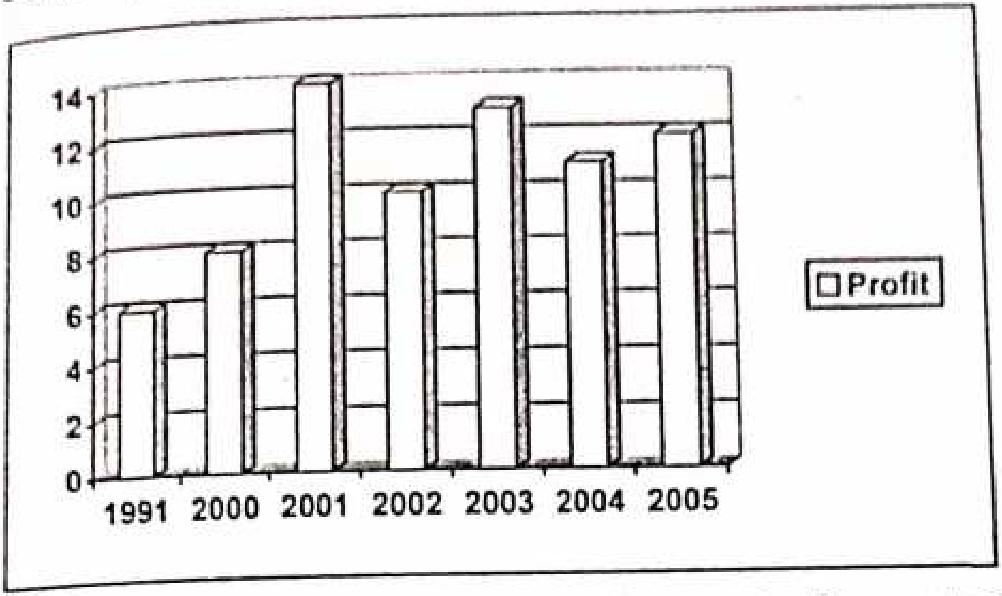
(d) The data regarding sales, profits, costs of production, units of production, population etc. for various years can be represented in bar charts.

Illustrative Examples

Ex 4. Taking the data of example I, draw a bar chart.

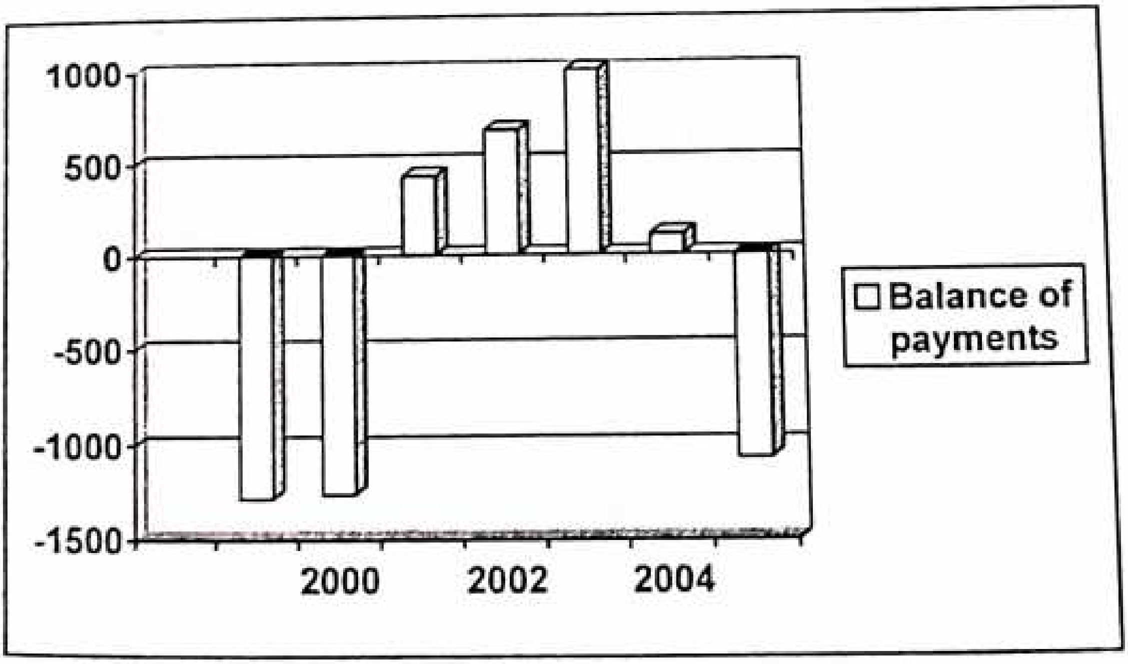
Solution :

Profit earned by Kalinga Ltd. from 1999 to 2005



Ex. 5. The information regarding the balance of payments of a country is given below. Prepare bar chart.

Years	1999	2000	200 1	2002	2003	2004	2005
Balance of payments (Rs. in million)	(-)1280	(-)1260	430	670	1000	110	(-)1070



Besides simple bar chart, there are also multiple bar charts and sub-divided bar char

Multiple Bar Chart

Multiple bar chart or multiple bar diagram is used when comparison is to be made betwe two or more sets of related statistical data. In this chart, separate bars are drawn for e phenomenon. The procedure for drawing such a diagram is the same as that of a simple diagram. The different bars for a period or related phenomena are placed together. After leav some gap, another set of bars for the next period or related phenomena are placed toge After leaving some gap, another set of bars for next period are shown. In order to distinguis bars of a period, different colours should be used or different types of dottings. This chart is a known as compound bar chart.

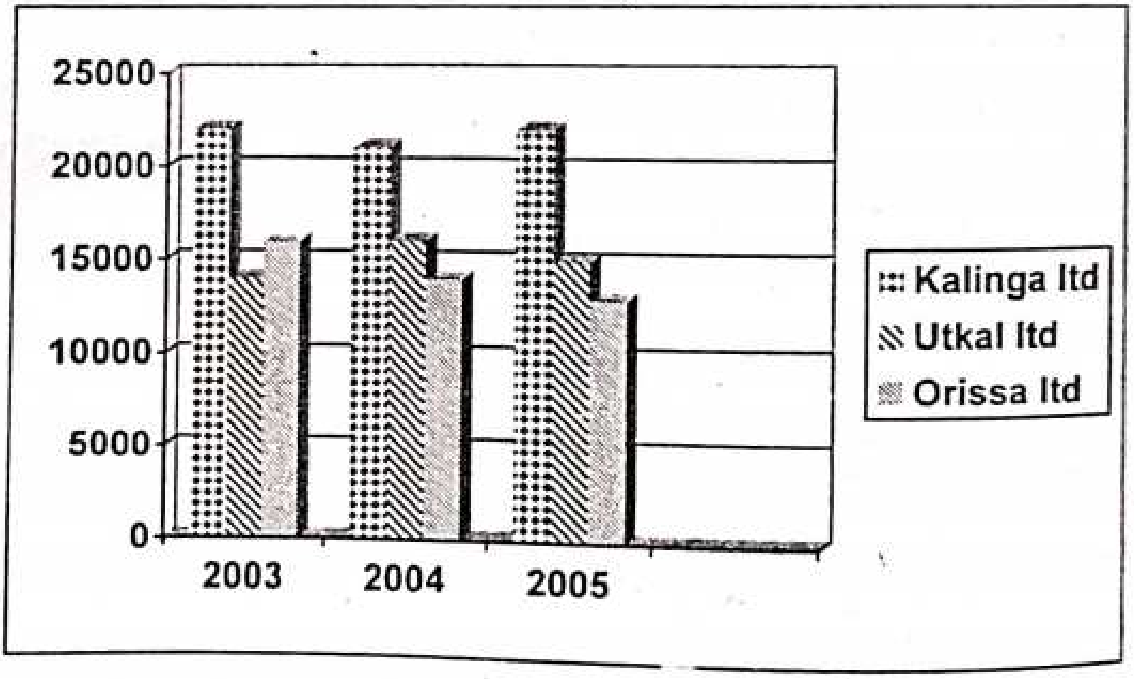
For example, if we want to compare the production or sale or profits of public sector st plants for the last three years a multiple bar chart should be prepared.

Illustrative Example

Ex. 6. Production of the following three companies for the last three years are given bel Construct a bar chart.

Years :	2003	2004	2005
Kalinga Ltd (in Units) :	22,000	21,000	22,000
Utkal Ltd (in Units) :	14,000	16,000	15,000
Orissa Ltd (in Units) :	16,000	14,000	13,000

Production of Kalinga Ltd., Utkal Ltd., and Orissa Ltd. from 1982 to 1984



Sub-divided Bar Chart

In a sub-divided bar chart, each bar represents the total magnitude of a given phenomenon, which is sub-divided into various parts in proportion to the values given. The sub-divisions are distinguished by different colours or dottings. The sub-divided bar chart is useful for comparing the sizes of various components parts among themselves, and also the relation between each component and the whole. It is also called component bar chart.

Illustrative Example

Ex. 7. From the following results of H.S.C. examinations for the four years, prepare a sub-divided bar chart.

Year	2002	2003	2004	2005
First class	5,000	7,000	8,000	10,000
Second class	10,000	13,000	12,000	15,000
Third class	20,000	15,000	17,000	14,000
Failed	25,000	30,000	35,000	36,000
Total	60,000	65,000	72,000	75,000

Solution :

Results of H.S.C. Exam. 2002 to 2005

1.8 PIE CHART

Pie chart or pie diagram is a circular diagram used for representing the total value with its components. The area of the circle represents the total value. The component parts or the total value are represented by different sectors of the circle.

Steps for construction a pie chart :

(a) First the data of component parts are expressed in the form of percentage of the total.

(b) Then the percentages are converted into degrees around the center of the circle. As the total angle at the center is 360°,3.6° represents 1% of the total value.

(c) A circle of appropriate size is drawn.

(d) With the help of a protractor, the points of the circle representing the size of each component part is found out.

Illustrative Example

Ex. 4. Mr. Patnaik gets a salary of Rs. 1800. Out of this, he spends Rs. 1080 on food, Rs. 90 on education of his children and Rs. 145 on miscellaneous activities. Show the above data on a pie chart.

Sub-divided Bar Chart

In a sub-divided bar chart, each bar represents the total magnitude of a given phenomenon, which is sub-divided into various parts in proportion to the values given. The sub-divisions are distinguished by different colours or dottings. The sub-divided bar chart is useful for comparing the sizes of various components parts among themselves, and also the relation between each component and the whole. It is also called component bar chart.

Illustrative Example

Ex. 7. From the following results of H.S.C. examinations for the four years, prepare a sub-divided bar chart.

Year	2002	2003	2004	2005
First class	5,000	7,000	8,000	10,000
Second class	10,000	13,000	12,000	15,000
Third class	20,000	15,000	17,000	14,000
Failed	25,000	30,000	35,000	36,000
Total	60,000	65,000	72,000	75,000

Solution :

Results of H.S.C. Exam. 2002 to 2005

1.8 PIE CHART

Pie chart or pie diagram is a circular diagram used for representing the total value with its components. The area of the circle represents the total value. The component parts or the total value are represented by different sectors of the circle.

Steps for construction a pie chart :

(a) First the data of component parts are expressed in the form of percentage of the total.

(b) Then the percentages are converted into degrees around the center of the circle. As the total angle at the center is 360°,3.6° represents 1% of the total value.

(c) A circle of appropriate size is drawn.

(d) With the help of a protractor, the points of the circle representing the size of each component part is found out.

Illustrative Example

Ex. 4. Mr. Patnaik gets a salary of Rs. 1800. Out of this, he spends Rs. 1080 on food, Rs, 90 on education of his children and Rs. 145 on miscellaneous activities. Show the above data on a pie chart.

Sol. First the data are expressed in the form or percentage and then percentages are converted into degrees around the center of the circle

	Rs.	%	Degree
Food	1080	60	216
Rent	215	12	43.2
Clothing	270	15	54
Education	90	5	18
Miscellaneous	145	8	28.8
			360

Ex. 9. A factory employs different types of workers, as given below : represent the data on a pie chart.

Workers	Foremen	Clerk	Fathers	Mechanics	Labourers
Number	4	12	9	38	21

1.9 HISTOGRAM

Histogram is a common method of presenting a frequency distribution graphically. Class intervals are taken on X-axis and the frequencies on Y-axis. Rectangles proportional to the area are erected. There is no space between the rectangles.

For example, if we want to show the continuous frequency distribution of heights of the students in a class, we have to draw a histogram.

The bar chart and histogram are not same. The following are the main differences between the two.

(a) In bar chart, equal spaces are left between two rectangular bars. But in histogram, there is no space between two bars.

(b) In bar chart, the length of the bar is material as it sis a one dimensional diagram. But a histogram is a two-dimensional diagram where both the length and width of the bar are material.

Illustrative Example.

Ex. 10 The following table shows the frequency distribution of the weights of the students in a class of 40 students. Construct a histogram.

Class intervals (weight in kg)	No of students
40-45	12
45-50	12
50-55	14
55-60	8
60-70	4

Graph

Note : The scale along the y-axis should begin from zero. But the scale along the x-axis may not begin from zero.

1.10 Pictogram

A pictogram is a chart which represents data by using simple pictures. Pictogram (i.e. pictures) are very useful in attracting the attention of the user. As they are very easy to understand, the are used for presenting data to common men. In a Pictogram each symbol or picture represents a definite numerical value. If a fraction of the numerical value represented by a symbol or picture occurs, then the proportional part of the symbol or picture from the left is drawn.

Ex.12 The following data shows the number of car is owned by a master company in different years. Represent the data by Pictogram.

Year	2001	2002	2003	2004	2005
No. of car	174	231	255	330	359

Solution :

Car

XXXXXXXXX Graph

1.11 CARTOGRAM OR MAPOGRAPH

In a cartogram or mapograph, a map is accompanied by various types of diagramtic presentation. On a map, data may be shown either (a) by paints, dots or crosses, or, (b) by deepening the colour in proportion to the magnitude. So cartograms are statistical maps.

❖❖❖

UNIT – II

DEFINITION AND CONCEPT OF MEASURES OF CENTRAL TENDENCY

A measure of central tendency or an average may be defined as a single figure calculated to represent a statistical series for a comparative study. From the above definition it follow that an average possesses the following characteristics :

(a) It is a single figure and not an aggregate of facts.

(b) It is calculated on the basis of a set of data.

(c) It is a figure capable of representing the entire series,

(d) It is calculated for the purpose of making a comparative study.

Concept :

An average is thus, a single figure. It is calculated mainly with the following two objects

(a) To reduce the complexity of the data

(b) To make the data comparable.

(a) To reduce the complexity of the data :

The statistical data are complex by nature. Human mind is not capable either to remember them all or to understand them properly. Thus if the marks of 128 students of Section A and 123 students section B will be presented before a person, it will on his part to have a clear idea about the standard of students of both the sections, nor it will be possible on his part to remember them all for any future purpose. But instead of the actual marks if the average of the marks of both the Sections will be presented before him, it will be quite easy on his part to understand the standard of the students of both the sections and, also it will be possible on his part to remember them all for any future purpose. In this way an average reduces the complexity of data and makes them intelligible.

(b) To make the data comparable :

Another object of an average is to make the data comparable with similar type of other data. From the absolute data no one can have any clear idea about he problem. It is the process of comparison that makes the data more intelligible and more useful on the part of an investigator. With this object in view the average of various figures are taken into consideration in most of the fields of statistical analysis.

2. TYPES OF CENTRAL TENDENCIES :

The various measures of central tendencies that are used in various fields of statistical analysis can be broadly divided into two types viz. 1. Mathematical averages, and 2. Positional average.

Mathematical averages :

This class of averages includes :

1. Arithmetic average or the mean, 2.Geometric average, and 3 Harmonic average. These are called mathematical averages because in their calculations the mathematical principles and procedures are strictly followed.

Positional averages :

This class of averages includes the following important ones : 1.Median.2.Quartles, 3.Deciles 4.Percentiles, 5.Septiles, 6.Octiles, 7.quintiles, and 8.Mode. Among these the Median, Mode and Quartiles are found to be most popularly used. These averages are called positional averages because their values are ascertained by mean of location of their position in the distribution to items.

3. DESIDERATA OR ESSENTIAL QUALITIES OF AN IDEAL MEASURE OF CENTRAL TENDENCY

Before we proceed to the actual calculation procedure of each fo the averages mentioned above, it will proper to mention here the essential qualities or the desiderata which an ideal average should have. These are :

(a) The average should be rigidly defined and should have a definite value. This means that there should be ambiguity as to its meaning and that the determination of its value would not be left to estimation.

(b) The average should be calculated on the basis of all the items of the series. This means no item of the series is ignored in the calculation of an average.

(c) The average should not be affected by the values of the extreme items of the series. This means, its values of the greater items of the smaller items of the series. In that case it would not be truly representative of the series.

(d) The average should not be greatly affected by fluctuation of Sampling. This means, in the field of sampling where samples are collected at two or more different times or by two or more different investigators the values of the averages thus found out from time to time should not differ very much from each other.

The average should be capable of further algebraic treatment. This means the calculation procedure of an average should be such that the algebraic principles are not where deviated so that further and further calculations can be made on the basis of the averages thus obtained.

4. ARITHMETIC AVERAGE OR THE MEAN

Definition :

Arithmetic average or the mean may be defined as the value which is obtained by dividing the total of the values of a series by the total number of items of the series. Thus, it the values of the items in a series are 2,3,4,5 and 6 its arithmetic average will be

$$\left(\frac{2+3+4+5+6}{2}\right)$$

Arithmetic average is popularly known as the mean, it is also the most popular average of all the averages that are used in the various fields. In fact, when the idea of average is presented before a layman he immediately refers it to the arithmetic average. Symbolically such an average is stated as follows :

$$a = \frac{\sum m}{n}$$

Where, a stands for the arithmetic average, m for the sum of the values of the m variable and n for the total number of items in the series.

Calculation of Arithmetic Average in Simple Series :

In a simple series there will be no difficulty in calculáting the value of the arithmetic average. In such a case the value of the mean can be ascertained by any of the following two methods : 1. Direct method and 2. Short-cut method.

Direct Method :

Under this méthod the mean will be calculated with the application of the following formulae :

$$a = \frac{\sum m}{n}$$

and for this value of the items will be simply totaled which will constitute the value of the m. This total value will be divided by the total number of the items. The resultant figure thus obtained would given the value of the desired mean the following illustration would clarity the above method.

Example - 1

From the following data relating to the marks of five students ascertain the average of marks secured by a student

Roll Nos.	1	2	3	4	5
Marks	30	60	40	20	80

Solution

Calculation of Average

Roll Nos.	Marks
	(m)
1	30
2	60
3	40
4	20
5	80
N=5	Σ m=230

By the formulae we have :

$$a = \frac{\sum m}{n} = \frac{230}{5} = 46$$

Hence the average mark secured by a student is 46.

N.B.: Roll Nos. Column represents each individual items of the series and so it should be confused with the values of the variables. He re it is the mark column, which represents the values of the variables of which only the average is being calculated.

Short – cut Method :

This method will be application where the values of the items appear to be of big size and are in large number. In such a case the process of addition would entail a bit of difficulty and to do away with such difficulty the short-cut method should, advantageously, be used. Under this method the following steps are to be taken :

1. Assume a value to be the value of the arithmetic average. Such assumption should, preferably, be made from the middle part of the series.

Example - 1

From the following data relating to the marks of five students ascertain the average of marks secured by a student

Roll Nos	1	2	3	4	5
Marks	30	60	40	20	80

Solution

Calculation of Average

Roll Nos.	Marks
	(m)
1	30
2	60
3	40
4	20
5	80
N=5	Σ m=230

By the formulae we have :

$$\bar{x} = \frac{\sum m}{n} \quad = \frac{230}{5} = 46$$

Hence the average mark secured by a student is 46.

N.B.: Roll Nos. Column represents each individual items of the series and so it should be confused with the values of the variables. He re it is the mark column, which represents the values of the variables of which only the average is being calculated.

Short – cut Method :

This method will be application where the values of the items appear to be of big size and are in large number. In such a case the process of addition would entail a bit of difficulty and to do away with such difficulty the short-cut method should, advantageously, be used. Under this method the following steps are to be taken :

1. Assume a value to be the value of the arithmetic average. Such assumption should, preferably, be made from the middle part of the series.

2. Calculate the deviations of the various items of the series from the assumed average (with + and – sings).

3. Calculate the step-deviations by dividing each of the deviations with a common factor (Of course, if possible and necessary).

4. Find the total of the deviations or that of the step deviations, as the case may be, and the total of the number of the items.

5. Put the values in the following formulae and get the arithmetic average or the mean :

$$a = \frac{\sum dx}{n} \quad \text{(when deviation are not found)}$$

$$\text{or } a = \frac{\sum dx/c}{n} xc \quad \text{(when step deviations are found)}$$

Where, a stands for the actual arithmetic average, X for the assumed average Σ dx for the sum of the deviations from the assumed average, Σ dx/c for sum of the steps deviations from the assumed average, c for the common factor and n for the total number of the items.

The following example would illustrate the above methods.

Example – 2 :

From the following data relating the income of a certain person find out the average income of a person. In come in Rs. 180, 260, 750, 880, 940.

Solution :

Calculation of the average income of a person

Income in Rs.	Deviations from step deviations the assumed. Mean with c = 10	
(m)	(dx)	(dx/c)
180	-570	-57
260	-490	-49
750	0	0
880	130	13
940	190	19
N=5	dx=740	dx/c=74

According to the first formula :

$$a = X \frac{\sum dx}{n} = 750 + \frac{740}{5} = 750 + -148 = 602$$

According to the second formula :

$$a = X + \frac{\sum dx/c}{n} Xc = 750 + \frac{-74}{5} x10 = 750 - 148 = 602$$

Thus, in both tho cases the average income of a person comes to be Rs 602.

Calculation in a Discrete Series :

In case of a discrete series the calculation of the mean value can be made either under Direct or Short-cut method as discussed above. The special consideration to be given, here is the treatment of the frequency column. Under Direct method the frequency column will simply be multiplied with the value or them column. But under the short-cut method the frequency column will be multiplied either with the deviation of with the step deviation column. Thus the symbolic representation of the formulae of the arithmetic average will be modified here as under:

Direct Method :

$$a = \frac{\sum mf}{n}$$

Where, a stands for arithmetic average, mf for the sum of the product of the m variables and their respective frequencies, a n for the total number of the items.

Short-cut Method :

$$a = X + \frac{\sum dfx}{n} \quad \text{(if step deviations are not found)}$$

or $$a = X + \frac{\sum dfx/c}{n} Xc \quad \text{(if step deviations are found)}$$

Where, a stands for the actual mean, x for the assumed mean, fdx for sum of the product of deviations from the assumed mean and the frequencies, fdx/c for sum of the product of the step deviations from the assumed mean and the frequencies, c for the common factor of step deviations and n for the total number of items.

The above procedure can be well understood from the following example.

Example - 3

Form the following frequency distribution find out the value fo the mean under both the Direct and Short-cut methods.

Values :	10,	15,	5,	20,	3,
Frequency :	20	40	15	16	5

Solution under Direct Method :

Calculation of Arithmetic Average

(m)	(f)	(m f)
10	20	200
15	40	600
5	15	75
20	16	320
3	5	15
	Σ f=96	Σ m f=1210

Here Σ f or n = 96 and Σ mf = 1210

By the formula we have :

$$a = \frac{\sum mf}{\sum f} = \frac{\sum mf}{n} = \frac{1210}{96} = 12.6 \quad \text{approx}$$

Solutions of Arithmetic Average :

(m)	(f)	(dx) x = 15	(fdx)
10	20	-5	-100
15	40	0	0
5	15	-10	-150
20	16	5	80
3	5	-12	-60
		Σ f=96	Σ fdx= - 230

By the formula we have :

$$a = X + \frac{\sum fdx}{n} = 15 + \frac{(-230)}{96} = 15 + (-2.395) = 12.6$$

Thus, under both the methods the value of the arithmetic average is 12.6 approximately.

Calculation of Arithmetic Average in a Continuous Series :

The procedure of calculating the arithmetic average in a continuous series under both the methods discussed above will remain the same. The only special treatment required here is to find out the mid values of each of the class intervals and to treat these mid values as the values of the m variable.

Where, however, the class magnitudes of all of the class intervals appear to be equal, the following shortest method may be advantageously used

Formula Under Shortest Method :

a = M - i (F, - 1)

Where a stands for the arithmetic average, M for the mid value of the last class interval, i for the class magnitude and F, for he average of the cumulative frequencies

The following example would illustrate the calculation of the arithmetic average in a continuous series under each of the above three methods.

Example - 4

Ascertain the value of the arithmetic average from the following group data

Class intervals	0-10	10-20	20-30	30-40	40-50
Frequency	7	5	4	8	16

Solution :

Under Direct Method :

Calculation of Arithmetic Average

Class interval	Mid values	Frequencies	Product of M.V. & Frequency
	(m)	(f)	(mf)
0-10	5	7	35
10-20	15	5	75
20-30	25	4	100
30-40	35	8	280
40-50	45	16	720
		n = 40	mf = 1210

By the formula we have :

$$a = \frac{\sum mf}{n} = \frac{1210}{40} = 30.25$$

Under Short-cut Method :

Calculation of Arithmetic Average

Class	M.V.	Frequency	Deviation	Step Devn,	Frequ. & Step
intervals	(m)	(f)	(dx) 25	(dx/c) 10	(fdx/c)
0-10	5	7	-20	-2	-14
10-20	15	5	-10	-1	-5
20-30	25	4	0	0	0
30-40	35	8	10	1	8
40-50	45	16	20	2	32
			n = 40		Σ fdx/c=21

By the formula we have :

$$a = X + \frac{\sum fdx/c}{n} = 25 + 2\frac{1}{40} = 30.25$$

Under short-cut method

Calculation of Arithmetic Average

Class intervals	Frequencies	Cumulative frequencies
0-10	7	7
10-20	5	12
20-30	4	16
30-40	8	24
40-50	16	40
	n=40	Σ cf=99

Here $F_1 = \frac{\sum cf}{n} = \frac{99}{40} = 2.475$

M = 45, I = 10, n = 40

Thus, a = m – i (F$_1$ – 1)

= 45 – 10 (2.475-1) = 45-10 (1.475) = 45 – 14.75

= 30.25.

Hence, the value of the arithmetic average is 30.25.

ALGEBRAIC TREATMENT

Arithmetic average possesses many algebraic properties, for this, it is capable of further algebraic treatment in different matters. A few cases of its algebraic treatment are discussed here as under :

1. The aggregate of items of a series can be known. Thus, if in a series, the value of 'a' and 'n' are respectively 20 and 40 the aggregate of the items of the series can be found as under.

2 The number o items of a series can be ascertained, if the value of its arithmetic average and the aggregate of the items are known.

3. The Arithmetic average of a series can be found out if the aggregate of the items and their numbers are given.

4. The compound or the combined average of two or more series, or two or more component parts of a series can be ascertained easily, if the value of the arithmetic average and the number o items of each of the value of the series or that of each of the component parts of the series (as the case may be) are given. For ascertaining such compound average the following formula has to be simply applied.

$$\frac{a_{12} = (a_1 x n_1) + (a_2 x n_2)}{n_1 x n_2}$$

Where, a_{12} sands for the compound average of the 1st and the second series, a_1 and n_1 for the arithmetic average and the number of items of the 1st series and a_2 and n_2 for the arithmetic average and the number of items of the second series.

Example – 5

From the following data ascertain the value of the compound average

	1st series	2nd series	3rd serie
Arithmetic average	30	20	10
Number of items	25	50	60

Solution

By the formula we have :

$$\frac{a_{123} = a_1.n_1 + a_2.n_2 + a_3.n_3)}{n_1 + n_2 + n_3}$$

$$= \frac{(30x25) + (20x50) + (10x60)}{25 + 50 + 60} = \frac{750 + 1000 + 600}{135} = \frac{2350}{135} = 17.4$$

Hence the value of the compound average of the three series taken together is 17.4 approximately.

Merits and Demerits of Arithmetic Average

Having thus understood the meaning and the procedure of the arithmetic average now it would be possible to lay down it various merits and demerits in the light of the characteristics of an ideal measure of central tendency. These are discussed as under :

Merits

1. It is rigidly defined and its value is always definite.
2. Its calculation is based on all the items of the series.
3. It is very simple to understand.
4. It is easy to calculate and its calculation does not need arraying of the data.
5. It is capable of further algebraic treatment.
6. It is not greatly affected by the fluctuation and sampling.
7. It sit he center of gravity balancing the values on either side of a series.

Demerits

1 It is greatly affected by the values of the extreme items of the series.

2. It is greatly dependent upon all the items of the series.

3. It involves the mathematical intricacies in its calculation.

4. It is likely to give absurd results at times. For example, the average of 2 and 3 children would be 2.5 under its procedure which is never practicable.

5. It has an upward bias. This means, one big item among five items four of which are small would push up the average considerably but the reverse is not true.

GEOMETRIC MEAN

Geometric mean is defined as the nth root of the product of n items of a series. Symbolically it is stated thus :

$$g - \left(\sqrt[n]{m_1} x m_2 x m_3 x \ldots\ldots x m_n\right)$$

Where, g stands for the geometric mean, n for the number of items and m for the values of the variables.

Thus, if in a series there are three items viz. 1,2, and 4 its geometric mean would be

$$g = \left(\sqrt[3]{1x}2x4\right) = 2:$$

This method of calculating the geometric mean will be possible only where the number of items and their size are small. But there the number of items and their size are large, the method would be impracticable. In such a case the logarithmic procedure has to be followed.

Logarithmic Procedure

Under this procedure, the geometric mean can be found out by both the direct and the short cut method these are discussed as under.

Direct Method

Under this method the logs of the various items to be found out first. These logs would then be totaled and the total thus found out would be divided by the total number of items. This would give us the average of the logs of the variables. Now, the anti log of such average would be found out which would be the value of the geometric mean. Symbolically it would be stated as follows :

$$g = Anti\log of \frac{\sum \log X}{n}$$

Where, g stands for the geometric mean, logs for sum of the logs of the m variables and n for the total number of items.

Short cut Method

Under this method the logs of the various items will first, be found out. Then a log will be assumed to be the log of the geometric mean. From this assumed log the deviations of each of the logs would be found out and totaled. This total will then be divided by the number of items and the result thus obtained will be added to the logs of the assumed geometric mean. Now the anti log of the resultant log will be found out which give the value of the desired geometric mean. Symbolically it will be stated as follows :

$$g = A.L.\frac{(x + \sum \log dx)}{n}$$

Where, g stands for the geometric mean, A.L for the anti log, x for the log of the assumed geometric mean, S log. dx for sum of log deviations from the assumed geometric mean and n for the total number of the items. The following example would clarify the calculation of geometric mean under each of the above procedures.

Direct Method

Example – 6

From the following data ascertain the geometric mean under the direct method.

Calculation of Geometric Mean

(m)	(Logs)	(Log deviation) x = 3
.10	1.000	2.0000
.020	2.3010	1.3010
.0030	3.4771	0.4771
.00040	4.6021	1.6021
.000050	5.6990	2.6990
n=5		Σ log dx = 2.0792

By the formula we have :

$$g = A.1\frac{(x+\sum \log dx)}{n} = A.L.\frac{(3.+2.0792)}{5} = A.L.3.4158 = .002605$$

Hence, the geometric mean of the series is .002605.

Algebraic treatment of Geometric Mean

Just like the arithmetic average, Geometric Mean possesses many algebraic properties for which it is capable of algebraic treatment in various ways. Some of these ways are described hear as under :

1. The product of the items of a series can be known if the values of the G.M. and their numbers are given. Thus the G.M. of the items 2, 4 and 8 would be 3"2x4x8=4.The product of the items will also be the same i.e. 4x4x4=64. Hence it is seen that the products of the original items can be easily found out it the values of the G.M. and their numbers are known.

74356., 6745., 326., 3., .14, .0250, .00304.

Solution

Calculation of geometric Mean

(m)	(Logs)
74356.	4.8714
6745.	3.8290
326.	2.5132
54.	1.7324
3.	0.4771
.14	*1.1461
.0250	*2.3979
.00304	*3.4829
n=8	Σ log = 8.4500

(*bar signs indicate that the whole number is a negative number)

(i.e. 1 = 1)

By the formula we have :

$$g = A.L\frac{\sum \log dx)}{n} = A.L.\frac{8.4500}{8} = A.L.31.05625 = 11.391$$

Hence, the G.M. of the series is 11.391.

Short cut Method

Example – 7

From the following data find out the geometric mean under the short cut method. .10, .020, .0030, .00040, .000050.

The combined G.M. of two or more series can be found out if the separate G.Ms. and number of items of each of the series are given. For this however, the following formulae will have to be applied.

$$g_{12} = A.L\frac{(n_1.\log g_1 + n_2.\log g_2)}{n_1 + n_2}$$

Where, $g_{1\,2}$ stands for the geometric mean of both the 1st and the 2nd series, n_1 and n_2 for the number of times in the respective series and g_1 and g_2 for the geometric means of both the series respectively. The calculation of the combined G.M. described above can be well understood from the following example :

Example – 8

Find out the combined G.M. of the two series from the following data.

	1^{st} series	2^{nd} series
G.M.	10	50
Number of items	5	4

Solution

By the formula we have :

$$g = A.L\frac{(n_1.\log g_1 + n_2.\log g_2)}{n_1 + n_2}$$

$$g = A.L\frac{(5x\log 10 + 4x\log 50}{5+4}$$

$$= \frac{A.L.(5x1.0000)+4}{9}x\frac{1.6990}{9} + A.L.\frac{(5+6..7960)}{9} + A.L.\frac{11.7960}{9} = A.L.1.3107$$

$$= 20.45$$

Hence the combined G.M. of both the series is 20.45.

3. The product of the ratios of the items to the G.M. on either side of the G.M. will be equal. This can be understood from the following example.

Example – 9

From the following data prove that the product of the corresponding ratios on either side of the G.M. is equal :

3,6,8,9

Solution

Hence, the G.M. would be $g = 4\sqrt{3x6x8x9 = 6}$

The product of the ratios of the G.M. to be the items whose values are more than the G.M. $= \frac{8}{6} x \frac{9}{6} = 2$

Hence, it is proved that the product of the corresponding ratios on either side of the G.M. $= \frac{8}{6} x \frac{9}{6} = 2$

Hence, it is proved that the product of the corresponding ratios on either sides of the G.M. is equal.

The G.M. of the ratios of the corresponding items in two series is equal to the ratios of their G.Ms. Say, there are two series as follows :

	B
A	2
3	4
6	4
8	8
9	

The ratios of the corresponding items would be :

A/B

1.5

1.5

2.

1.125

The G.M. of the series A, series B and that of the series A/B will be as follows.

G.M. of A : $g = 4\sqrt{3x6x8x9 = 6}$

G.M. of B : $g = 4\sqrt{2x4x4x8 = 4}$

G.M. of A/B : $g = 4\sqrt{1.5x1.5x2x1.125 = 1.5}$

Now, the ratios of the G.Ms. of the series A and that of the series B would be

$\frac{6}{4} = 1.5$

Solution

Here, the G.M. would be $g = 4\sqrt{3x6x8x9} = 6$

The product of the ratios of the G.M. to be the items whose values are less than the G.M. $= \frac{8}{6} x \frac{9}{6} = 4$

Hence, it is proved that the product of the corresponding ratios on either side of the G.M. $= \frac{8}{6} x \frac{9}{6} = 4$

Hence, it is proved that the product of the corresponding ratios on either sides of the G.M. is equal

The G.M. of the ratios of the corresponding items in two series is equal to the ratios of their G.Ms. Say, there are two series as follows :

A	B
3	2
6	4
8	4
9	8

The ratios of the corresponding items would be :

A/B

1.5

1.5

2

1.125

The G.M. of the series A, series B and that of the series A/B will be as follows.

G.M. of A : $g = 4\sqrt{3x6x8x9} = 6$

G.M. of B : $g = 4\sqrt{2x4x4x8} = 4$

G.M. of A/B : $g = 4\sqrt{1.5x1.5x2x1.125} = 1.5$

Now, the ratios of the G.Ms. of the series A and that of the series B would be

$\frac{6}{4} = 1.5$

Hence, it is proved that G.M. of the ratios of the corresponding items in two series is equal to the Geometric means.

1. The G.M. of the products of the corresponding items in two series is equal to the product of their G. means. Thus, if in the above example we multiply the corresponding items of A and B series the products would be respectively, 6, 24, 32 and 72 and their G.M. equal to 24. The product of the G.M. of these two series A and B is also 24 i.e. (6x4).

2. In the calculation of the average rate of increase of any sum at compound process the use of G.M. is highly useful. Thus in the case of compound interest the process of G.M. is inevitably used. For example, if Rs.1000 at compound interest amounts to Rs. 1500 after 10 years, the compound rate would be

$A = P\ (1+i)^n$ (By the formula of compound interest)

$A = P\ (1+i)^n = A$, or $(1+i)^n = \frac{A}{P}$ or, $(1+i) = {}^{n}?\ \frac{A}{P}$

Or $I = {}^{n}?\ \frac{A}{P} - 1 = 10\frac{1500}{1000} - 1 = {}^{10}?1.5 - 1 + 1.04\ 1 - 1 = .041 = 4.1\ \%$

Merits and Demerits of the Geometric Mean

The G.M. as a mathematical measure of dispersion can be said to have the following merits and demerits :

Merits

1. It is rigidly defined and its value is always definite.
2. it is capable of further álgebraic treatment.
3. it is based on all the observations of the series.
4. It is not affected much by the fluctuation and sampling.
5. It gives comparatively more to the sampling.
6. It is a suitable measure for relative study like that of Index number.

Demerits :

1. It is not simple to follow on the part of a man of ordinary prudence.
2. It is difficult to calculate for its algebraic procedure.

3. It cannot be calculated if any clue of the series is either zero or a negative number.

4. It may give a value, which may not be found in the series.

Harmonic Mean

This is another mathematical measure of central tendency, which is defined as the reciprocal of the arithmetic average of the reciprocals of the values of the variables. Thus, in its calculation the following steps are to be adhered to :

1. Finding the reciprocals of the various values of the variables.
2. Finding the average of the reciprocals thus obtained.
3. Finding the reciprocals of the average of the reciprocals thus found out.

The symbolic representation of the Harmonic mean, in the light of the above steps would be

$$h = r\frac{\sum r}{N}$$

Where, h stands for the harmonic mean, r for the reciprocals, r for sum of the values of the variables and n for the number of items.

Procedure for finding out the reciprocals :

The important matter connected with the harmonic mean is the finding of the reciprocals. One way is to divide the figure by a particular figure. Thus, to find out the reciprocals of the figure 5 we have to find out the result of 1/5 or to find out the reciprocals of .03 we have to ascertain the value of 1/.03. This way of finding out the reciprocals would be preferable only in case of small sized figures, where the division.can be made at an ease. But in case of big figures viz. .009017, 137.5708 the division of one (1) by such figures would be very difficult. In such cases the another way of finding out the reciprocals of the figures would be very much easier as well as practicable. Such way of finding of reciprocals with the help of the Reciprocal Table. This way of finding the reciprocals will of course, need an understanding of the technique involved in it. This is discussed as under :

Procedure of finding the Reciprocals from the Table

- There is a Table called reciprocal Table, which is usually provided to pick up the reciprocals of the various numbers. Just like Log Table, this Table also contains a vertical line to the left and a horizontal line on the top. The vertical lien contains the first two digits of the natural number and the horizontal line contains the reciprocal numbers for different natural numbers stated in the vertical line. Thus, the Table takes the following shape :

Reciprocal Table

0123456789 Mean differences	
1.0	
1.1	
1.2	8130
1.3	
1.4	
1.5	

To find out the reciprocal of a figure the first two digits of the figure will be located at the vertical column and the reciprocal of this will be picked up from the horizontal line with reference to the third digit of the figure. Thus to find out the reciprocal oft h figure 123 we have to locate 1.2 in he vertical column first, and then refer to the reciprocal number that stands on the horizontal lines against the third digit .3 of the figure 123 which is 8130 in this case. Before referring to the Table, however, the following steps are to be followed first.

1. If the item is a positive number (i.e. having some digits before the decimal point), put the decimal point first, and then put the zero equal to one less than number of digits in the natural number before the decimal point. Thus, for, 123 the first procedure will be to write .C0. After this the Table figure will eb put at the right hand side. Thus, the reciprocal of 123 will be .008130.

2. If the item is a negative number, (i.e. having no digit before the decimal point, pick up the Table figure first, and the put the decimal point after one more place than the number of the zeros appearing after the decimal point put before any significant digit in the given item. Thus, in the case of 123, the Table figure to be picked up is 8130. In this the decimal point will be put after 8, because there is no zero after the decimal point and before the significant digit '1' in the item 1.123 thus the resultant reciprocal would be 8.130.

3. If the item consists only the digit '1' whether positive or negative and there is not other significant digit except zeroes, whether before or after, the decimal point thus put as half year both the above steps will only be shifted by one place to the right. Thus, for the figure 100, the step would have been to put .00 first, (because there are three digits before the

decimal point), but as the figure consists no other digit than a single 1, the preliminary step will be to shift the decimal point by one place more to the right. Thus, the preliminary step will be to put here .0, (here the decimal point has crossed the first zero to the right). In the above way the reciprocals of all types of figures, staggering they may be found out at an ease. The following example would clarify the calculation of Harmonic mean discussed above.

Example – 10

From the following series find out the value of the H.M.

145236, 78543, 9327, 456, 17, 10, 8.

Solution.

Calculation of Harmonic Mean

(m)	(Reciprocals)
145236	.00000689665
78543	.00001273890
9327	.00010729600
456	.00219298000
17	.05882350000
10	.10000000000
8	.12500000000

Here n=7, R=28614341155 By the formula we have :

$$h = r\frac{\sum r}{n} = rof\,\frac{.28614341155}{7}$$
$$= rof\,.04087763022 = 24.5098 = 2451(approx)$$

Example – 11

From the following data ascertain the Harmonic Mean.

100

10

1.1

.01

.001

.55

.056

.007

Solution

Calculation of Harmonic Mean

(m)	Reciprocals
100	.01000
10	.10000
1	1.00000
.1	10.00000
01	100.00000
.001	100.0000
.55	1.81818
056	17.85710
.007	142.85700
n=9	1273.64220

By the formula we have

$$h = rofr = rof\frac{1273.64220}{9} = rofr141of141of141.51580 = .00709230$$

Example – 12

From the following frequency distribution find out the value of harmonic mean.

m=0.4, 4.8, 8.12, 12.16, 16.20

f=4 18 20 42 36

Solution

Direct Method

Calculation of Harmonic Mean

(m)	(m.v).	(f)(Reciprocal)		(R.f)
0.4	2	5000	2.0000	
4.8	6	18	.1667	3.0006
8.12	10	20	.1000	2.0000
12.16	14	42	.0714	2.9988
16.20	18	36	.0556	2.0016
		N=120		Σr=12.0010

By the formula we have :

h=r. $h = r\frac{\sum r}{n} = rof\frac{12.0010}{120}$ = r of .100000 = 10.00

Shortcut Method :

Calculation of Harmonic mean.

(m)	(m.v)	(f)	(f/n)
0-4	2	4	2
4-8	6	48	3
8-12	10	20	2
12-16	14	42	3
16-20	18	36	2
		n=120	Σ f/n = 12

By the formula we have :

$$h = \frac{N}{\sum \frac{f}{n}} = \frac{120}{12} = 10.$$

Merits and Demerits of Harmonic Mean

H.M. has the following merits and demerits respectively.

(1) It has a definite meaning and its values are always definite.

(2) Its calculation is based on the items of the series.

(3) It is capable of further algebraic treatment.

(4) It is not affected very much by the fluctuation and sampling.

(5) It gives greater w.-ight to small items and to a single item cannot push up the value.

(6) It is very much helpful for relative studies like ratios and rates.

Demerits

(1) It is not easily understood.

(2) Its calculation is bit difficult.

(3) It is not useful for analysis of economic data as it gives more importance to smaller values.

(4) It gives a value, which usually does not exist in the series.

(7) WEIGHTED AVERAGE :

When all the items of a series are of equal importance, it would be proper to calculate the simple averages. But when the different items deserve different importance in a particular field of analysis, it would be desirable to find out the weighted average of the series. In such cases the simple average will fail to give the correct figure. For example, the average income of the employees of a factory obtained on the basis of simple average shall not fairly represent he income of all the employees. Because the elative importance of the figures of their income is not the same. The weighted average, in this case, will give a fair representation.

The process of finding out the weighted average does not require anything more than they are required in the case of a frequency distribution series. It will be enough to remember this much that the frequencies of the items will take the shape of the weights and accordingly the difficult formula on different averages will be modified.

Thus in the case of arithmetic average, Geometric average and that of Harmonic average, the formulas will be modified as follows :

(a) Weighted Average $= a = \frac{\sum mw}{\sum w} or x - \frac{\sum wdx}{\sum w}$

(b) Weighted Geometric Average $= g = AntiLog.\frac{\sum \log w}{\sum \log w}$

(c) Weighted H.M. $= h = r\frac{\sum r.w}{\sum w}$

In all the above cases it will be observed that "W" has taken the place of "f" only and that there is nothing more as change to be followed for calculating the weighted averages.

(8) RELATIVE IMPORTANCE AND LIMITATIONS OF THE ARITHMETIC AVERAGE, GEOMETRIC AVERAGE AND HARMONIC AVERAGE.

The relative importance and limitation of the arithmetic average, geometric average and harmonic average can be viewed in the light of the characteristics of an ideal measure of Central Tendency. These are highlighted as follows :

1. ***Capability of being rigidly defined :***

From this point of view all the three averages referred to are satisfactory.

2. *Capability of being easily followed :*

From this point of view it is only the arithmetic average, which is satisfactory. A layman will understand the implication of an arithmetic average. But the Geometric mean and Harmonic mean are very difficult to be understood as a common man of ordinary prudence.

3. *Capability of being based on all the observations:*

From this point of view all the three referred to satisfy the requirement. None of the averages are found out by ignoring any item of the series.

4. *Capability of algebraic treatment :*

From this point of view, all the three satisfy the requirement.

5. *Capability of not being affected by the values of the extreme items:*

From this point of view, arithmetic averages suffer from the drawback as much as it is greatly affected by the values of the extreme items of the series. But the geometric mean and the harmonic means are loss affected by the values of the extreme items.

6. *Capability of not being affected by the fluctuations of sampling :*

From this point of view all the three satisfy the requirement.

I. MEDIAN

According to Yule and Kendall, Median is defined as "the middle most or central value of the variables when the values are arranged in order of magnitude, or as the value such that greater and smaller values occur with equal frequency".

From the above definition it follows that Median is the middle item of a series arranged in ascending or descending order. It divides a series into two equal parts and takes the middle position in the series. Its value is always located with reference to its position in the middle part of the series and for this called as a positional average. Thus, if in a series there are 5 items viz. 5, 20, 40 and 125, the value of he median would be the value of the 3rd item i.e. 25. In case the number of items are in even numbers the value of the median would be determined as the half of the two middle values. Thus, if there are 6 items in a series viz. 5, 20, 25, 30, 40 and 45 the value of the median would be ½ of (25+30)=27.5.

From the above analysis the formula for determining the value of median can be represented as follows

2. ***Capability of being easily followed :***

From this point of view it is only the arithmetic average, which is satisfactory. A layman will understand the implication of an arithmetic average. But the Geometric mean and Harmonic mean are very difficult to be understood as a common man of ordinary prudence.

3. ***Capability of being based on all the observations:***

From this point of view all the three referred to satisfy the requirement. None of the averages are found out by ignoring any item of the series.

4. ***Capability of algebraic treatment :***

From this point of view, all the three satisfy the requirement.

5. ***Capability of not being affected by the values of the extreme items:***

From this point of view, arithmetic averages suffer from the drawback as much as it is greatly affected by the values of the extreme items of the series. But the geometric mean and the harmonic means are loss affected by the values of the extreme items.

6. ***Capability of not being affected by the fluctuations of sampling :***

From this point of view all the three satisfy the requirement.

I. **MEDIAN**

According to Yule and Kendall, Median is defined as "the middle most or central value of the variables when the values are arranged in order of magnitude, or as the value such that greater and smaller values occur with equal frequency".

From the above definition it follows that Median is the middle item of a series arranged in ascending or descending order. It divides a series into two equal parts and takes the middle position in the series. Its value is always located with reference to its position in the middle part of the series and for this called as a positional average. Thus, if in a series there are 5 items viz. 5, 20, 40 and 125, the value of he median would be the value of the 3rd item i.e. 25. In case the number of items are in even numbers the value of the median would be determined as the half of the two middle values. Thus, if there are 6 items in a series viz. 5, 20, 25, 30, 40 and 45 the value of the median would be ½ of (25+30)=27.5.

From the above analysis the formula for determining the value of median can be represented as follows

M = value of $\frac{(n+1)th}{2}$ items where 'M' stands for median and 'n' stands for the number items.

Of course, it should be remembered here that in case for a continuous series the additio of 1 to n in he formula is not required.

The detailed procedure of finding out the value of median under different series runs a under :

Determination of median from a simple series

For the determination of median from a simple series the following steps are to be followe :

1. Arrange the different items either in ascending or descending order

2. Put the formula : M = value of $\frac{(n+1)th}{2}$ item and get the median item.

3. Locate the value from the series with reference to the median item.

Example - 1

From the following observation determine the value of the median

15, 3, 21, 2, 7, 81, 31, 37, 50, 48, 72.

Solution

Determination of median

Items arranged in ascending order :

2, 3, 7, 15, 21, 31, 37, 48, 50, 72, 81

By the formula we have :

M = value of $\frac{(n+1)th}{2}$ item

= Value of $\frac{(11+1)th}{2}$ item

= Value of 6^{th} item

Hence value of median item = 31.

Example – 2

Determine the value of Median from the following data

Wages in Rs. 50, 40, 60, 80, 30, 20

Solution

Determination of median

Wages arranged in descending order

Rs. 80, 60,50, 40, 30, 20.

By the formula we have :

M = value of $\frac{(n+1)th}{2}$ item = value of $\frac{(6+1)th}{2}$ item

= Value of 3.5th item = ½ of (value of 3rd + value of 4th item)

= ½ of (50 + 40) = 45.

Determination of median from a discrete series

The determination of median from a discrete series will involve the following steps :

1. Arrange the items together with their corresponding frequencies either in ascending or descending order.
2. Find out the cumulative frequencies in a separate column.
3. Put the formula : M = $\frac{(n+1)th}{2}$ item and get the median item.
4. Locate the value of Median item in the series with reference to cumulative frequency within which the median item falls

Example – 3

Ascertain the value of median from the following series :

Marks : 30, 80, 40, 70, 50, 60, 25.

No. of students : 5, 15, 6, 40, 30, 21.

Calculation of median

Marks arranged in ascending order	No.of Students	c.f.
25	21	21
30	5	26
40	15	41
50	40	81
60	30	111
70	6	117
80	3	120

By the formula we have :

M = value of $\frac{(n+1)th}{2}$ item = value of $\frac{(120+1)th}{2}$ item = value of 60.5th item

Which lies against the cumulative frequency of 81.
Hence the value of median = 50.

Calculation of median from a continuous series

In a continuous series, the values are always stated in an orderly manner that is either in ascending or in descending order. Thus, there will be no more necessity of arranging them again. However, if the class intervals are given in an inclusive manner, viz. 10-19, 20-29,etc. they are to be adjusted according to the exclusive manner, viz. 9.5-19.5, 19.5-29.5 etc. This is necessary because in case of a continuous series the value of median will be located from the median class interval by the formula of interpolation and for this exclusive class limits of the median class will be necessary.

Subject to above adjustments the following steps are to be followed in determining the value of median from a continuous series.

1. Find out the cumulative frequencies.

2. Put the formula : M = value of $\frac{(n)th}{2}$ item and there by get the median item and locate its cumulative frequency in the c.f. column.

3. Locate the median class with reference to the c.f. thus ascertained.

4. Put the formula of interpolation as follows and ascertain the particular value of median

$$M = L_1 + \frac{L_2 - L_1}{f_1}(m - c)$$

Where, 'M' stands for median, 'L_1' and 'L_2' for the lower limit and the upper limit of the median class, 'f_1' for the frequency of the median class. 'm' for the median items and 'c' for the cumulative frequency of the class proceeding to the median class.

Example – 4

Compute the median from the following data :

Class interval	7.5 – 12.5	12.5 – 17.5	17.5 – 22.5	22.5 – 27.5	27.5 – 32.5
Frequency	4	6	5	3	2

Solution

Computation of the Median

Class intervals	f	c.f
7.5-12.5	4	4
12.5-17.5	6	10
17.5-22.5	5	15
22.5-27.5	3	18
27.5-32.5	2	20

M = Value of $\frac{n}{2}$ item = Value of $\frac{20}{2}$ = Value of 10th item

This lies in the 12.5 – 17.5 class interval. Thus, by putting the formula of interpolation we have :

$$M = L_1 \frac{L_2 - L_1}{f_1} (m - c) = 125.5 + 5/6(10 - 4) = 17.5$$

Merits and Demerits of Median

MERITS

(1) It has a rigid definition and so it satisfies the first requirement of an ideal average.

(2) It is easily understood and easily calculated.

(3) It is not affected by the values of the extreme items.

(4) In an open end series there will be no difficulty in ascertaining its value. This means, even if the values of the extreme items are not known, it can be calculated if the number of items is known.

(5) It gives best results in a study of those phenomena, which are incapable of quantitati measurement, for example greatness.

(6) It can be determined graphically.

DEMERITS

(1) If fails to give a representative figure, when there are wide variations between the value different items.

(2) It is not suitable for further algebraic treatment. For example, we cannot find out the tot values of the items, if we know their number and median.

(3) In case of continuous series, it is determined by interpolation with the assumption that the frequencies or the class intervals are uniformly spread over their values in the clas interval. This may not be true in most of the cases.

(4) It does not give the greater importance to the big or small items as and when require because it ignores the extreme items.

(5) It is more likely to be affected by the fluctuations of sampling.

(6) It requires the arrangement of items in ascending or descending order, which is somethin tedious.

2. QUARTILES

Meaning

Quartile is a positional measure of dispersion, which divides a series into four equal parts It is a fact that one point can divide a series into two parts. In other words, to get two parts from a series, we have to divide it by one point. Therefore, to get four equal parts from a series as prescribed by the quartile we have to divide the series by three points. From this it follows that there are three quartiles, which divide a series into four equal parts. These three quartiles are signified as Q_1 or the 1st Quartile, Q_2 or the 2nd quartile and Q_3 or the third quartile. Q_1 and Q_3 are also otherwise known as lower quartile and upper quartile respectively. Q_2 being the middle quartile is equal to the value of the median discussed in the proceeding paragraphs.

Symbolically the above three quartiles are stated as under :

Q_1 = value of $\frac{(n+1)th}{4}$ item

Q_2 = value of $\frac{2(n + 1)th}{4}$ item

Q_3 = value of $\frac{3(n + 1)th}{4}$ item

It should, however be remembered that as in case of median, the addition in the above formula will not be necessary in the case of continuous series

Procedure of Calculation :

The procedure of calculating the different quartiles will remain the same as it was in case of median. The following illustration would clarify the procedure.

Example - 6

From the following variable determine the value of quartiles.

Variables : 1, 11, 3, 4, 5, 7, 8, 15, 2.

Solution :

Determination of Quartilities

Q_2 = value of $\frac{2(n + 1)th}{4}$ item

= value of $\frac{2(9 + 1)th}{4}$ item = value of 5th item = 5

Q^3 = value of $\frac{3(n + 1)th}{4}$ item

= value of $\frac{3(9 + 1 th}{4}$ item = value of 7.5th item

= value of 7th item + ½ (value of 8th – value of 7th item)

= 7 + ½ (11 – 7) = 9

Hence the value of Q_1, Q_2, Q_3 are 2.5, 5 and 9 respectively.

Example - 7

From the following series find out the values of lower quartile and the upper quartile ?

Size of the items :	0-5	5-10	10-15	15-20	20-25
Frequencies	5,	15,	30,	25,	10

Size of the item	(f)	(c.f)
0-5	5	5
5-10	15	20
(Q₁) 10-15	30	50
(Q₃) 15-20	25	75
20-25	10	85

Q_1 = Value of $\frac{(n)th}{4}$ item = value of $\frac{(85)th}{4}$ item = value of 21.25th item which lies in 10-15 class interval standing against the c.f. group of 50, By putting the formula of interpolation :

$$Q_1 = L_1 + \frac{L_2 - L_1}{f}(q_1 - c) = 10 + \frac{15-10}{30}(21.5-20)$$

$$= 10 + \frac{5}{30}(1.5) = 10 + \frac{7.5}{30} = 10.25$$

Q_3= value of $\frac{3(n)th}{4}$ item = value of $\frac{3(85)th}{4}$ item

= Value of 64.5th item which lies in 15.20 class interval standing against the c.f. of 75 group. Putting the formula of interpolation :

$$Q_3 = L_1 + \frac{L_2 - L_1}{f_1}(q_3 - c) = 15 + \frac{20-15}{25}(64.5-50)$$

$$= 15 + \frac{5}{30}(14.5) = 15 + 2.9 = 17.9$$

Hence the value of the lower quartile and the upper quartiles respectively are 10.25 and 17.9.

3. DECLIES

A decile is another positional measure of dispersion, which divides a series into 10 equal parts. As such there are nine deciles in a series viz. D_1, D_2, D_3........D_2. Symbolically representation of the Deciles can be stated as follows :

D_1 = Value of $\frac{(n+1)th}{10}$ item

D_2= Value of $\frac{9(n+1)th}{10}$ item.

Procedure of calculation :

The procedure of the calculation of the values of the various deciles will be the same as it was in case of quartiles and median. The following example would give a brief idea of its calculation.

Example – 8

Determine the values of the lower and the upper deciles from the following observations

Variables : 2,3,18,7,4

Frequencies : 10, 12, 5, 6, 16

Solution :

Determine of the lower and upper deciles

Variables arranged in ascending order	(f)	(c f)
3	12	12
4	16	28
5	10	38
7	6	44
18	5	49

D_1 = Value of $\frac{(n+1)th}{10}$ item = value of $\frac{(49+1)th}{10}$ item

= Value of 5th item = 3 which stands against the c.f. of 12

D_2 = Value of $\frac{9(n+1)th}{10}$ item = Value of $\frac{9(49+1)th}{10}$ item

= Value of 45th item = 18, which stands against the c.f. of 49.

Hence, the values of D_1 and D_2 are 3 and 18 respectively.

4. MODE

Mode is defined as the most common item of a series. In short, this refers to the value of a variable against which the number of items or the frequency is the maximum. Thus, the calculation of mode entirely depends upon the distribution of frequencies. The item or the value, which has the maximum frequency, is taken to be the model value of the series. However, in some cases, it so happens that more than one item possesses the maximum frequencies of size. In such cases different values are taken to be the model values of the series and the mode is said to be ill defined here. However to day away with such indefiniteness, an attempt is made to determine the mode by the method of grouping the frequencies in various groups. Such groupings of frequencies are usually made in twos and threes i.e. for two times and three times respectively. After grouping the frequencies in various rows, the maximum totals are thickly marked and then an analysis table is prepared to find out the particular value against which the maximum frequencies cluster from maximum times. In case of continuous series, however, the above process would reveal the modal class instead of the modal value. From the modal class thus located the value of the mode will be determined by the formula of interpolation, which is as follows.

$$Z = L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2}(L_2 - L_1)$$

Where, Z stands for the mode, L_1 for the lower limit and L_2 for the upper limit of the modal class, f_1 for the frequency of the modal class, f_0 for the frequency of the class succeeding the modal class.

Example – 8

Determine the values of the lower and the upper deciles from the following observations

Variables : 2,3,18,7,4

Frequencies : 10, 12, 5, 6, 16

Solution :

Determine of the lower and upper deciles

Variables arranged in ascending order	(f)	(c f)
3	12	12
4	16	28
5	10	38
7	6	44
18	5	49

D_1 = Value of $\frac{(n+1)th}{10}$ item = value of $\frac{(49+1)th}{10}$ item

= Value of 5th item = 3 which stands against the c.f. of 12

D_2 = Value of $\frac{9(n+1)th}{10}$ item = Value of $\frac{9(49+1)th}{10}$ item

= Value of 45th item = 18, which stands against the c.f. of 49.

Hence, the values of D_1 and D_2 are 3 and 18 respectively.

4. MODE

Mode is defined as the most common item of a series. In short, this refers to the value of a variable against which the number of items or the frequency is the maximum. Thus, the calculation of mode entirely depends upon the distribution of frequencies. The item or the value, which has the maximum frequency, is taken to be the model value of the series. However, in some cases, it so happens that more than one item possesses the maximum frequencies of size. In such cases different values are taken to be the model values of the series and the mode is said to be ill defined here. However to day away with such indefiniteness, an attempt is made to determine the mode by the method of grouping the frequencies in various groups. Such groupings of frequencies are usually made in twos and threes i.e. for two times and three times respectively. After grouping the frequencies in various rows, the maximum totals are thickly marked and then an analysis table is prepared to find out the particular value against which the maximum frequencies cluster from maximum times. In case of continuous series, however, the above process would reveal the modal class instead of the modal value. From the modal class thus located the value of the mode will be determined by the formula of interpolation, which is as follows.

$$Z = L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2}(L_2 - L_1)$$

Where, Z stands for the mode, L_1 for the lower limit and L_2 for the upper limit of the modal class, f_1 for the frequency of the modal class, f_0 for the frequency of the class succeeding the modal class.

The following example would illustrate the calculation of the model.

Example - 9

Ascertain the value of the mode from the following frequency distribution.

Wages (in Rs.) :	30,	40,	50,	60,	70,	80,	90
Frequency :	5,	10,	20,	25,	3,	15,	30

Solution

Calculation of mode by the method of grouping

Wages Rs.	(f)	Grouping in twos Grouping in threes				
	(1)	(2)	(3)	(4)	(5)	(6)
30	5	15				
'0	10			35		
			30			
50	20	45				
60	25				55	48
70	3	18	28	43		
80	15	18		48		
90	30		45			

Analysis Table
Columns of values of the variables

	30	40	50	60	70	80	90	
1							*	
2			*	*				
3							*	
4				*	*	*		
5			*	*	*			
6				*	*	*		
Total			1	3	4	2	2	2

From the above analysis table it appears that the wage value of Rs.60 in having the maximum frequency for maximum lines. Hence the model wage in this case would be Rs.60.

Example - 10

Determine the value of mode from the following series

Class interval	0-5	5-10	10-15	15-20	20-25
Frequencies	22	18	45	50	20

Solution

Determine of mode by the method of grouping

Class intervals	(f)	Grouping in twos		Grouping in Threes		
	(1)	(2)	(3)	(4)	(5)	(6)
0-5	22	40		85		
5-10	18		63			
10-15	45	95			113	
15-20	50		70			115
20-25	20					

Analysis Table

Columns	Class intervals					
	0-5	5-10	10-15	15-20	20-25	
1				I		
2			I	I		
3					I	I
4		I	I	I		
5			I	I	I	
6				I	I	I
Total		1	2	2	5	42

From above analysis table it appears that the mode lies in the class interval of 15-20.

By putting formula of the interpolation we have :

$$Z = L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2}(L_2 - L_1)$$

$$= 15 + \frac{50 - 45}{2 \times 50 - 45 - 20}\frac{(20 - 15)}{(20 - 15)}$$

$$= 15 + \frac{5}{35} \times 5 = 15 + \frac{5}{77} = 15 + .71 = 15.71$$

Hence, the value of the mode is 15.71 (approx)

Note - It should be remembered that where by the method of grouping, there appears more than one value of the class interval as the modal value of the mode should be determined by the following equation which is based on the empirical relationship between mean, median and mode.

$Z = 3m - 2a$

For this purpose the values of the mean and median are to be found out first.

Merits and Demerits of Mode

MERITS

(1) It is easy to calculate. In most of the cases of discrete series mode can be calculated even by inspection

(2) It is easily understood by a common man. Mode is an average, which people use in their day-today expressions. The average size of the shoes and averages size of the garments are the examples of common use.

(3) Unlike arithmetic average it does not give a value, which is not found in a series.

(4) It is not affected by the values of the extreme items of series

(5) It does not need all the items of a series. If the point of a norm or maximum concentration is known, it will be enough for the determination of the mode.

DEMERITS

(1) It is ill defined and in certain cases of bimodal, trimodal etc. it is indeterminate and indefinite

(2) Its calculation is not based on all the observations of the series.

(3) It is not capable of further algebraic treatment.

For Median, Quartile and Decile

(1) In case of individual and discrete series, arrange the items in an ascending order, (descending order may be applicable only in case of Median).

(2) In case of the discrete and continuous series only, find out the cumulative frequencies.

(3) Apply the appropriate formula and locate the required value there by with reference to the item thus ascertained.

(4) In case of continuous series, after ascertaining the particular item as above, apply the formula of interpolation to determine the exact value from he concerned class interval.

For Mode

(1) Mode cannot be calculated from the individual series. If at all it is asked to calculate the mode from such a series, it will be necessary to convert the series first either into a discrete or a continuous series.

(2) If the maximum frequency is far above the next greater frequency, ascertain the modal item or the classes by mere inspect or the maximum frequency in doubtful cases applies the method or grouping the frequencies and preparing the analysis table.

(3) In continuous series the formula of interpolation will have to be applied.

For all the above

(1) If the continuous series is given in an inclusive manner. It will be the first duty to convert them into the exclusive manner.

SELF-TEST – 9

1. From the following series determine the value of the median, lower quartiles and upper declies : 5, 9, 18, 35, 42, 15, 7, 3.

2. Determine the mode, median, upper quartile and lower decile from the following series

   Wages (in Rs.) 1, 2, 3, 4, 5, 6, 7, 8, 9

   No. of workmen   8, 10, 11, 1, 20, 25, 15, 9, 6.

3. Calculate the mode and the median from the following

Value	Frequencies
0-4	328
5-9	350
10-19	720
20-29	664
30-39	598
40-49	524
50-59	378
60-69	244

MEASURES OF DISPERSION

(Range, Inter-Quartile Range, Mean Deviation)

INTRODUCTION :

By this time you know why it is necessary to tabulate and classify statistical series and to condense them into a single figure called average. The average has its own limitations and even an ideal average can represent a series only " as best as a single figure can" No doub average have a very great utility in statistical analysis but they fail to reveal the entire story of phenomenon. Averages alone cannot adequately describe a set of observations, unless all the observations are the same. There may be a dozen series whose central values may be the differ from each other in a number of ways.

Suppose there are three series of nine items each as follows

Series A	Series B	Series C
30	26	1
30	27	9
30	28	10
30	29	20
30	30	30
30	31	40
30	32	45
30	33	55
30	34	60
270	270	270

Arithmetic Mean

30	30	30

Since the arithmetic mean is the same in all the three series one is likely to conclude tha these series are alike in nature. But a close examination shall reveal that these series diff widely from one another. In the first series the mean or average is 30 and the value of all items identical. The items are not all scattered, and the mean though the mean is 30 yet all the items have different values. But they are not very much scattered as the minimum value of the series 26 and the maximum is 34. In this case also mean is a good representative of the series. In the third series the mean is also-30 and the values are very widely scattered and the mean is 3 times of the smallest value of the series and half of the maximum value. Obviously the averag here does not satisfactorily represent the individual items in the group. In all these three serie the averages are identical (i.e., 30) and yet the series widely differ from each other in the formation. The scatter in the first series is nil, in the second series it varies within a small range while in the third case the values range between a very big span and they are widely scattered It is obvious that in order to get a better idea about the composition of a series we should stud the extent of the scatter around the average. The name given to this scatter is dispersion.

Dispersion refers to the variability in the size of items, it indicates that the series of item in a series is not uniform. The value of various items differs from each other. If the variation is substantial dispersion is said to be considerable and if the variation is little dispersion is insignificant. This is the general sense in which the term dispersion is used.

The terms dispersion not only gives a general impression about he variability of a series, t also a precise measure of this variation Usually in a precise study of dispersion, the deviations the size of items from a measure of central tendency are found out and these deviations are veraged to give a single figure representing the dispersion of the series. This figure can be ompared with similar figures representing other series

verages of the Second Order :

Since for a precise study of dispersion we have to average deviations of the values of the arious items, from their average, various measures of dispersion are called AVERAGE OF HE SECOND ORDER. Measures of Central tendency are called averages of the first order.

bsolute and Relative Dispersion :

Measures of dispersion may be either absolute or relative. Absolute measures of ispersion are expressed in the some statistical unit in which the original data are given, uch as rupees, kilograms, tonnes etc. If we calculate dispersion of a series relating to he income of a group of persons in absolute figures, it will have to be expressed in the ame unit in which the original data are, say rupee. Thus we can say that the income of a roup of person is Rs.150/- per month and the dispersion is Rs.30/-. This is called absolute ispersion. If, on the other had, dispersion is measured as a percentage or ratio of a neasure of absolute dispersion to an appropriate average it is called Relative Dispersion. t is not expressed in the unit of the original data. In the above case the average income vould be referred to as Rs.150/- and the relative dispersion as 2 or 20%. He relative neasure of dispersion is sometimes called a coefficient of dispersion, because' coefficient neans a pure number that is independent of the unit of measurement".

In a comparison of the variability of two or more series, it is the relative dispersion that has o be taken into account as the absolute dispersion may be erroneous or unfit for comparison if he series are originally in different units such as pounds of sugar versions tonnes of sugarcane etc.

Properties or Qualities of a Good Measure of Dispersion :

A good measure of dispersion should possess the same qualities which a good measure f central tendency possesses. A good measure of dispersion should possess, as far as possible, he following properties or qualities :

(i) It should be simple to understand.

(ii) It should be easy to compute.

(iii) It should be rigidly defined.

(iv) It should be based on each and every item of the distribution.

(v) It should be capable of further algebraic treatment.

(vi) It should be affected much by the fluctuations of sampling

(vii) It should not be unduly affected by extreme items of a series.

Types of Measures of Dispersion :

There are various types of measures of dispersions; of them the following are in common use.

(1) Range (2) Inter Quartile Range (3) Semi-Inter-Quartile Range or Quartile Deviation (4) mean Deviation (5) standard Deviation.

1. RANGE

Range is a measure of dispersion, which can be determined in both the Absolute and the relative way.

Absolute Range :

Absolute Range is the difference between the values of the two extreme items of a series. Symbolically Range = L – S where, L- largest value. S – Smallest value. Thus, If, in a series the minimum and maximum values are 5 and 40 respectively, the absolute range would be: 40 – 5 = 35, If the values are stated in a continuous manner, the range would be the difference between lower limit of the lowest class and the upper limit of the highest class.

Range as calculated above is unfit for purposes of comparison, if the distributions are in different units. For example, the range of the weights of śtudents cannot be compared with the range of their height measurements, as the range of weights would be in Kgs and that of heights in centimeters.

It should be remembered here, that the calculation of range has no relation with the frequencies of a series.

Relative or the Co-efficient of Range

If absolute range is divided by the sum of the extreme items, the resulting figure is called relative or the co-efficient of the range. Coefficient of Range = $\frac{L-S}{L+S}$ Thus on the above example the coefficient of range would be:

$\frac{40-5}{40+5} or \frac{35}{45} or \frac{7}{9}$ or 78% (approx.) Such type of range would or helpful for making comparative study. Coefficient of range is also called "The Ratio of the Range" or "The Coefficient of the Scatter".

Merits and Demerits of Range

MERITS

(1) It is easily calculated.

(2) It is readily understood by a common men even.

(3) It is very much useful in the field of Quality Control of manufactured goods.

DEMERITS

(1) It is greatly affected by the fluctuation of sampling. It its value is never stable and it varies from sample to sample.

(2) A single variation in the value of an extreme items would affect the value of the range.

(3) It is not based on all observations of the series.

(4) It is not capable of further algebraic treatment.

2. INTER QUARTILE RANGE

Just like range an inter-quartiles range can be studied in both the absolute and the relative ways.

Absolute inter-Quartile Range

Absolute Inter-quartile range is the difference between the two extreme Quartiles of a series. In other words, it is the difference between the Q_3 and the Q_1 $IQR = Q_3 - Q_1$. Thus to find out the value of Inter-Quartile Range it is required to ascertain first the values of the lower quartile and the upper quartile.

Relative or the Coefficient of Inter Quartile Range

Coefficient of Inter Quartile Range is the ratio of Inter Quartile Range to the sum of the two quartiles. Symbolically this may be stated as follows :

$$\text{Coefficient of Inter Quartile Range} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Merits and Demerits

MERITS

(1) it is easy to calculate provided that procedure of calculation of Quartiles is felt easier.

(2) It is readily understood.

(3) Unlike Range, it is not affected by the values of the extreme items.

(4) It gives a fair measure of variability as 50% of the values of a variable lie between the two quartiles.

DEMERITS

(1) It is affected by the fluctuation of sampling.

(2) It is not based on all the observations of a series.

(3) It is a measure of location and so its value is not always stable.

(-) A change in the value of an extreme item may affect its value.

(5) It is not capable of further algebraic treatment.

3. QUARTILE DEVIATION

Quartile deviation is otherwise known as the Semi-Inter-Quartile Range. It can be calculated in both the ways as follows :

(1) Absolute Quartile Deviations is the mid point of the Inter-Quartile Range in other words it is one half of the difference between the upper Quartile and the lower quartile. Symbolically, it is stated as follows :

Quartile Deviation = $\frac{Q_3 - Q_1}{2}$. Where, Q_3 stands for upper quartile and Q_1 for the lower quartile.

2. RELATIVE OR THE CO-EFFICIENT OF QUARTILE DEVIATION

It is ascertained by dividing the difference between the Q_3 and Q_1. Symbolically Co efficient of Q.D. = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$

Merit and Demerits

Merits

(1) It is easy to calculate, as its calculation does not involve any mathematical intricacies.

(2) It is easy to understand even for a common man.

Demerits

(1) It is not based on al the observations of the data.

(2) It is not capable of further algebraic treatment.

(3) It is affected to a great extent by the fluctuations of sampling.

(4) A single change in the value of an extreme item will affect its value greatly.

Note

It should be remembered here that the three measures of dispersions, discussed above, i.e., Range, Inter-Quartile Range and the Quartile Deviation are known otherwise as the measures of dispersion by the method of limits. These are known so because their calculations are made purely on the basis of the extreme limits of the different values. Thus, when it is asked to find out the Dispersion by the method of limits-attempt should always be made to find out all these Dispersions.

The following examples would show how the dispersions are calculated by the method of limits under different series.

Example

The following are the marks obtained by a batch of 9 students in a certain examination.

Roll No :1, 2, 3, 4, 5, 6, 7, 8, 9

Marks : 68, 49, 32, 21, 54, 38, 59, 66, 41

Find out the dispersion and their co-efficients by the method of limits.

Solution

This is a case of simple series. Before ascertaining the values of the various determinan it is necessary to rearrange the series in an ascending order. Thus the calculation will procee as follows :

Calculation of Range. I. Q. R., Q.D. and their coefficients.
Marks secured in ascending order : 21, 32, 38, 41, 49, 54, 59, 66, 68

1. Range = x-y where 'x' stands for the maximum value and 'y' for the minimum value.

= 68.21 the minimum value = 47

2. Coefficient range = $\frac{x-y}{x=y}=\frac{68-21}{68+21}=\frac{47}{89}=.53$ (approx)

3. I.Q.R. = $Q_3 - Q_1$ = 62.5-35 = 17.5 Where I.Q.R. stands for Quartile Range, Q_3 for upper an Q_1 for lower quartile.

4. Coefficient of Q.I.R. = $\frac{Q_1 - Q_1}{Q_3 + Q_1}=\frac{62.5-35}{62.5+35}=\frac{17.5}{97.5}$=.18 (approx)

Here Q_3 = Value of item

= Value of item = Value of 7.5th item

= Value of 7th item +1/2 (value of 8th item value of 7th item

= 59+1/2 (66+59) = 59+3.5 = 62.5

Again, Q_1 = Value of the item = Value of item = value of 2.5th item = Value of 2nd iten +1/2 (value of 3d item – value of 2nd item) = 32+1/2 (38 – 32) = 32+3=35

Example – 2

Compute the dispersions and their coefficients by the method of limits from the followinc series :

No.of colds suffered :	0,	1,	2,	3,	4,	5,	6,	7,	8,	9
Persons :	15,	46,	91,	162,	100,	95,	62,	26,	13,	2

Solution

Calculation of Dispersion and their coefficients by the method limits

(m)	(f)	(c.f.)
0	15	15
1	46	61
2	91	152
3	162	314
4	110	424
5	95	519
6	82	601
7	26	627
8	13	640
9	2	642
	n=642	

2. Coefficient range = $\frac{x-y}{x=y} = \frac{68-21}{68+21} = \frac{47}{89} = .53$ (approx)

3. I.Q.R. = $Q_3 - Q_1$ = 62.5-35 = 17.5 Where I.Q.R. stands for Quartile Range, Q_3 for upper and Q_1 for lower quartile.

4. Coefficient of Q.I.R. = $\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{62.5-35}{62.5+35} = \frac{17.5}{97.5}$ =.18 (approx)

Here Q_3 = Value of item

= Value of item = Value of 7.5th item

= Value of 7th item +1/2 (value of 8th item value of 7th item

= 59+1/2 (66+59) = 59+3.5 = 62.5

Again, Q_1 = Value of the item = Value of item = value of 2.5th item = Value of 2nd item +1/2 (value of 3d item – value of 2nd item) = 32+1/2 (38 – 32) = 32+3=35

Example – 2

Compute the dispersions and their coefficients by the method of limits from the following series :

No.of colds suffered :	0,	1,	2,	3,	4,	5,	6,	7,	8,	9
Persons :	15,	46,	91,	162,	100,	95,	62,	26,	13,	2

Solution

Calculation of Dispersion and their coefficients by the method limits

(m)	(f)	(c.f.)
0	15	15
1	46	61
2	91	152
3	162	314
4	110	424
5	95	519
6	82	601
7	26	627
8	13	640
9	2	642
	n=642	

1. Range = x − y = 9 − 0 = 9

2. Coefficient of range = $\frac{x-y}{x+y} = \frac{9-0}{9+0} = 1$

3. I.Q.R. = $Q_3 - Q_1 = 5 - 3 = 2$

4. Coefficient of I.Q.R. = $\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{5-3}{5+3} = 2 = .25$

5. Q.D. $\frac{Q_3 - Q_1}{2} = 5 - 3 = 1$

6. Coefficient of Q.D $\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{5-3}{5+3} = \frac{2}{8} = .25$

(Here, Q_3 = Value $3\frac{(n+1)}{4}$ item = Value of $3\frac{(642+1)th}{4}$ item = value of 482nd item = 5. Q_1 = Value of $3\frac{(n+1)th}{4}$ item = value of $3\frac{(642+1)th}{4}$ item = Value 161st item = 3.)

Example – 3

The following table gives weights of one hundred persons.

Compute the coefficient of dispersions by the method of limits

Solution

Calculation of the coefficients of dispersion by the method of limits

Weight in Lbs.	No. of persons
85-95	4
95-105	13
105-115	8
115-125	14
125-135	9
135-145	16
145-155	17
155-165	9
165-175	8
175-185	2

Weight in Lbs	No.of Persons	(m)	(f)	(c.f)
		85-95	4	4
		95-105	13	17
		105-115	8	25
		115-125	14	39
		125-135	9	48
		135-145	16	64
		145-155	17	81
		155-165	9	90
		165-175	8	98
		175-185	2	
			n = 100	

1. Co-efficient of range = $\frac{x-y}{x-y} = \frac{185-85}{185+85} = \frac{100}{270} = .37$ (approx.)

2. Coefficient of I.Q.R. = $\frac{Q_3-Q_1}{Q_3+Q_1} = \frac{151.47-115=36.47}{151.47+115=66.47} = .14$ (approx.)

3. Coefficient of Q.D.= $\frac{Q_3-Q_1}{Q_3+Q_1} = \frac{151.47-115}{151.47+115} = .14$ (as above)

Where 'x' stands for the lower limits of the lowest class and 'y' for the upper limit of the highest class.

Q_3 value of $3\frac{(n+1)th}{4}$ item = Value of $3\frac{(100)}{4}$ item = Value of 75th item = this lies in the 145-155 class interval.

Thus, $Q_3 = 1_1+1_2-1_1\,(q_3-c) = 145 + 155 - 145\,(75-64)\,f_1$

$$=145\frac{10}{17}\times 11 = 145+\frac{110}{17} = 145+6.47 = 151.47$$

Q_1 = Value of $\frac{n}{4}$ the item = Value of $\frac{100}{4}$ the item = Value of 25th item.

This lies in 105- 115 class interval.

Thus, $Q_1 = 1_1 + \frac{1_2-1_1)th}{1_1}(Q_1-c) = 105+\frac{115-105}{8}(25-17)$

$$=105+\frac{10}{4}\times 8 = 115.$$

4. MEAN DEVIATION

Mean deviation is otherwise known as the arithmetic average of the above deviations. It is a mathematical measure of dispersions, which is calculated in the line of the calculation of arithmetic average. It is defined as the arithmetic average of the deviations of various items from a measure of central tendency (either mean, median or mode), thus Mean deviation is always calculation, attempt is made to know as the what is the average of the deviations of the various items of the series. For this purpose, the separate deviations of each of the items of the series from an average are first found out. In finding out such deviations the plus and minus signs are completely ignored and mean, median or mode is taken as the central value. Of the three averages, Median is considered as the most suitable for calculation the deviation. Mode sometimes being ill defined and interminate is not considered suitable for the purpose. However, in practical fields, the deviations are usually taken either from Median or Mean. The symbolical presentations of mean deviation are as follows.

1. Where deviations are taken from mean.

$\delta.\frac{\sum d}{n}$ Where Σ stands for the mean deviation from mean, d for the deviations from the mean, and n for the number of items

2. Where deviations are taken from median.

$\delta m\frac{dm}{n}$ Where m stand for the mean deviation from medium, dm for the deviations from the medium and n for the number of items.

3. Where deviations are taken from Mode.

$\delta=\frac{dz}{n}$ Where z stands fro the mean deviation from mode, dz for the deviations from mode & n is the number of items.

It should be noted here that the above formulas of mean deviation are for the absolute results. For comparative studies, such results may not be helpful and so it will be necessary to find out the co-efficient of the mean deviation. The coefficient of mean deviation would be calculated as a ratio of the mean deviation to the particular average from which the deviations have been calculated. Thus the symbolic presentation of the coefficients of mean deviation in each of the above three cases would be:

(1) When deviations are taken from Mean :

Coefficient of $\delta = \frac{\delta}{a}$

(2) When deviations are taken from Median :

Coefficient of $\delta m = \underline{\delta m}$

(3) When deviations are taken from Mode :

Coefficient of $\delta z = \delta z / z$

The details procedure of its calculation under different series and under different methods or discussed as under.

(1) Calculation of mean deviation and its coefficient in a series of individual observations. In case of an individual series, mean deviation whether form mean or from Median can be calculated under two methods, vis : (1) Direct Method and (2) Short-cut Method.

Direct Method :

Under direct method the following steps are to be taken into effect .

(1) Find out the value of mean or median

(2) From the mean or median find out the deviations of items ignoring plus and minus signs.

(3) Get the total of the deviations and divide the same by the numbers of items. This will give the required result of mean deviation.

(4) To obtain the coefficient of mean deviation divide the mean deviation so obtained by the mean or the median from which the deviations have been found out.

Short cut Method

Under this method the following steps are necessary :

(1) Arrange the values in an ascending order.

(2) Ascertain the value of the mean or median.

(3) Aggregate the values of the items whose values are more than the value of the mean or the median., and represent it by 'ay' or 'mx' (as the case may be)

(4) Aggregate the values of the items whose values are less than the value of the mean or the median and represent it by 'ar' or 'mx' (as the case may be)

(5) Substitute the values in the following formula and get the result :

(i) $\delta \frac{ay - ax}{n}$

(ii) $\delta m \frac{my - mx}{n}$

Where δ stand fro the mean deviation from mean, 'ay' for sum of the values 'greater than mean 'ax' for sum of the values smaller than mean, 'n' for the number of items, for mean deviation from median, 'my' for sum of the values greater than the median and 'mx' for sum of the values smaller than the median.

The following illustration would clarify the calculations.

Example – 4

From the following marks obtained by a batch of 7 students calculate the mean deviation and its coefficient under both the methods :

Roll nos.	1	2	3	4	5	6	7
Marks :	78	59	42	31	64	48	69

Solution

Under Direct Method

Calculation of mean deviation from mean

Marks	Deviation from mean 56 ('+' and '-' signs are ignored)
78	22
59	3
42	14
31	25
64	8
48	8
69	13
Σm=391	Σd=93

δ=Σm=391=56 (approx.)

$\delta = \frac{d}{n} = \frac{93}{17} = 13$ (approx.)

Under Short-cut Method :

Calculation of mean deviation from mean

Marks arranged in ascending order
(m)
31
42
48
48
59
64
69
78
Σm=391,n=7

a= $\frac{\sum m}{m} = \frac{391}{7} = 56$ (approx.)

'ay' or the sum of the values greater than mean :

(59-56)+64+69+78 = 241.

'ax', or sum of the values smaller than the mean :

31+42+48=121.

Thus = $\frac{ay - ax}{n} = \frac{214 - 121}{n} = \frac{93}{7} = 13$ (approx)

N.B.: In getting the value of 'ay' 56 has been deducted from the first great items 59 as the case 56 is the average item which will remain in the middle and will not be included in either of the groups. The excess of 59 over 56 will be added with higher group i.e. ay.

Calculation of mean deviation from median

Marks arranged in ascending order	Deviation from m = 59 (signs ignored)
31	28
42	17
48	11
59	0
64	5
69	10
n=7	Σdm=90

M = value of (n+1)th item = value of (7+1)th item = 59

$\delta=\frac{\sum dm}{n}=\frac{90}{7}13$ (approx.)

Under Short cut Method

'my' for sum of the items greater than median :

64+96+78=211

'mx' for sum of the items smaller than median

Thus, $\delta m=\frac{my-mx}{n}=\frac{211-121}{7}=\frac{90}{7}=13$ (approx)

Coefficient of mean deviation from mean

Coefficient $\delta m=\frac{\delta}{a}=\frac{3}{59}=.23$

Coefficient of mean deviation from median :

Coefficient $\delta m=\frac{\delta m}{m}=\frac{13}{59}=.22$

ii) Calculation of Mean Deviation in Discrete Series :

In a discrete series, mean deviation can be calculated either from mean, median or mode and under any of the two methods: (1) Direct and (2) Short-cut Method.

Direct Method :

(1) Ascertain the value of mean, median or the mode (as the case may be)

(2) From the mean, median or the mode thus ascertained find out the deviations of each of the items ignoring plus and minus signs.

(3) Multiply the deviations with the corresponding frequencies.

(4) Get the total of the frequency column and that of the product of the deviations and the frequency column.

(5) Divide the product-total by the total of the frequency and get the required value of mean deviation.

Under Short-cut Method

Calculation of mean deviation from mean

(m)	(f)	(dx) (m-20)	(fdx)	dx (± ignored)	fdx (± ignored)
10	5	-10	-50	10	50
15	8	-5	-40	5	40
20	15	0	0	0	0
25	16	5	80	5	80
30	6	10	60	10	60
		50	50		230

$$a = x = \frac{fdx}{n} 20 + \frac{50}{50} = 21$$

$$\frac{\sum = \sum fdx + \left[\sum f(a-x) - \sum f(a-x) - \sum f(a-x)\right]}{n}$$

$$= \frac{230 + 28(21-20) - 22(21-20)}{50}$$

$$= \frac{230 + 28 - 22}{50} = \frac{230 + 6}{50} = 4.72$$

B. ***Under Direct Method***

Calculation of mean deviation from median

Wages (m)	(f)	(cf) (m-20)	dm from 2n (± 'ignored)	(fmd) (± ignored)
10	5	10	10	50
15	8	13	5	40
20	15	28	0	0
25	16	44	5	80
30	6	50	10	60
				230

M = value of (50+1)th item = value of 25.5th item = 20

$$m = \frac{fdm}{n} \frac{230}{50} = 4.6$$

Example – 5

Ascertain the mean deviation from mean, median, and mode form the following data :

A. ***Solution :***

Under Direct Method

Calculation of mean deviation from mean

Wages (m)	(f)	(mf)	D from 21	fd
10	5	50	11	55
15	8	120	6	48
20	15	300	1	15
25	16	400	4	64
30	6	180	6	54
Total	50	1050		236

$$a=\frac{\sum mf}{n}=\frac{1050}{50}=21$$

$$=\frac{\sum fd}{n}=\frac{236}{50}4.72$$

Under Short-cut Method

(m)	(f)	(dx)/m-20	(fdx)	dx(2) (± ignored)	fdx (± ignored)
10	5	-10	-50	10	50
15	8	-5	-40	5	40
20	15	0	0	0	0
25	16	5	80	5	80
30	6	10	60	10	60
	50		50		230

$$a=x=\frac{fdx}{n}20+\frac{50}{50}=21$$

$$\sum=\frac{\sum fdx+\left[\sum f(a-x)-\sum f(a-x)-\sum f(a-x)\right]}{n}$$

$$=\frac{230+\left[28(21-20)-22(21-20)\right]}{50}$$

$$=230+\left[28+22\right]=230+6=4.72$$

Short-cut Method :

Under this method the following steps are to be followed :

(1) Assume a value to be the value of the mean, median or the mode.

(2) From the assumed average, find out the deviations of the items ignoring the ± signs.

(3) Multiply the deviations with the corresponding frequencies and get the total of the products as the total of the deviations from the assumed average.

(4) Ascertain the actual mean, median or the mode as the case may be.

(5) Find the difference between the actual and the assumed mean, median ort he mode.

(6) Multiply such difference with the total number of items whose values are less than the values of the actual average.

(7) Multiply such difference with the total number of items whose values are less than the value of the actual average.

(8) Deduct the value of the step no.7 from the value of the step no.6 and add the result there of the total of the deviations from the assumed average as arrived under step no.3.

(9) Divide the result thus obtained by the number of items and get the value of the mean deviation.

Under this method the symbolic representation of the mean deviation would be :

1) $$\delta=\frac{\sum fdx+\left[\sum f(a-x)-\sum f(a-x)\right]}{n}$$

2) $$\delta m=\frac{\sum fdmx+\left[\sum f(m-x)-\sum f(m-x)\right]}{n}$$

3) $$\delta z=\frac{\sum fdzx+\left[\sum f(z-x)-\sum f(z-x)\right]}{n}$$

Where, fdx stands fro total of deviations from assumed mean, for sum of the frequencies of the values smaller than actual, F for sum of the frequencies of the values greater than the actual mean, 'a' for actual mean 'x' for assumed, median or mode as the case may be, 'n' for total of frequencies, fdmx for total of deviations from assumed median and fdzx for total of deviations from assumed mode.

The following examples would illustrate the calculations of medan deviation in discrete series.

Under Short-cut Method :

Calculation of mean deviation from Median :

Wages	(f)	dmx = 15 (± ignored)	fdmx
10	5	5	25
15	8	0	0
20	15	10	75
25	16	10	160
30	6	15	90
			350

$$\delta m = \frac{\sum f dmx + \left[\sum f(m-x)\right] - \sum f(m-x)}{n}$$

$$= \frac{350 + 13(20-15) - 37(20-15)}{50}$$

$$= \frac{350 \_ 65 - 185}{50} = \frac{350 - 120}{50} \frac{230}{50} = 4.6$$

C. Calculation of Mode by grouping Method :

(m)	(f)	Grouping in twos		Grouping in threes		
10	5	13		28		
15	8		23		39	
20	15	31				
25	16		22			37
30	6		22			

Analysis Table

Values Column	10	15	20	25	30
1				*	
2				*	
3			*		
4		*	*		
5		*	*	*	
6			*	*	*
Total	1	3	5	4	1

From the analysis table it appears that modal value in 20.

N.B.- Calculation of mean deviation from mode under both the methods will be made in the similar manner as it is shown above in case of mean deviation from median.

(iii) Calculation of Mean Deviation in Continuous Series :

In the continuous series the calculation procedure of mean deviation from mean, median or mode will remain the same except that under short cut method the following special adjustment will be made only when the actual average and the assumed average lie in two different class intervals :

(1) Multiply the frequency of the class in which the actual average lies, with the difference between the deviation of the mid value of the mean class from the total average one such deviation from the assumed average.

(2) Deduct the above result from the total of the deviations from the assumed average.

$$\delta = \frac{\sum fdx + [f(a-x)] - [f(d-dx)] - [f(d-dx)]}{n}$$

Here, the additional abbreviations – 'f' stands for the frequency of class in which the actual average lies, 'd' for the deviation of the mid value of the actual average class from the actual average and 'dx' for the deviation of the mid value of the actual average class from the assumed average.

The formulations for the mean deviation from median and mode will be adjusted accordingly.

The following illustrations would clarify the above procedure

Example – 6

Ascertain the mean deviation from mean and median from following observations.

M^1	0-10	·10-20	20-30	30-40	40-50
F^1	7,	12,	5,	18,	8

Solution

Under Direct Method :

Calculation of mean deviation from mean

(m)	(mc)	(f)	(mf)	d from 26.6± ignored	fd
10-10	5	7	35	21.6	151.2
10-20	15	12	180	11.6	139.2
20-30	25	5	125	1.6	8
30-40	35	18	630	8.4	151.2
40-50	45	8	360		147.2
		50	1330		596.8

$$a=\frac{mf}{n}=\frac{1330}{50}=26.6$$

$$\delta=\frac{\sum fd}{n}=\frac{596.8}{50}=11.94\text{ (approx)}$$

Under Short-cut Method :

Calculation of mean deviation from mean

(m)	(mv)	(f)	dx from 15 (±ignored)	fdx (± ignored)
0-10	5	7	10	70
10-20	15	12	0	0
20-30	25	5	10	50
30-40	35	18	20	360
40-50	45	8	30	240
				720

$$\delta=\frac{\sum fdx+\left[\sum f(a-x)-\sum f(a-x)-\left[f(d-dx)\right]\right]}{n}$$

$$=\frac{720+19(26.6-15)-26(26.6-15)(1.6-10)}{50}$$

$$=\frac{720+220.4-301.6-42}{50}=\frac{720+220.4-343.6}{50}$$

$$=\frac{720-123.2}{50}=\frac{596.8}{50}=11.94$$

Under Direct Method :

Calculation of mean deviation from median

(m)	(f)	(cf)	(mv)	dm from 30.56 ± ignored	fdm ± ignored
0-10	7	7	5	25.56	186.72
10-20	12	19	15	15.56	186.72
20-30	5	24	35	4.44	79.92
40-50	8	50	45	14.44	115.52
					588.88

explained properly as and when the calculation of standard deviation under different methods and different series will be dealt with.

2) COEFFICIENT OF STANDARD DEVIATION :

before proceeding to the actual work of computation it would be worthwhile to mention here that there are two types of results whcihmay be found out in connection with the standard deviation. They are : (1) Absolute standard deviation and (2) Co-efficient of standard deviation. The procedure discussed above relates only to the calculation of the absolute deviation, which would not be properly useful for the comparative studies. For making a comparative study of the variability between two or more series it is the coefficient of standard deviation which will be very much help full because this is expressed in relation to the arithmetic average in the form of a ratio symbolically, it is stated as follows :

$$\text{Coefficient S.D.} = \frac{S.D}{a}$$

Where standard deviation stands for standard deviation and 'a' for arithmetic average.

Thus after the absolute standard deviation is calculated the value of the coefficient of standard deviation can be ascertained without any difficulty.

3. Calculation of standard deviation in a series of individual observation. The calculation of standard deviation from a series of individual observation can be made by any of the following three method :

1. Direct method
2. Short-cut method
3. Alternative short-cut method.

A brief analysis of the above methods is made as follows :

Direct Method :

Under this method the following procedure are to be followed

First, Calculate the actual mean.

Next, find out the deviations of the items from the actual mean (with + & - signs)

Next, find out the step deviations by dividing each of the deviation with a common divisible factor if necessary and possible.

Next, square up the step deviations or the deviations as the case may be and get the total thereof.

(4) It facilitates the comparison between the formation of two or more series as the deviations are calculated form the central value.

DEMERITS :

(1) It ignores the algebraic signs and hence it is not capable of further algebraic treatment.

(2) Its value is not always definite as it can be calculated either from mean, median or from mode.

DEFINITION AND CONCEPT OF STANDARD DEVIATION :

Standard deviation is defined as "the square root of the arithmetic average of the square of the deviations measured from the mean".

From the above definition it follows that the standard deviation of a series is calculated with the following procedure :

1. Calculation of the arithmetic average or the mean, of the series.
2. Finding the deviations of the different items of the series from the mean
3. Squaring up of the deviations and getting their total.
4. Dividing the total of the squares of the deviations by the total number of the items and there by getting the average of the squares of the deviations or the second moment about he mean as it is other wise called.
5. Finding the square root of the above average of the squares of the deviations and the resulting figure is the standard deviation of the series.

It may be mentioned, here, that the measure of standard deviation has been introduced by Prof. Karl Pearson in the year 1893 as a development over the mean deviation which suffered from certain algebraic defects. Standard deviation is represented by the (sigma) a Greek letter.

Symbolically,

$$\delta = \sqrt{\frac{fd^2}{n}}$$

Where, δ stands for the standard deviation, fd² for sum of the squares of this deviations from the mean and n for total number of the items.

The above symbolic representation of the standard deviations formula will, of course, be little bit modified in different methods of calculations and in different types of series. This will be

M = value of $\frac{50^{th}}{2}$ item = value of 25th which lies in 30-40 class interval.

Thus, M = $l_1 = \frac{l_2 l_1}{f_1}(m-c) + 30 + 10\frac{(25-24)}{18}$

=30+ = 30+.56 = 30.56

m= $\frac{\sum fdm}{n} = \frac{588.88}{50} 11.78$ (approx)

Under Short-cut Method :

Calculation of mean deviation from median

(m)	(f)	(mv)	Dmx from 25 ± ignored		fdmx
0-10	7	5	20	140	
10-20	12	15	10	120	
20-30	5	25	0	0	
30-40	18	35	10	180	
40-50	8	45	20	160	600

$$\delta = \frac{\sum fdx + \left[\sum f(a-x) - \sum f(a-x) - [f(d-dx)]\right]}{n}$$

$$= \frac{600 + 24(30.56-25) - 26(30.56-25)}{50}$$

$$= \frac{600 + (24 \times 5.56)}{50} = \frac{(26 \times 5.56}{} = 600 - 11.12$$

$$= \frac{588.88}{50} = 11.78$$

N.B.: Here m the actual median m i.e. 30.56 and the assumed median, i.e. 25 remain almost in one class interval and the frequency against the value 30.56 is insignificant. Hence the last adjustment in the formula was not necessary here.

Merits and Demerits of Mean Deviation

MERITS

(1) It is relatively simple to understand and casy to calculate

(2) It is based on all the observations of the data

(3) It is less affected by the values of the extreme items

Lastly, substitute the values in the following formula and get the result of Standard deviation.

S.D = $\sqrt{\frac{ED^2}{n}}$ (if step deviation is not taken)

Or S.D = $\sqrt{\frac{ED^2}{n} / cxc}$ (if step deviations is taken)

The following example may be carefully studied to understand the procedure.

Example – 1

Calculate the Standard deviation from the following data relating to the marks.

10, 20, 30, 40, 50, 60, 70, 80

Solution

Calculation of Standard Deviation of Marks

Marks	Deviation from means .45	Step deviations by common factors	Square of step deviations
	(d)	(d/c)	$(d/c)^2$
10	-35	-7	49
20	-25	-5	25
30	-15	-3	9
40	-05	-1	1
50	05	1	1
60	15	3	9
70	25	5	25
80	35	7	49
360			168

a=$\frac{\sum / n}{n} = \frac{260}{8} = 45$

Where 'a' stands for the actual arithmetic average, 'm' for the sum of values of 'm' variables and 'n' for the total of the items.

Hence, the value of the arithmetic average is 45.

Now, W.D =

$$\frac{\sqrt{\sum (d/c)^2}}{n} \times C$$

$$\sqrt{\frac{168}{8} \times C}$$

$$\sqrt{21 \times 5}$$

$$= 4.6 \times 5 = 23(approx)$$

Where standard deviation stand for standard deviation, SE(d/c)² for sum of the squares of step deviations, 'c' for the common factor for step deviations and 'n' for the total number of the items

2. Short-cut Method.

Under this method the following procedures are to be followed :

First, assume a figure to be the value of the mean,

Next, find the deviations of the items from the assumed mean.

Next, find the step deviations with a common factor, if possible

Next square up the steps deviations and get the total.

Next, substitute the values in the following formula and get the standard deviation.

$\delta=\sqrt{\frac{dx^2}{n}-\frac{(dx)^2}{n}}$ xc where dx stands for the deviation from the assumed mean.

The following example would clarify the method described above.

Example – 2

Give the data below find the standard deviation by short cut method

4.5, 6.5, 75, 8.5, 9.5

Solution

Calculation of Standard Deviation

Variables (m)	Deviation from assumed mean 7.5	Squares of deviations (dx)
4.5	-3	9
6.5	-1	1
7.5	0	0
	0	0
8.5	1	1
9.5	2	4
	-1	15

Now standard deviation = $\sqrt{\frac{dxp^2}{n}}-\frac{(dx)^2}{n}=\sqrt{\frac{15}{5}}-\frac{(1)^2}{5}=\sqrt{3\text{-}1.44}$ (approx)

Where, standard deviation stand for the standard deviation, dx^2 for sum of the squares of deviations calculated from the assumed mean, dx for sum of the deviations from the assumed mean and for the total number of items.

N.B. : Here step deviation was not necessary as the value of some of the deviations was very small in size.

3. Alternative Short-cut Method

This method of finding the value of standard deviation will be advisable only where the figures of m variable are of smaller size. Under this method the value of mean is assumed to be the values of deviations from the assumed mean. As such the calculation under this method will proceed as follows :

1. Squaring up of the m variables and getting their total
2. Getting the total of the m variables.
3. Substituting the values in the following formula

$$S.D=\sqrt{\frac{\sum m^2}{n}-\frac{(m)^2}{n}}$$

From the following example the calculation of standard deviation under the method discussed above will be clear.

Example – 3

As certain the value of standard deviation from the following series assuming that the value of the mean is zero.

Variables : 5, 10, 15, 20, 25

Calculation of standard deviation

(M)	$(M)^2$
5	25
10	100
15	225
20	400
25	625
75	1375

$$\text{Now } S.D=\sqrt{\frac{\sum m^2}{n}-\frac{(m)^2}{n}}$$

Now, S.D. $\sqrt{\frac{\sum f.d^2}{n}}$

Where, S.D. stands for the standard deviation, fd^2 for sum of the product of the squares of deviations and their frequencies, and 'n' for total of the frequencies.

Thus, S.D. = $\sqrt{\frac{579}{100}}$ = √5.79 = 2.4 (approx)

Short-cut Method :

Example – 6

From the following data ascertain the value of standard of deviations by short cut method :

Values :	0,	5,	10,	15,	20
Frequencies :	10,	4,	6,	8,	22

Solution

Calculation of standard deviation by short – cut method

(m)	(f)	(dx) x = 10	(dx/c)	($_{-20}$fds/c)	($_4$dx/c)^2	(fdx/c^2
0	10	-10	-2	20	1	10
5	40	-5	-4	4	0	4
10	6	0	0	0	0	0
15	8	5	1	8	1	8
20	22	10	2	44	4	88
	50			28		140

By the formula we have :

$$S.D = \sqrt{\frac{\sum f.dx/c^2}{n} - \frac{(f.dx/c)2 \times c}{n}}$$

Where, standard deviation stands for the standard deviation, $f.dx/c^2$ for sum of the product of squares of step deviations and their frequencies, f.dx/c fro sum of the product of the step deviations and the frequencies c for the common factor of step deviation and n for the total of frequencies.

Thus, S.D = $\sqrt{\frac{140}{50} - \frac{(28)^2}{50}} \times 5$ = "2.8 - .3x5x5

= 1.6 x 5 = 8 approx.

Alternative Short-cut Method.

Calculation of Standard deviation in a Discrete Series

The above procedure of calculation of standard deviation in a series of Individual observation will also remain the same in case of a Discrete series only with the exception that the frequencies of the values found here, will be multiplied with the respective deviations and the total of frequencies will be taken to be the number of items. N other words in discrete series the square of the deviations from the arithmetic average are multiplied by the total frequencies and the square root of this is the standard deviation of the series. Symbolically.

$$\delta = \sqrt{\frac{\sum fd^2}{n}}$$

Keeping this much in view, the calculation of standard deviation in a discrete series will be made as follows :

Under Direct Method

Example – 5

From the following frequency distribution find out the standard deviaion under discrete method

Variables :	4,	6,	8,	10,	12
Frequencies :	21,	25,	30,	16,	8

Solution

Calculation of standard deviation

Variables	Frequencies	m.f	Deviation from mean	Squares of deviations	Frequency sqs. Of deviation
(m)	(f)	·	(d)		(fd²)
4	21	84	-3.3	10.89	228.69
6	25	150	-1.3	1.69	42.25
8	30	240	0.7	+.749	14.70
10	16	160	.7	7.29	116.64
12	8	96	4.7		179.72
	100	730			579.00

$$a = \frac{mf}{n}\frac{730}{100} = 7.3$$

Thus, arithmetic average is 7.3

Where, 'a' stands for the arithmetic average, 'mf' for sum of the product of the values of variables and their corresponding frequencies and 'n' for total of frequencies.

Where, standard deviation stands for the standard deviation, for sum of the squares of the m variables, m for sum of the m variables and n fro total number of items.

Thus, $S.D = \sqrt{\frac{1375}{5} - \frac{(75)^2}{5}} = \sqrt{275 - 225} = \sqrt{50} = 7$ (approx)

Example – 4

Taking the data given in the example no.2 and example no.3 above into consideration, state which of the series is more consistent.

Solution

This is question of comparative study where the coefficient of standard deviation will give better result. By the formula we have.

Co-efficient of S.D.= $\frac{S.D}{a}$

In the solution to the example no.2 and 3 the standard deviations of the two series were 1.44 and 7 respectively. The arithmetic average of the two series will be ascertained as under.

1st series

$$a = x + \sqrt{\frac{\sum dx}{n}} = 7.5 + \frac{1}{5} = 7.3$$

2nd series

$$a = x + \frac{\sum dx}{n} = 0 + \frac{75}{5} = 15$$

Where 'a' stands for the actual arithmetic average, 'x' for the assumed arithmetic average, dx for sum of the deviations from the assumed arithmetic average and 'n' for the total number of items.

Now, the coefficient of standard deviation of the two series will be found as under

1st series

Coefficient of S.D= $\frac{S.D}{a}$ =7= $\frac{1.44}{7.3}$ =.2 (approx)

2nd series

Coefficient of S.D.= $\frac{S.D}{a} = \frac{7}{15}$=.47 (approx)

From the values of the coefficient of standard deviations of both the series now, it is clear that the series no.1 is having less coefficient of standard deviation. Hence, this series is more consistent that the series no.2.

Example – 7

Taking the figures given in the example – 6 above find the standard deviation under the alternative short cut method.

Solution

Calculation of standard deviation

By the formula we have :

$$S.D. = \sqrt{\sum f.m^2} + (\sum f.m^2)$$

Where, standard deviation stands for the standard deviation, $f.x^2$ for sum of the product of squares of m and the frequencies, f.m for sum of the product of m and f. and n fro the total of the frequencies.

Thus, S.D. = " $\frac{11300}{50} - \frac{(640)^2}{50}$ ="226 – $(12.8)^2$ ="226-164 = "62 = 8 (approx)

Calculation of Standard Deviation in a Continuous Series :

The calculation of standard deviation in a continuous series can be made in any of the following three methods :

1. Direct method
2. Short-cut method
3. Shortest method

The procedures of calculation under direct method and short cu' method will remain same here, as it was in case of discrete series only with the exception that he mid values of the class intervals will have to be found out which will be considered as the values of m variables for the purpose of calculation. The following examples would clarity the methods of calculation in a continuous series.

Direct Method

Example – 8

From the following data find the standard deviation under direct method

Class intervals :	0-10	10-20	20-30	30-40	40-50
Frequencies :	16	4	5	10	25

Solution

Calculation of standard deviation under direct method

Class intervals Mid values	(m)	(f)	m.f	(d)	d/c	$(d/c)^2$	$(f.d^2/c)$
0-10	5	16	80	-24	-12	144	2304
10-20	15	4	60	-14	-7	49	196
20-30	25	5	125	-4	-2	4	20
30-40	35	10	350	6	9		90
10-50	45	25	1125	16	8	64	1600
		60	1740				4210

$$a = \frac{\sum m.f}{n} = \frac{1740}{60} = 29$$

Hence, the value of the mean is 29

Now, S.D. = $\sqrt{\frac{f.d/c^1}{n}}$

$=? \frac{4210}{60} \times 2 =?70.17 \times 2 = 8.37 \times 2 = 16.74$

Hence, the standard deviation is 16.74

Short cut Method

Example – 9

Calculate standard deviation from the following data under the short cut method :

Class intervals :	0-5	5-10	10-15	15-20	20-25
Frequencies :	3	7	5	4	9

Solution

Calculation of standard deviation

Class interval	(m)	(f)	(dx)	(dx/c)	$(f.dx/c)^2$	$(f.dx/c)^2$	
0-5	2.5	3	-10	-2	-6	4	12
5-10	7.5	7	-5	-1	-7	1	7
10-15	12.5	5	0	0	0	0	0
15-20	17.5	4	5	1	4	1	4
20-25	22.5	9	10	2	18	4	36
		28			9		59

By the formula we have :

$$\text{S.D} = \sqrt{\frac{\sum f.dx/c^2}{n} - \frac{(f.dx/c^2}{n}} \times c$$

Where, standard deviation stands for the standard deviation, f.dx/c² for sum of the product of squares of stop deviations from the assumed mean and the respective frequencies f, dx/c fro sum of the product of step deviation from the assumed mean and the frequencies, 'c' for the common factor of step deviation, and 'n' for the total of frequencies.

Thus, S.D. = $\sqrt{\frac{59}{28} - \frac{(9)^2}{28}} \times 5 = \sqrt{2.1 - .1} \times 5 = 1.4 \times 5 = 7$

Hence, the standard deviation of the above series is 7 approximately.

Shortest Method

This method is applicable only where all the class intervals given in the series are of equal class magnitude. However, under this method the following steps are to be followed :

First, Find the cumulative frequencies in a separate column and get he total thereof.

Next, Cumulative the cumulative frequencies again in a separate column and get the total thereof.

Next, Find the value of F_1 by dividing the sum of the cumulative frequencies by the sum of the frequencies with the application of the following formula :

$$\sqrt{\frac{\sum f.d^2}{n}}$$

Next, Find the value of F_2 by dividing the sum of the cumulations of the cumulative frequencies by the sum of the frequencies with the application of the following formula.

Next, Ascertain the class magnitude by looking at thé class intervals.

Lastly, Put the formula and substitute the values therein to get the desired standard deviations:

$S.D. = 1 \times 2F_2 - f_1 - f^2_1$

Where, i stands for class magnitude, F_2 for average of the cumulation of the cumulative frequencies, F_1 for average of the cumulative frequencies.

The above method would be clarified from the following example.

Example – 10

Using the data given in the example – 9 find the standard deviation under the shortest method.

Solution

Calculation of standard deviation

Class interval	Frequencies	Cumulative frequencies	Cumulation of the cumulative frequencies
	(f)	(cf)	(ccf)
0-5	3	3	3
5-10	7	10	13
10-15	5	15	28
15-20	4	19	47
20-25	9	28	75
Total	28	75	166

By the formula we have :

S.D = I x$\sqrt{F_2 - F_1}$=5 x $\sqrt{2 \times 6 - 2.7 - 7.29}$ = 5 x $\sqrt{2}$ = 5 x 1.47

Workings

$F_2 = \frac{ccf}{n} = \frac{166}{28} = 6$ (approx)

$F_1 = \frac{cf}{n} = \frac{75}{28} = 2.7$ (approx)

Hence, the value of the standard deviations is 7 which is just equal with the result that was arrived at under the short cut method.

7. *Merits and Demerits of Standard Deviation.*

Having thus studied the meaning, nature and the calculation procedure of the standard deviation now, it will be possible to lay down here the outstanding merits and demerits that go for and against such a measure of dispersion. In doing so, however, it will be required to keep in view the characteristics of an ideal measure of dispersion. Thus, in the light of these characteristics the merit and demerits of the standard deviation can be drawn as under.

MERITS :

(1) It is rigidly defined and its value is always definite

(2) Its calculation is based on the observation of all the items of the series. This means, no items of the series is ignored in its calculation as it happens in the case of certain other measures of dispersion viz. Quartile, deviation.

(3) It is less affected by the fluctuation of sampling and for this reason the standard is used as a key instrument in the matter of Sampling analysis.

(4) It is less affected by the fluctuation fo sampling and for this reason the standard is used as a key instrument in the matter of sampling analysis.

DEMERITS :

(1) It is difficult to understand on the part of a layman because it involves with the mathematical intricacies.

(2) It is difficult to calculate in comparison to the other measures of dispersion.

(3) It is greatly affected by the values of the extreme items because it involves with the procedure of squaring up of the deviations thereby making the big items still bigger and ht small items still smaller in proportion to each other.

SUMMARY

To sum up, the standard deviation is a mathematical measure of dispersion enunciated by Prof. Karl Pearson in the year 1893 as a development over the mean deviation. It gives an idea about he standard of deviations of the items of the series from the central tendency of the series. However, in the determination of its value the deviations of the various items of the series are found out only from the arithmetic average and in doing so the and signs of the deviations are duly taken into consideration. As defended, it is the square root of the average of the square of deviations obtained from the mean. It possesses many mathematical properties and satisfies most of the qualities of the ideal measure of dispersion. Despite a few drawbacks that go against, it is regarded as the best measure of dispersions just as the mean is regard as the best measure of central tendencies.

SKEWNESS, MOMENTS AND KURTOSIS

SKEWNESS

1. MEANING

Skewness means lack of symmetry in the distribution of a statistical series. It gives an indication of the extent to which the items of the series concentrate around or scatter away from the central value.

DEFENITION :

(1) "When a series is not symmetrical, it is said to be asymmetrical or skewed" – Croxton & Cowden

(2) "A distribution is said to be 'skewed' when the mean and the median fall at different points in the distribution, and balance (or center of gravity) is shifted to one side or the other – to left or right".

2. Concept of Skewness

The concept of skewness will be clear from the following there diagrams showing (a) Symmetrical distribution and he tow types of skewed distribution such as (b) positively skewed distribution, (c) Negatively skewed distribution.

(a) *Symmetrical Distribution :*

The above diagram on the right shows us that the symmetrical distribution if plotted on a graph paper gives us a perfectly bell shaped curve where the values of mean, median and mode coincide. Here the quartiles are equidistant and the spread of the frequencies is same on both sides of the central line of the course.

(b) *Skewed Distribution*

The distribution, which is not symmetrical is called a skewed distribution. There are two types of skewed distribution.

(i) ***Positively skewed Distribution :*** In this type of distribution we find most of the frequencies towards the right side of the curve. That is why the right tail of this curve is longer than left tail and mean > median > mode.

(ii) ***Negatively skewed Distribution :*** In this type of skewness we find most of the items towards the left side of the curve and the left tail of the curve is longer than right tail and mode > median > mean.

3. OBJECTIVES OF SKEWNESS

Skewness of a distribution helps us in finding out :

(i) The nature and degree of concentration in a series i.e. whether the concentration is with higher or the lower value,

(ii) The empirical relations of median, mode and mode in a moderately skewed distribution 3 median – 2 mean = mode holds good or not.

(iii) To know if the distribution is normal.

4. DISTINCTION BETWEEN DISPERSION AND SKEWNESS

DISPERSION

1. It deals with spread of individual values around and central value in a distribution
2. It needs in finding out the degree of variability in data.
3. It indicates how far mean is representative of the series.
4. It deals with the amount of variation.

Remark

1. It the modal class or modal value cannot be clearly found just by observation, grouping table and analysis table should be prepared to find the modal value.

2. If mode is ill defined or the value of mode cannot be found out. The Karl Pearson's coefficient of skewness is by $\frac{3(Mean - Median)}{\delta}$

(b) Bowley's Coefficient of Sewness :

Bowley's coefficient of skewness is also unknown as quartile coefficient skewness. It is especially useful :

(i) When the mode is ill defined and extreme observations are present in the data.

(ii) When the distribution has open-end classes and unequal class intervals.

(a) It's absolute measure is $Q_3 + Q_1 - 2$ median

(b) Its relative measure or Bowley's coefficient of skewness is given by the formula

$$\frac{Q_3+Q_1}{Q_3-Q_1} - 2md$$

The range of variation under this method is ± 1. It is therefore, easy to interpret but the main drawback of this measure is that it is based on only 50% of the data.

Example – 2

Find a suitable measure of skewness from the following distribution:

Annual Sales (Rs.'000)	0-20	20-50	50-100	100-250	250-500	500-1000
No. of firms	20	50	69	30	25	19

Solution

Bowley's measure of skewness based on quartiles will be more appropriate here, since class-intervals are of unequal width.

Annual sales (Rs.5000)	No.of firms	Cumulative frequencies
0-20	20	20
20-50	50	70
50-100	69	139
100-250	30	169
250-500	25	194
500-1000	19	213

Solution

Calculation of mean standard deviation and mode

No. of rejects per operator	No. of operators (f)	Mid-value (x)	d=x-38/5	fd	Fd$_2$
21-25	5	23	-3	-15	45
26-30	45	28	-2	-30	60
31-35	28	33	-1	-28	28
36-40	42	38	0	0	0
41-45	15	43	1	15	15
46-50	12	48	2	24	48
51-55	3	53	3	9	27
Total	120			-29	223

$$\text{Mean } (x) = A + \frac{\sum fd}{n} \times h$$

$$= 38 + \frac{(-25)}{120} \times 5 = 36.96$$

Mode : By observation, it can be found from the table given above that the modal class is 36-40 or 35.5 – 40.5. And the values of , f and f$_2$ are 28.42, and 15 respectively.

The modal value can be found from the formula $L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$

$$\text{Mode (mo)} = 35.5 + \frac{42-28}{(42-28)+(42-15)} \times 5$$

$$= 35.5 + \frac{14}{14+47} \times 5 = 37.21$$

Standard deviation can be turned from the formula

$$\sqrt{\frac{\sum fd^2}{n} - \frac{(\sum fd)^2}{n}} \times h$$

$$= \sqrt{\frac{223}{120} - \frac{(-25)^2}{120}} \times 5$$

$$= \sqrt{1.8583} - 0.0434 = 6.74$$

Hence, Karl Pearson's coefficient of skewness is given by :

$$Sh = \frac{x - mo}{\delta} = \frac{36.96 - 37.21}{6.74} = -0.037$$

Ans.

SKEWNESS

1. It deals with symmetry of distribution of values on both sides of the central value.
2. It helps in finding whether the concentration is in higher or lower values.
3. It indicates if the distribution is normal.
4. it deals with the direction of the distribution.
5.
MEASURES OF SKEWNESS

Skewness can be computed in two ways viz a9i) absolute way and (ii) relative way.

Absolute measure is found not by deducting the value of mode from mean. The difference whether positive or negative indicates that the distribution is ositively skewed or negatively shelved.

However, absolute measure is not helpful while comparing the skewness between two series with different units. Here relative measure is to be calculated which is also otherwise known as coefficient of skewness.

There are several measures of the coefficient of skewness. But the to:d most commonly used measures are Karl-Pearson's coefficient of skewness and Bo:viey's coefficient of skewness.

a) Karl Pearson's Skewness

(a) Its absolute measure is Arithmetic mean – mode

(b) The relative measure or

Karl Peason's coefficient of skewness $= \dfrac{Arithmeticmean - Mode}{S.D}$

Where the mode is ill defined, the formula is $\dfrac{(3A.M - Median)}{S.D}$

Ex: In analysis of production rejects resulted in the following figures.

Calculate Karl Pearson's coefficient of skewness

No. of rejects per operator	No. of operators	No. of rejects per operator	No. of operators
21-25	5	41-45	15
26-30	12	46-50	12
31-35	28	51-55	3
36-40	42		

Calculate Karl Pearson's co-efficient of Skewness

Quartile items are

Q_1= 213 or 53.25 the or it lies in the 20 – 50 class

Q_2 or Md = $\frac{213}{2}$ or 106.5th or it lies in the class 50-100

Q_3 = $\frac{3\times213}{4}$ or 159.75th or it lies in the class 100-250

Using the formula for quartiles and median, we get

$$Q_1 = \frac{20 + 53.25 - 20 \times 30}{50} = 39.95$$

$$\text{Md or } Q_2 = 50 + \frac{106.5 - 70}{69} \times 50 = 76.45$$

$$Q_3 = 100 + \frac{159.75 - 139}{30} \times 150 = 203.75$$

Bowle's co-efficient of skewness is given by :

$$Sk_b = \frac{Q_3 + Q_1 - 2md}{Q_3 - Q_1}$$

$$= \frac{203.75 + 39.95 - 2 \times 76.45}{203.75 - 39.95} = 0.554$$

(C) Kelly's Measure of Skewness

Kelly's measure is a compromise between Pearson's and Bowley's measure. Here the distribution between 90th and 10th percentile or 9th decile and 1st decile is considered.

Kelly's absolute measure of skewness is given by

$P_{90} + P_{10} - 2P_{50}$ or $D_9 + D_1 - 2D_5$

The relative measure or the coefficient of Kelly's

$$\text{Skewness is } \frac{P_{90} + P_{10} \; 2P_{50}}{P_{50} + P_{10}} \text{ or } \frac{D_9 + D_1 - 2D_5}{D_9 - D_1}$$

Example – 3

From the data given below, calculate a coefficient of skewness based on percentile :

Marks	No. of students	Marks	No. of studnets
0-10	4	30-40	10
10-20	6	40-50	7
20-30	20	50-60	3

Solution

Marks	Frequency	Cumulative frequency
0-10	4	4
10-20	6	10
20-30	20	30
30-40	10	40
40-50	7	47
50-60	3	50

Median is the n/2th or 50/2th or 25th item which is in 20-30 class. P_{10} is the value of $\frac{10(50)}{100}$ or 5th item, which is in 10-20 class. P_{90} is the value of $\frac{90(50)}{100}$100 or 45th items which is in 40-50 class.

$$\text{Median} = 20+\frac{10}{20}(25-10)=27.5$$

$$P_{10}=10+\frac{10}{6}\times(5-4)=11.67$$

$$P_{10}=40+\frac{10}{7}(45-40)=47.14$$

$$\text{Kelly's skewness is } \frac{P_{10}+P_{90}-2_m}{P_{90}-P_{10}}$$

$$=\frac{11.67+47.14-(2\times 27.5)}{47.14-11.67}=\frac{3.81}{34.47}=0.11$$

(D) ***Measure of Sk'wness Based on Moments :***

A measure of skewness can be obtained by suing the third moment about the mean. This method is explained in the moment section.

MOMENTS

6. Meaning

The word moment is a mechanical term, which refers to a measure of force. In statistics it means "the arithmetic average of certain power of the deviations of the items from their arithmetic mean "the 4th moment is denoted ur (mur). It means finding out the average of rth powers of deviations.

Thus, Ur= $\frac{\sum(x-x)^2}{n}$ (for ur-grouped data)

Ler = $\frac{\sum f(x-x)^{'}}{\sum f}$ (for grouped data)

Here, it may be observed that

Li$_1$= $\frac{\sum f(x-\bar{x})}{n} = \frac{Ex}{n} - \frac{E\bar{x}}{n} = \bar{x} = \bar{x} = 0$

However, the first moment about the origin is given by $\frac{\sum(x-0)}{n} = \bar{x}$

The second moment is $U_2 \frac{\sum(x-\bar{x})^2}{n}$

Thus the second moment about he mean is the variance.

Similarly, the third moment is $U_3 \frac{\sum(x-\bar{x})^3}{n}$

And fourth moment is $U_4 \frac{\sum(x-\bar{x})^4}{n}$

The third moment about the mean is considered as measure of skewness and the fourth moment is used as a measure of Kurtosis.

We can summarise the above discussion as follows :

As we have defined moment about mean we can also define moments about any point say A, the rth moment about any point A is denoted by lur¹.

Thus Ur = $\frac{\sum(X-A)^2}{n}$ (for grouped data)

And Ur= $\frac{\sum f(X-A)^2}{\sum f}$ (for ungrouped data)

Example – 4

Calculate the second, third and the fourth moments for the following distribution of service time at the registration counter of a local post office.

Service time (in minutes)	2.0	2.5	3.0	3.5	4.0	4.5
Nuclear of Frequencies :	5	30	40	15	5	5

UNIT – III

Lesson - I

The objective of this unit are:

3.1.0 Objective

3.1.1 Explain the meaning of Regression.

3.1.2 Explain the definition of Regression.

3.1.3 Explain utility or significance of Regression.

3.1.4 Difference between correlation and regression analysis.

3.1.5 Describe the types of Regression.

3.1.6 Explain the Regression lines.

3.1.7 When are these Lines coincide.

3.1.8 Why Two Regression Lines.

3.1.9 Explain the Regression equation and Regression Coefficient.

3.1.10 Explain the Properties of Regression Coefficient.

3.1.11 Solved Examples

3.1.12 Self Assessment.

STRUCTURE:

- Explain the meaning and definition of Regression.
- Difference between correlation and Regression analysis.
- Explain utility or significance of Regression.
- Describe the types of Regression.
- Explain the Regression lines.
- Explain the Regression equation.
- Describe Regression coefficient.
- Explain the Properties of Regression Coefficient..
- Solved Examples
- Self assessment.

According to Prof. Karl Pearson, Kurtosis of a series can be measured by B_2 wh equal to

$$\frac{\mu_4}{\text{[illegible]}}$$

Indications :

If $\beta_2 = 3$, it gives normal or measured curve

If $\beta_2 > 3$, it gives a lepto kurtic curve

If $\beta_2 < 3$, it gives a platykurtic curve

2.γ_2 (Gamma two) :

According to R.A. Fisher, Kurtosis is measured by γ_2 which is equal to $\beta_2 - 3$.

❖❖❖

Solution Calculation for moments are given below

x	f	fx	$x - x$	$f(x - x)^2$	$f(x - x)^3$	$f(x - \bar{x})^4$
2.0	5	10.0	-1.0	5.0	-5.0	5.0
2.5	30	75.0	-0.5	7.5	-3.75	1.87
3.0	40	120.0	0	0	0	0
3.5	5	52.5	0.5	3.75	1.87	0.94
4.0	5	20.0	1.0	5.0	5.0	5.0
4.5	5	22.5	1.5	11.25	16.87	25.31
	100	3000		32.5	14.99	38.12

$$U_2 = \frac{\sum f(x-x)^2}{\sum f} = \frac{32.5}{100} = 0.325$$

$$U_3 = \frac{\sum f(x-x)^3}{n} = \frac{14.99}{100} = 0.15$$

$$U_4 = \frac{\sum f(x-x)^4}{\sum f} = \frac{38.12}{100} 0.381$$

Since U_3 # 0, we can deduce that the distribution is skewed.

KURTOSIS

8. Meaning

Kurtosis in Greek language mean 'bulginess' it measures the flatness of the curve. There terms are used for indicating flatness. Measourtic stands for a normal curve, leptokurtic for a peaked curve and platykurtic fcr a curve less peaked than normal as shown in the figure below :

Fig : 4

9.Kurtosis Measurements :

There are two constant or coefficients through which Kurtosis of a series is determined. They are (1) β2 and (2)γ_2

1.β_2 (Beta two) :

REGRESSION ANALYSIS

1.1 MEANING OF REGRESSION:

The dictionary meaning of the word Regression is "stepping back" or going back to the mean value. The word has first used by sir Francis Galton in 1877 in his study based on heredity. He studied the relationship between the heights of father and their Sons and arrived at some very interesting conclusions which are as follows:

i) Tall Fathers have tall sons and short fathers have short sons.

ii) The mean heights of sons of Tall father is less than the mean height of their fathers

iii) The mean height of the sons of short fathers is more than the mean height of their fathers Thus sir Galton found that the deviation in the mean height of the race was less than the deviations in the mean height of the son from the mean height of the race was less than the deviations in the mean height of the father from the mean or below the mean, the sons tended to go back or regress towards the mean.

Thus, regression means going back or returning towards the mean. Galton depicted the average relationship between the aforesaid two variables graphically. The line showing the relationship is called the line of regression.

1.2 DEFINITIONS OF REGRESSION ANALYSIS:

I. According to Morris Hamburg, "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 variable and to the measurement of the errors involved in the estimation process".

II. According to Ya –Lun chou,"Regression analysis attempts to establish the nature of the relationship between variables that is ,to study the functional relationship between the variable and thereby provide a mechanism for prediction or forecasting."

Thus, regression analysis is a measure of the average relationship between two or more variables in terms of the original units of the data. Moreover, reg ession analysis is a statistical method with the help of which we are in a position to estimate or predict the unknown values of one variable from the known values of another variable.

The variable which is used to predict the variable of interest is called independent variable or the explanatory variable and the variable we are trying to predict is called the dependent variable or the explained variable. The independent variable is denoted by X and the dependent variable is denoted by Y. The analysis is called simple linear regression.

1.3 UTILITY OR SIGNIFICANCE OF REGRESSION ANALYSIS:

Regression analysis is very widely used in almost all scientific disciplines. In Economics it is the basic technique for determining the relationship between the economic variables. More uses of Regression analysis are:

i) In Business:

In the field of Business this statistical tool is very widely used. Businessmen are interested in predicting future production, consumption, investments, prices, profits, sales etc.

ii) Social and Economical study:

In social and Economic study and planning- projection to population ,Birth rate ,Death rate and other rates are required. These variables can e estimated by Regression analysis.

iii) Used in calculating the correlation coefficient

With the help of regression analysis we also can calculate the coefficient of correlation. The sqare of coefficient of correlation (r) is called the coefficient of Determination. It measures the degree of association of correlatio that exist between two variables. In general, greater the value of r^2, better is the fit and the more useful the regression equations as a predictive device.

iv) It helps in estimating the value of dependent variable from the value of independent variables.:

The regression line describes the average relationship existing between X and Y variables. In other words, it depicts the mean values of X for given values of Y. The regression equation provides estimates of the dependent variables when the values of independent variables are substituted into the equation.

v) Helps in measuring error in using regression line.:

Regression analysis is very useful in obtaining a measure of the error involved in using the regression line as a basis for estimation. For this purpose the standard error of estimates is calculated.

1.4 DIFFERENCE BETWEEN CORRELATION AND REGRESSION ANALYSIS.

Correlation and regression analysis helps in studying the relationship between two variables yet they differ in their approach and objectives.

While correlation analysis tests the closeness with which two or more phenomenon co-vary, regression measures the nature and extent of the relation ,thus enabling us to make prediction. The main differences are ;

1. Correlation is a linear relationship between two variables so that they move either in the same or in the reverse direction where as regression is a measure of the average relationship between two or more variables.

2. Correlation only shows the existence of some association in the movement of variables where as regression presumes one variable as a cause and other variable as its effect.

3. Correlation is a measurement showing association between two variables. The coefficient of correlation (r) falls in the range of -1 to +1. Where as regression is a measurement of relationship with given constant of the equation. We can define the value of the dependent variable by substituting independent variable i.e. Y=a+bX.

$$Y=a+bX+cX^2$$

4. Correlation is applied for testing and verification of relationship only and gives only a limited information where as regression is applied for estimation and prediction and furnish us more comprehensive information.

5. Correlation is not capable of further mathematical treatment where as in regression various order of differentiation helps us in finding out the rates of change in the dependent variable with a give change in one or more independent variables.

1.5 TYPES OF REGRESSION.

Generally regression is of three types:

1. Simple and Multiple Regression
2. Linear and Non-Linear Regression.
3. Total and Partial Regression.

SIMPLE AND MULTIPLE REGRESSION:

The average relationship between exactly two variables in known as simple regression. In this analysis one of the variable is dependent and the other is the independent one.

On the contrary, under Multiple regression there are more than one independent variable e.g. the output of wheat is the result of fertilizers, rainfall and climate etc.

LINEAR AND NON-LINEAR REGRESSION:

In Linear regression ,the dependent variable varies at a constant rate with a given range of the independent variable. The constant rate of change can be expressed in absolute terms or in terms of percentages. It always gives a straight line whenever put in a graph paper.

$\Sigma X, \Sigma Y, \Sigma XY, \Sigma X^2$ indicate the totals which are computed from the observed pairs of values of two variables X and Y to which the Least squares estimating line is to be fitted and N is the number of observed pairs of values.

Regression Equation of X on Y:

The regression equation of X on Y is expressed as follows:

$$X= a + bY$$

To determine the values of a and b , the following two normal equations should be solved;

$$\Sigma X= Na+ b\Sigma Y$$
$$\Sigma XY= a \Sigma Y + b \Sigma Y^2$$

Deviations taken from arithmetic means of X and Y:

This method of finding out regression equation is very difficult . The calculations can be simplified if instead of dealing with the actual values of X and Y we take the deviations of X and Y series from their respective means . In such a case the two regression equations can be expressed as follows:

(i) Regression Equation of X on Y ;

$X\text{-}X = r [\sigma_x / \sigma_y](Y - Y)$

$r \sigma_x / \sigma_y$ is known as the regression coefficient of X and Y.

X is the mean of X –series

Y- is the mean of Y series.

The Regression coefficient of x and Y is denoted by the symbol b_{xy} It measures the change in X corresponding to a unit of change in Y. When deviations are taken from the means of X and Y , the regression coefficient of X and Y is obtained as;

$$b_{xy} = r[\sigma_x / \sigma_y] = \Sigma xy/\Sigma y^2$$

Instead of finding out the value of correlation coefficient σ_x , σ_y , etc , we can find the value of regression coefficient by calculating Σxy and Σy^2 and dividing the former by the later.

(i) Regression equation of Y on X :

$Y\text{-}Y = r [\sigma_y / \sigma_x](X - X)$

$r [\sigma_y / \sigma_x]$ is the regression coefficient of Y on X .It is denoted by b_{yx}. It measures the change in Y corresponding to a unit change in X. When deviations are taken from actual means, the regression coefficient of Y on X can be as follows:

$$b_{yx} = r[\sigma_y / \sigma_x] = \Sigma xy/\Sigma X^2$$

1.8 WHY TWO REGRESSION LINES

One regression line can not minimize the sum of square of deviations for both the variables. i.e., X and Y unless the relationship between them indicates perfect positive or negative correlation.

In case of perfect correlation one regression line is sufficient though X and Y have the same type of deviations.

1.9 REGRESSION EQUATIONS AND REGRESSION COEFFICIENT :

Regression equations, also known as estimating equations , are algebraic expression of the regression lines. Since there are two regression lines , there are two regression equations— the regression equation of X on Y is used to describe the variations in the values of X for given changes in Y and the regression equation of Y on X is used to describe the variation in the values of Y for given values of X.

Regression equation of Y on X

The regression equation of Y on X is expressed as follows:

$$Y= a+ bX$$

It may be noted that in this equation 'Y' is a dependent variable i.e., its value depends on X . 'X' is independent variable ,i.e., we can take a given value of X and compute the value of Y.

'a' is Y-intercept because its value is the point at which the regression line crosses the Y-axis, that is the vertical axis. 'b' is the "slope" of the line which represents change in Y for a unit change in X variable. 'a' and 'b' are called the numerical constants.

If the values of the constants 'a' and 'b' are obtained , the line is completely determined. By the method of Least Square ,the line should be drawn through the plotted points in such a way that , the sum of the squares of the deviations of the actual Y values from the computed Y values is the least, or in other words , in order to obtain a line which fits the points best $\Sigma(Y-Yc)^2$, should be minimum. Such a line is known as a line of best fit.

With a little algebra and differential calculus it can be shown that the following two equations, if solved simultaneously, will yield values of the parameters 'a' and 'b' such that the Least squares requirement is fulfilled.

$$\Sigma Y= Na+ b\Sigma X$$

$$\Sigma XY= a\ \Sigma X + b\ \Sigma X^2$$

These equations are usually called the normal equations.

In Non-Linear regression , the dependent variable changes at varying rate with a given change in independent variable.

TOTAL AND PARTIAL REGRESSION:

In case of total regression , all important variables are considered. For Example—all business activities are affected by multiplicity of causes.

while incase of partial regression , three or more variables are considered. For example—keeping constant, the effect of the influencing variable , i.e, the output of gram (y) is influenced by three variables , fertilizers(x),rainfall(t) and the number of ploughing(z), the total relationship can be defined as—

Y=f(x,t,z) (total regression)

Partial regression,

Y=f(X but not t and z)

Y=f(t but not x and z)

Y=f(z but not x and t)

1.6 REGRESSION LINES:

When two variables have linear relationship, the regression lines can be used to find out the vales of dependent variables. These regression lines are based on two equations called regression equations which gives the best estimate of one variable when the other is exactly known or given.

If we study the case of two variables i.e., X and Y, we shall have two regression lines as the regression of X on Y and the regression Y on X. The regrèssion line f Y on x gives the most probable values of X for given values of Y.

1.7 WHEN ARE THESE LINES COINCIDE:

When there is either perfect positive or perfect negative correlation between two variables (-1<=r<=+1) the regression lines will coincide means to have only one line.

The farther the two regression lines ,the lesser is the degree of correlation. The nearer the two regression lines to each other the higher is the degree of correlation. If the variables are independent i.e, r = o, the lines of regression are at right angles i.e, parallel to OX and OY.

It should be noted that the regression lines cut each other at the point of average of X and Y. if from that point where both the regression lines cut each other a perpendicular is drawn on the X axis, the mean value of X can be obtained. Similarly if from that point a horizontal line is drawn on the Y axis, mean value of Y can be obtained.

It should be noted that, the under-root of the product of two regression coefficients gives us the value of correlation coefficient. Symbolically,

$$r = \sqrt{b_{xy} \times b_{yx}}$$

Example :

The given data

A	X	Y
Arithmetic Mean	36	85
Standard deviation	11	8

Correlation co-efficient between X and Y=0.66

(i) Find two regression equations.

(ii) Estimate the value of X and Y=75.

Solution :

(i) Regression equation X on Y :

$$(x-\bar{x}) = r\frac{\sigma_x}{\sigma_y}(Y-\bar{y})$$

$$\bar{X} = 36, r = 0.66, \sigma_x = 11, \sigma_y = 8, \bar{Y} = 85$$

$$x - 36 = 0.66 \frac{11}{8} (y-85)$$

$$\text{or } x - 36 = 0.66 \frac{11}{8} (y-85)$$

or, x =.9075y – 771375 +36

or, x = = -41.1375 + .9075y

Regression Equation Y on X :

$$(x-\bar{y} = r\frac{\sigma_x}{\sigma_y}(x-\bar{y})$$

$$y - 85 = 0.66 \frac{8}{11} (x-36)$$

or, 4 – 85 = .48 (x – 36)

or, y – 85 = .48x – 17.82

∴y = 67.82 + 0.48x

(ii) From the regression equation of X on Y, we can get the estimated value of X and Y=75.

X = (0.9075 x 75) = 41.1375

∴y 75 = 26.925.

mple – 2

The average weekly wages of working class in Mumbai and Kolkata are Rs.12 and Rs.18 respectively, their standard deviations equal to Rs.2 and Rs.3 respectively and the efficient of correlation between them is +0.67. Find out the most likely wage in Kolkata of it 04 Rs.20 in Mumbai

Solution :

Let the wages in Mumbai be x and wages in Kolkata be y. Thus,

$\bar{x} = 12, \bar{y} = 18, \sigma_x = 2, \sigma_y = 3, \pi = o.67$

Regression equation of y on x is given by,

$$(y-\bar{y}) = \pi\frac{\sigma_x}{\sigma_y}(x-\bar{x})$$

$y - 18 = 0.67 \times \frac{3}{2}$ (x-12)

y – 18 = 1.005 (x-12)

=y=1.005x + 5.94

Thus, most likely wage in Kolkata of it is rs.20/- in Mumbai -

Y = (1.005 x 20) + 5.94

Y = (1.005 x 20) + 5.94

= 20.1 x 5.94

= 26.04.

Regression co-efficient of y on X i.e. by x

$$byx = \pi\frac{\sigma_y}{\sigma_x}$$

$$=0.6 \times \frac{1.8}{1.5} = \frac{1.08}{1.5} \cdot :0.72$$

(ii) Regression equation x on y :

$(x-\bar{x} = bxy(y-\bar{y})$

(x-43) = 0.5 (y – 38)

or, x – 43 = 0.54 – 19

or, x =0.5y – 19 + 43

∴x = 0.54 + 24

Regression equation of y on x.

$(y-\bar{y} = bxy(x-\bar{x})$

(x-38) = 0.72 (x-43)

or, y-38=0.72x – 30.96

or, y = 0.72x – 30.96 + 38

Y = 0.72x + 7.04.

(iii) Estimated value of x when y = 42

X = (.5x x 42) + 24

= 21 + 24

=45.

(iv) Estimated value of y when x=38

X=(.72 x 38) + 7.04

=27.36 + 7.04

=34.40.

Examples – 3

The following data is given;

$\sum x = 301, \bar{x} = 43, \sum y = 226$

Variance of x = $\sigma_x = 2.25$

Variance of y = $\sigma_y = 3.24$

Coefficient of correlation between x and y is 0.6.

Find out –

i) Both coefficient of regression

ii) Both regression equation

iii) The estimated value of x when y = 42

iv) The estimated value of y when x = 38

Solution

$$\bar{x} = \frac{\sum x}{N} \Rightarrow N = \frac{\sum x}{X} = \frac{301}{43} = 7$$

Thus, $\bar{y} = \frac{\sum y}{N} = \frac{266}{7} = 38$

Since the standard deviation is square root of variance, therefore,

$\sigma_x = \sqrt{2.25} = 1.5$

gression coefficient of x on y.

$$b_{xy}=\frac{\sum dxdy-\frac{\left(\sum dx\right)\left(\sum dy\right)}{n}}{\sum dy^2-\frac{\left(\sum dy\right)^2}{n}}$$

$$=\frac{57-0}{60-0}=0.95$$

gression line of y on x is

$y-\bar{y}=b_{yx}(x-\bar{x})$

or, y-12=0.95 (x-5)

or, y=0.95x + 7.25

Regression line of x on y is,

$x-\bar{x}=b_{xy}(y-\bar{y})$

or, x-5=0.95 9y-12)

or, x=0.95y – 6.4

Putting x=6.2 in the regression equation of y on x, we get,

Y=0.95 x 6.2 + 7.25

=5.84 + 7.25 =13.14

Thus, the estimate of y corresponding to x=6.2 is 13.14.

Example – 6

Find out regression equations from the following data when the deviations are taken from assumed mean.

x	27	27	27	28	28	18	29	29	30	31
y	18	18	19	20	21	21	22	23	24	25

Solution:

Computation of regression equation

x	dx=x-A A=27	dx^2	Y	dy=y-A A-21	dy^2	dxdy
27	0	0	18	-3	9	0
27	0	0	18	-3	9	0
27	0	0	19	-2	4	0
28	1	1	20	-1	1	-1
28	1	1	21	0	0	0
18	-9	81	21	0	0	0
29	2	4	22	1	1	2
29	2	4	23	2	4	4
30	3	9	24	3	9	9
31	4	16	25	4	16	16
N=10	Σdx=4	Σd^{2}x=116		Σdy=1	Σdx^2=53	Σdxdy=30

∴ The equation to the line of regression of y on x is

$y-\bar{y}=b_{yx}(x-\bar{x})$
Or, y – 7.08 = 0.10 (x-30)
Y = 0.10x + 4.08
When x = 40,
Y = 0.10 (40 + 4.08) = 8.08

Thus most likely yield of rice per acre when ruin is 40 inches is 8.08 tons.

Example – 5

Calculate the coefficient of correlation and obtain the lines of regression for the following data :

x:	1	2	3	4	5	6	7	8	9
y:	9	8	10	12	11	13	14	16	15

Obtain an estimate of y which should correspond on the average to x =

Solution :

Calculation of π

x	(x-5)dx	dx^2	Y	(y-12) dy	dy^2	dxd
1	-4	16	9	-3	9	12
2	-3	9	8	-4	16	12
3	-2	4	10	-2	4	4
4	-1	1	12	0	0	0
5	0	0	11	1	1	0
6	1	1	13	1	1	1
7	2	4	14	2	4	4
8	3	9	16	4	16	12
9	4	16	15	3	9	12
N=9	Σdx=0	Σdx^2=60	N=9	Σdy=0	Σdx^2=60	Σd

$$\therefore \bar{x} = ax + \frac{\sum dx}{n} = 5+0=5$$

$$\bar{y} = ax + \frac{\sum dy}{n} = 12+0=12$$

$$\pi = \frac{57-0}{\sqrt{60-0}\sqrt{60-0}} = \frac{57}{60} = \frac{19}{20} = 0.95$$

$$byx = \frac{\sum dxdy - \frac{\left(\sum dx\right)\left(\sum dy\right)}{n}}{\sum dx^2 - \frac{\left(\sum dx\right)^2}{n}}$$

$$= \frac{57-0}{60-0} = 0.95$$

$\sigma_y = \sqrt{3.24} = 1.8$

Thus, we are given,

$\bar{x} = 43, \bar{y} = 38, \sigma_x = 1.5, \sigma_y 1.8\pi = 0.6$

egression coefficient of x on y i.e. bxy,

$$bxy = \pi \frac{\sigma_x}{\sigma_y}$$

$$= 0.6 \times \frac{1.5}{1.8} = \frac{0.9}{1.8} = 0.5$$

ample – 4

Obtain the equation to the line of regression of yield of rice on water from the data given in e following table. Estimate the most probable yield of rice for 40 inches of water.

ater (Inches)	12	18	24	30	36	42	48
ield (tons)	5.27	5.68	6.25	7.21	8.02	8.71	8.42

Solution :

Water is the independent variable and therefore we denote it by x the yield being denoted by y.

x	y	(x-30) dx	(y-7.08) dy	dxdy	Dx^2
12	5.27	-18	-1.81	32.58	324
18	5.68	-12	-1.40	16.80	144
24	6.25	-6	-0.83	4.98	36
30	7.21	0	0.13	0	0
36	8.02	6	0.94	5.64	36
42	8.71	12	1.63	19.56	144
48	8.42	18	1.34	24.12	324
210	49.56	0	0	103.68	1008

$n = 7, \bar{x} = \frac{210}{7} = 30, \bar{y} = \frac{49.56}{7} = 7.08$

Coefficient of regression of y on x,

$$byx = \frac{\sum dxdy - \frac{(\sum dx)(\sum dy)}{n}}{\sum dx^2 - \frac{(\sum dx)^2}{n}}$$

$$= \frac{103.68 - 0}{1008 - 0} = 0.10 \text{ approx.}$$

$$\bar{x} = A + \frac{\sum d_x}{N} = 27 + \frac{4}{10} = 27.4$$

$$\bar{4} = A + \frac{\sum dy}{N} = 21 + \frac{1}{10} = 21 + 0.1 = 21.1$$

i) Regression coefficient of x on y;

$$bxy = \frac{\sum d_x d_y \times N - \left(\sum d_x . \sum d_y\right)}{\sum d^2 y . N - \left(\sum d_y\right)^2}$$

$$= \frac{(30 \times 10) - (4 \times 1)}{53 \times 10 - (1)^2} = \frac{300 - 4}{530 - 1}$$

$$= \frac{296}{529} = 0.56$$

ii) Regression co-efficient of y on x;

$$bxy = \frac{\sum d_x d_y \times N - \left(\sum d_x . \sum d_y\right)}{\sum d^2{}_x . N - \left(\sum d_x\right)^2}$$

$$= \frac{(30 \times 10) - (4 \times 1)}{116 \times 10 - (1)^2} = \frac{300 - 4}{1160 - 16}$$

$$= \frac{296}{1144} = 0.26$$

iii) Regression equations of x on y;

$x - \bar{x} = bxy\left(y - \bar{y}\right)$
(x-27.4) = 0.56 (y-21.1)
(x-27.4) = 0.564 – 11.82
or, x = 564 – 11.82 + 27.4
x = 0.56y + 15.58.

iv) Regression equations of y on x ;

$y - \bar{y} = byx\left(x - \bar{x}\right)$
(y-21.1) = 0.26 (x-27.4)
(y-21.1) = 0.26 – 7.12
or, y = 0.26x – 7.12 + 21.1
y = 0.26x + 13.98.

ample - 7 :

Obtain the two regression equations by the method of last square from the following data

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

Solution :

x	Y	X^2	Y^2	xy
6	9	36	81	54
2	11	4	121	22
10	5	100	25	50
4	8	16	64	32
8	7	64	49	56
$\Sigma x=30$	$\Sigma y=40$	$\Sigma x^2=220$	$\Sigma y^2=340$	$\Sigma xy=214$

i) Regression line of x on y;
To get the values of 'a' and 'b' the following formula are used;
$\Sigma x=Na+b\Sigma y$
$\Sigma xy = a\Sigma y + b\Sigma y^2$

Hence,
$30 = 5a + 40b$ ....(i)
$214=40a + 340b$.........(ii)

Multiplying equation 9i) by 8 and subtracting it from equation (ii)
$240 = 40a + 320b$
$214 = 40a + 340b$
$26 \;\; = \; -20b$
$\therefore b = -1.3$
By putting the value of b in equation (i), we have,
$30 = 5a + (40x - 1.3)$
$5a = 30 + 52$
$5a = 82$
$a = 16.4$

Now regression equation of x on y;
$x = a + by$
$x = 15.4 - 1.3y$

ii) Regression line of y on x;
To get the values of 'a' and 'b' the following formula is used –
$\Sigma y = Na + b\Sigma x$
$\Sigma xy = a\Sigma x + b\Sigma x^2$

Hence,
$40 = 5a + 30b$ ....(i)
$214 = 30a + 220b$ ....(ii)

Multiplying equation (i) by b and subtracting it from (ii).

$240 = 30a + 180b$
$214 = 30a + 220b$
$26 \quad = \ - 40b$
$\therefore b = - 0.65$

By putting the value of b in equation (i)
$40 = 5a + (30x - 0.65)$
$5a = 40 + 19.5$
$5a = 59.5$
$a = 11.9$

Now, regression equation of y on x ;
$Y = a + bx$
$Y = 11.9 - 0.65\,x$

❖❖❖

both X and Y are decreasing:

	80	74	60	57
	22	20	18	7

when both the variables move in opposite directions i.e. if as one variable is increasing, other on an average is decreasing and if as one variable is decreasing ,the other on an average is increasing ,then they are said to be inversely or negatively correlated.

X is increasing and Y is decreasing:

	80	84	90	99
	22	20	18	7

X is decreasing and Y is increasing:

	80	74	60	57
	22	3	45	67

Thus , generally Price and supply of goods are Positively correlated where as Price and demand of goods are negatively correlated.

(B) On the basis of ratio of change of variables —

i) Linear correlation

ii) Non-Linear correlation.

i) If the variation in the value of two variables are in a constant ratio,it is known as Linear correlation. In other words if the amount of change in one variable tends to bear a constant ratio to the said to be Linearly correlated.

Constant ratio in both X and Y:

X	10	12	14	16
Y	50	60	70	80

ii) If the variation in the value of the variable are not in a constant ratio then they are said to be non-linear correlation. In other words if the amount of change in one variable does not bear a constant ratio to the amount of change in other variable, then they are said to be non-linear or curvi-linear. for example ,if we doubled the amount of rainfall the production of rice or wheat etc would not necessarily be doubled.

X	10	12	16	18
Y	50	55	70	75

(C) On the basis of the no of variables——

i) simple correlation

ii) Multiple Correlation

iii) Partial Correlation

are said to be correlated. The degree of relationship between the variables is measured through correlation analysis. The measure of correlation called the correlation coefficient. It is used in measuring the closeness of the relationship between the variables.

1.2 DEFINITION:

correlation analysis deals with the association between two or more variables -Simpson and Kafka

Correlation is an analysis f the covariation between two or more variables.—Ya-Lun Chou

Correlation analysis attempts to determine the "degree of relationship" between variables.-A.M Tuttle

1.3 Significance of the study of Correlation:

Correlation analysis is very widely used due to the following reasons:

1) correlation analysis is used to define in one figure the relationship exist among several variables. For example-income and expenditure ,demand and price, price and supply

2) We can estimate the value of one variable given the value of another .

3) Correlation analysis contributes to the understanding of economic behaviour in business t helps the executive to estimate costs, sales, prices and other variables on the basis of some other series with which these costs, sales or prices may be functionally related.

4) Correlation analysis used to reduce the rage of uncertainty.

1.4 Types of Correlation:

(A) On the basis of direction of change in variables—

i) Positive Correlation

ii) Negative Correlation

i) When the variables tend to move in the same direction i.e. if as one variable is increasing the other on an average ,is also increasing and if as one variable is decreasing ,the other on an average is also decreasing, then the variables are said to be positively correlated.

If both X and Y are increasing:

X	80	84	90	97
Y	22	24	28	34

UNIT – III

Lesson - II

objective of this unit are:

2.0 Objectives

2.1 Explain the meaning of Correlation.

2.2 Explain the definition of correlation.

2.3 Explain the utility or significance of Correlation.

2.4 Explain the types of Correlation.

2.5 Explain the methods of studying Correlation.

2.6 Explain the properties of Correlation.

2.7 Solved Examples.

2.8 Self Assessment.

STRUCTURE:

- Explain the meaning Correlation.
- Explain the Definition of Correlation.
- Explain the Significance of Regression.
- Difference between correlation and regression analysis.
- Describe the typ·s of Correlation.
- Describe the methods of studying Correlation.
- Explain the properties of Correlation.
- Solved Examples.
- Self Assessment.

Correlation analysis:

1.1 MEANING OF CORRELATION:

In Practice we may come across a large number of problems involving the use of two or more variables. if the change in one variable indicates the change in other variable, then they

i) Simple correlation:

if only 2 variables are there it is a problem of simple correlation when three or more variables are studied, it is a problem of either multiple or partial correlation. for example, if we study together the relationship between production, rainfall , use of fertilizer, it is known as multiple correlation. so far as partial correlation is concerned we recognize more then two variables, but consider only two variables, to be influencing each other, the effect of all other influencing variables being kept constant.

1.5 METHODS OF STUDYING CORRELATION:

The method of studying correlation between two variables are ;

i) Scatter Diagram Method

ii) Graphic Method

iii) Karl Pearson's Coefficient of correlation

iv) Concurrent Deviation Method

v) Rank Correlation Method

vi) Method of Least Square

Among all the methods of studying correlation the first two methods are based on the knowledge of graphs ,where as rest of the methods are based on mathematical analysis.

I. SCATTER DIAGRAM METHOD:

This is the simplest device to ascertain correlation between two variables. In this method the values of the variables are plotted on the graph paper in the form of dot i.e., for each pair of X and Y values we put a dot and thus obtain as many points as the number of observations, the greater the scatter of plotted points on the chart, he lesser is the relationship between the two variables. The more closely the points come to a straight line, the higher the degree of relationship.

Different cases of Scatter Diagram Method:

i) If all the points lie on a straight line falling from the lower left hand corner to the upper right hand corner, correlation is said to be perfectly positive i.e.,r=+1.

ii) If all the points are lying on a straight line rising from the upper left-hand corner to the lower right hand corner of the diagram, correlation is said to be perfectly negative i.e.r=-1.

iii) if the plotted points lie on the straight line paralled to the X –axis or in a haphazard manner, it shows absence of any relation between the variables i.e.,r=0

If the plotted points fall in a narrow band there would be a high degree of correlation.

If the points are widely scattered over the diagram it shows very little relationships between the variables.

MERITS OF SCATTER DIAGRAM METHOD:

It is simple and non-mathematical method of studying correlation between the variables

It is very easy to draw scatter diagram

It is easy to understand and interprete.

It is not affected by the extreme items.

DEMERITS OF SCATTER DIAGRAM METHOD:

Though it is a non-mathematical method, we cannot know the exact degree of correlation between the variables.

I. GRAPHIC METHOD:

In this method the individual values of the two variables are plotted on the graph paper. We thus obtain two curves, one for X variable and another for Y variable. Depending on the direction and closeness of the two curves we can infer whether or not the variables are related. If both the curves drawn are moving in the same direction i.e., either upward or downward, correlation is said to be positive. on the other hand if the curves are moving in the opposite direction correlation is said to be negative.

Example:

II. KARL PEARSON'S COEFFICIENT OF CORRELATION:

This is one of the .est method of studying correlation. It gives a numerical measure that defines the direction and degree of relationship between two independent variables. It is otherwise known as Pearsonian coefficient of correlation.It is denoted by the symbol r. The formula for computing Pearsonian r is –

$r=\Sigma xy/N\ 6_x\ 6_y$

where, $x=(X-\bar{X})$; $y=(Y-\bar{Y})$

6_x=standard deviation of series X

6_y=standard deviation of series Y

N=number of pair of observations

This method is applied only where the deviations are taken from actual mean and not from assumed mean

The value of coefficient of correlation obtained by Karl Pearson Method always lies between -1 to +1 when r=+1,it means there exists perfect positive correlation between the variables when r=-1,there exists perfect negative correlations between the variables. when r=0,it means there is no relationship between the variables.

The above formula can also be transformed to the following form;

$r=\Sigma xy/ \sqrt{6x2x6y2}$

where , $x=(X-\bar{x})$ and $Y=(Y-\bar{Y})$

It is obvious that ,while applying this formula we have not to calculate separately the standard deviation of X and Y series.

DIRECT METHOD OF FINDING OUT CORRELATION:

Correlation coefficient can also be calculated without taking deviations of items either from actual mean or assumed mean, i.e., actual X and Y values. The formula in such case is,

$$r=\frac{N\Sigma XY-(\Sigma X)\ (\Sigma Y)}{\sqrt{N\Sigma X^2-(\Sigma X)^2}\ \sqrt{N\Sigma Y^2-(\Sigma Y)^2}}$$

When the deviations are taken from Assumed mean:

MERITS OF PEARSONIAN COEFFICIENT :

i) Karl Pearson's the most popular method of studying Correlation.

ii) It summarizes in one figure the degree of correlation and also the direction.

LIMITATIONS OF PEARSONIAN COEFFICIENT:

i) The correlation coefficient always assumes the linear relationship regardless of the fact whether that assumption is correct or not.

ii) This method is very time consuming .

iii) Grate care should be taken while interpreting the value of this coefficient.

iv) The value of this coefficient of correlation is affected by extreme values.

CONCURRENT DEVIATION METHOD:

This method of correlation coefficient is very simple method of studying the relationship between two variables. This method is very useful in Time series analysis. In this method the deviation are taken from the value in the preceding item than that of actual or assumed mean. In

But $\sum\left(\frac{x}{\sigma_x}+\frac{y}{\sigma_y}\right)$ is the sum of squares of real quantities and as such it can not be negative,

at the most it can be zero.

$\therefore 2N(1+\pi > 0)$

Hence π cannot be loss then -1; at the most it can be - 1.

Similarly, by expanding $\left(\frac{x}{\sigma_x}+\frac{y}{\sigma_y}\right)^2$ it can be shown that this is equal to 2N(π-1)

This again can not be negative; at most can be zero.

∴π cannot be greater than +1; at most can be +1.

Hence, $-1 \leq \pi \leq +1$

2. The co-efficient of correlation is independent of change of scale and origin the variable x and y.

Proof : By change of origin we mean subtracting some constant from every given value of x and y and by change of scale we mean dividing or multiplying every value of x and y by some constant.

We know that,

$$\pi y = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-)^2}}$$

Where $\bar{x}$ and $\bar{y}$ refers to actual means of x and y series.

Let us now change the scale and origin, deducting a fixed quantity a from x and b from y and dividing it by x and y series by a fixed value I and c. Then new values of x and y obtained from original x and y shall be.

$$x = \frac{x-a}{i}\ y = \frac{y-b}{c}$$

$$\text{Mean of } x = \frac{\frac{\sum(x-a)}{c}}{N}$$

$$= \frac{\frac{\sum x - Na}{i}}{N} = \frac{\sum x - Na}{Ni}$$

iii) This is the only method that can be used where we are given only the ranks and not the actual data.

iv) This method also can be used if only the actual data is given.

DEMERITS:

i) This method can not be used to find the correlation in grouped frequency distribution.

ii) Where the number of items exceeds 30,the calculation for this method becomes very tedious and require lots of time.

1.6 PROPERTIES OF CORRELATION COEFFICIENT:

The important properties of correlation are ;

1. The coefficient of correlation lies between -1 and +1. Symbolically, $-1 \le \pi \le +1$ or $|\pi| \le |$.

Proof : Let x and y be the deviations of x and y series from their mean and $\sigma_x \sigma_y$ be their standard deviation then;

$$\sum\left(\frac{x}{\sigma_x}+\frac{y}{\sigma_y}\right)=\sum\left[\frac{x^2}{\sigma_{x^2}}+\frac{y^2}{\sigma_{y^2}}+\frac{2xy}{\sigma_x\sigma_y}\right]$$

$$=\frac{\sum x^2}{\sigma_{x^2}}+\frac{\sum y^2}{\sigma_{y^2}}+\frac{2\sum xy}{\sigma_x\sigma_y}$$

But $\frac{\sum x^2}{\sigma_{x^2}}=N\left[\therefore \sigma_{x^2}=\frac{\sum x^2}{N} or \frac{\sum x^2}{\sigma_{x^2}}=\frac{\sum x^2}{\sigma_{x^2}}\times N=N\right.$

Similarly,

$$\frac{\sum y^2}{\sigma_{y^2}}=N$$

Also, $\frac{2\sum xy}{\sigma_x\sigma_y}=2N\pi \left[\therefore \pi=\frac{\sum xy}{N\sigma_x\sigma_y}\right]$

Hence,

$= N + N + 2N\pi$

$= 2N + 2N\pi$

$=2N(1+\pi)$

method the direction of change of X and Y variable is required to find out. The formula can defined as.

MERITS:

It is simplest method of all the methods.

If the number of items are very large, this method provides quick estimation of relationship between the variables.

DEMERITS:

The result obtained from this method is the rough estimation of the relationship.

It does not differentiate between the big and small item.

SPEARAN'S RANK CORRELATION COEFFICIENT:

This method is otherwise known as Rank coefficient of correlation, which is derived by Edward Spearman in 1904. This method is used to find out covariability or lack of covariability between two variables. In this method ranks are given to the observation . It really does not matter which way the items are ranked, number one may be the largest one or may be the smallest one.

This method is very useful when quantitative measures for certain factors (e.i,evaluation of ability of leadership or judgement of a beauty contest)can not be fixed. But the individual in the group can be arranged in order to there by obtaining for each individual a number indicating his or her rank in the said group. the formula can be defined as:

$R=1-(6\Sigma D^2)/N(N-1)$

Where,

R= Rank correlation coefficient

D= Difference of Rank between paired items in two series.

N= No of items.

MERITS:

i) This method is simple to understand and easy to apply.

ii) This method can be used with great advantage if the data is of qualitative nature like intelligence, beauty, honesty etc.

But $\frac{\sum x - Na}{Ni} = \frac{\bar{x}-a}{i}$

Thus, mean of $x = \frac{\bar{x}-a}{i}$

Similarly, the mean of y $= \frac{\bar{y}-b}{c}$

The value of the coefficient of correlation π for new set of values will be,

$$\pi xy = \frac{\sum\left(\frac{x-a}{i}-\frac{\bar{x}-a}{i}\right)\left(\frac{y-b}{c}-\frac{\bar{y}-b}{c}\right)}{\sqrt{\sum\left(\frac{x-a}{i}-\frac{\bar{x}-a}{i}\right)\left(\frac{y-b}{c}-\frac{\bar{y}+b}{c}\right)^2}}$$

$$= \frac{\sum\left(\frac{x-a-\bar{x}+a}{i}\right)\left(\frac{y-b-\bar{y}+b}{c}\right)}{\sqrt{\sum\left(\frac{x-a-\bar{x}+a}{i}\right)^2\sum\left(\frac{y-b-\bar{y}+b}{c}\right)^2}}$$

$$= \frac{\frac{\sum(x-\bar{x})(y-\bar{y})}{ic}}{\sqrt{\frac{\sum(x-\bar{x})^2}{i^2}\times\sum\frac{(y-\bar{y})}{c^2}}}$$

$$= \frac{\sum(x-\bar{x})(y-\bar{y})/ic}{\sqrt{\sum(x-\bar{x})^2(y-\bar{y})^2/ic^2}}$$

$$= \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})\sum(y-\bar{y})^2}}$$

Thus the coefficient of correlation is independent of change of scale and origin.

3. The coefficient of correlation is the geometric mean of two regression coefficients.

Symbolically,

$$\pi = \sqrt{bxy \times byx}$$

4. The degree of relationship between the two variables is symmetric as;

$$\pi xy = \pi yx$$

$$\pi xy' = \frac{\sum xy}{N\sigma_x\sigma_y} = \frac{\sum yx}{N\sigma_y\sigma_x}$$

$=\pi yx$

When deviations are taken from an assumed mean

When actual means are in fractions, the calculation of correlation by the usual method would involve too many calculations and may be very time consuming. In such a case we make use of the assumed mean method for finding out correlation. When deviations are taken from an assumed mean the following formula is applicable :

$$r = \frac{N\sum d_x d_y - \sum d_x \times \sum d_y}{\sqrt{N\sum d_x^2 - (\sum d_x)^2}\sqrt{N\sum d_y^2 - (\sum d_y)^2}}$$

Where, dx refers to deviations of x series from an assumed mean i., $(x - \bar{x})$ and dy refers to deviations of y series from an assumed mean i.e. $(y - y)$.

CONCURRENT DEVIATION METHOD :

$$r_c = \pm\sqrt{\pm\left(\frac{2c - n}{n}\right)}$$

Where, π stands fro coefficient of correlation by concurrent method; C st[illegible]ds fro the number of concurrent deviations or the number of positive signs obtained after m[illegible]ltiplying D_x with D_y. n=Number of pairs of observations compares.

Example – 1

Calculate the coefficient of correlation between the heights of father and son from the following table :

Heights of father (in inches) :
65 6667 68 69 70 71,

Heights of son (in inches)
67 6866 64 72 . 72 69

Solution :

Calculation of correlation coefficient

S.No.	Height of Fathers	Height of Sons	Deviation from mean 68	Deviation from mean 69	Product of deviation	Sq.of deviation	Sq. of deviation
	x	y	x^1	y^1	$x^1 x y^1$	x^{12}	x^{12}
1	65	67	-3	-2	6	9	4
2	66	68	-2	-1	2	4	1
3	67	66	-1	-3	3	1	9
4	68	69	0	0	0	0	0
5	69	72	1	3	3	1	9
6	70	72	2	3	6	4	9
7	71	69	3	0	0	9	0
n=7	476	483	0	0	20	28	32

$\bar{x}$ = Average height of father

$$= \frac{\sum x}{n} = \frac{476}{7} = 68$$

$\bar{y}$ = Average height of son

$$= \frac{\sum x}{n} = \frac{483}{7} = 69$$

From the table, $\sum x'y' = 20, \sum x'^2 = 28,$ and $\sum y'^2 = 32$

Substituting the above values in Karl Pearson's formula

$$\pi = \frac{\sum x'y'}{\sqrt{\left(\sum x'^2\right)\left(\sum y'^2\right)}}$$

$$\pi = \frac{20}{\sqrt{28 \times 32}} = 0.67 \text{ approx}$$

Thus, the correlation between the heights of father and son is positive and 0.67.

Example – 2

Psychological tests of intelligence and of arithmetical ability were applied to 10 children. Her is a record of ungrouped data showing intelligence ratio (I.R) and arithmetic ratio (A.R) calculate

Child	A	B	C	D	E	F	G	H	I	J
I.R	105	104	102	101	100	99	98	96	93	92
A.R	101	103	100	98	95	96	104	92	97	94

Solution :

Calculation of correlation coefficient

Child	I.R.	A.R					
	X	Y	$x' = (x - \bar{x})$	$y' = (y - \bar{y})$	$x'y'$	x'^2	y'^2
A	105	101	6	3	18	36	9
B	104	103	5	5	25	25	25
C	102	100	3	2	6	9	4
D	101	98	2	0	0	4	0
E	100	95	1	-3	-3	1	9
F	99	96	0	-2	0	0	4
G	98	104	-1	6	-6	1	36
H	96	92	-3	-6	18	9	36
I	93	97	-6	-1	6	36	1
J	92	94	-7	-4	28	49	16
	990	980	0	0	92	170	140

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = \frac{123}{\sqrt{138 \times 164}} = \frac{123}{150.439} = 0.818$$

Example – 6

Seven students have obtained the following ranks in two subjects, History and Geography. Find their rank correlation coefficient

Rank in History	7	1	4	6	5	3	2
Rank in Geography	5	1	2	3.5	3.5	7	6

Solution

Calculation of rank correlation coefficient

	R_2	$(R_1 - R_2)$ D	D^2
R_1	5	2	4
7	1	0	0
1	2	2	4
4	3.5	2.5	6.25
6	3.5	1.5	2.25
5	7	-4	16
3	6	-4	16
2			
			ΣD^2=48.5

$$R = 1 - \frac{6\left[\sum D^2 + \frac{1}{12}\left(M^3 - M\right)\right]}{N^3 - N}$$

$$= 1 - \frac{6\left[48.5 + \frac{1}{12}\left(2^3 - 2\right)\right]}{7^3 - 7} = 1 - \frac{294}{336} = 0.125$$

Example – 7

Calculation Karl Pearson's coefficient of correlation for the data given below, taking 66 and 63 as assumed means of x and y respectively

Height of Husband X	60	62	64	66	68	70	72
Height of Wives Y	61	63	63	63	64	65	62

Solution :

Calculation of coefficient of correlation

		$(x_1-\overline{x_1})$		x2	$(x_2-\overline{x_2})$		
Case	x_1	x_2	x_2^2		x_2	x_2^2	x_1x_2
A	10	0	0	9	1	1	0
B	6	-4	16	4	-4	16	16
C	9	-1	1	6	-2	4	2
D	10	0	0	9	1	1	0
E	12	2	4	11	3	9	6
F	13	3	9	13	5.	25	15
G	11	1	1	8	0	0	0
H	9	-1	1	4	-4	16	4
n=8	$\Sigma x_1=80$	$\Sigma x_1=0$	$\Sigma x_1^2=32$	$\Sigma x_2=64$	$\Sigma x_2=0$	$\Sigma x_2^2=72$	$\Sigma x_1x_2=43$

$$\overline{x_1}=\frac{\sum x_1}{N}=\frac{80}{8}=10, \overline{x_2}=\frac{\sum x_2}{N}=\frac{64}{8}=8$$

$$\pi_{12}=\frac{\sum x_1x_2}{\sqrt{\sum x_1^2\times\sum x_2^2}}$$

$\Sigma x_1x_2=43$, $\Sigma x_1^2=32$, $\Sigma x_2^2=72$

Substituting the values,

$$\pi_{12}=\frac{43}{\sqrt{32\times 72}}=\frac{43}{\sqrt{2304}}=\frac{43}{48}=+0.896$$

Example – 5

Calculate the Karl Pearson's correlation coefficients between x and y.

X:	23	27	28	28	29	30	31	33	35	36
Y :	18	20	22	27	21	29	27	29	28	29

Solution :

Calculation of Karl Pearson's correlation coefficient

X	(x-30) X	X²	Y	(y-25) Y	Y²	XY
23	-7	49	18	-7	49	49
27	-3	9	20	-5	25	15
28	-2	4	22	-3	9	6
28	-2	4	27	2	4	-4
29	-1	1	21	-4	16	4
30	0	0	29	4	16	0
31	1	1	27	2	4	2
33	3	9	29	4	16	12
35	5	25	28	3	9	15
36	6	36	29	4	16	24
$\Sigma x=300$	$\Sigma x=0$	$\Sigma x^2=138$	$\Sigma y=250$	$\Sigma y=0$	$\Sigma y^2=164$	$\Sigma xy=123$

Arithmetic mean of x-series $= \frac{\sum x}{n} = \frac{990}{10} = 99$

Arithmetic mean of y-series

$= \frac{\sum y}{n} = \frac{980}{10} = 98$

From the table, $\sum x'y' = 92, \sum x'^2 = 170$ and $\sum y'^2 = 140$

$$r = \frac{\sum x'y'}{\sqrt{(\sum x'^2)(\sum y'^2)}}$$

$$= \frac{92}{\sqrt{170 \times 140}} = \frac{9.2}{\sqrt{238}} = 0.62$$

Example – 3

Two series x and y with 50 items each have standard deviations 4.5 and 3.5 respectively if the summation of products or deviations of x and y series from their respective arithmetic means be 420 find the coefficient of correlation between x and y.

Solution :

We are given the following results :

$n = 50, \sigma_x = 4.5, \sigma_y = 3.5, \sum x'y' = 420$

$$r = \frac{\sum x'y'}{n\sigma_x\sigma_y} = \frac{420}{50 \times 4.5 \times 3.5}$$

$$= \frac{420 \times 2 \times 2}{50 \times 9 \times 7} = \frac{8}{15} = 0.53$$

4. Making use of the data summarized below, calculate the coefficient of correlation r_{12}.

Case	X_1	Y_2	Case	X_1	Y_2
A	10	9	E	12	11
B	6	4	F	13	13
C	9	6	G	11	8
D	10	9	H	9	4

Solution

Calculation of coefficient of correlation.

X	(x-66) dx	dx^2	y	(y-63) dy	dy^2	dxdy
60	-6	36	61	-2	4	12
62	-4	16	63	0	0	0
64	-2	4	63	0	0	0
66	0	0	63	0	0	0
68	2	4	64	1	1	2
70	4	16	65	2	4	8
72	6	36	67	4	16	24
	Σdx=0	Σdx^2=112		Σdy=5	Σdy^2=25	Σdxdy=46

$$\pi = \frac{N\sum d_x d_y - \sum d_x . \sum y}{\sqrt{N\sum dx^2 - \left(\sum dx\right)^2}\sqrt{N\sum dy^2 - \left(\sum dy\right)^2}}$$

$$= \frac{7\times 46 - 0\times 5}{\sqrt{7\times 112 - (0)^2}\sqrt{7\times 25 - (5)^2}} = \frac{322}{\sqrt{784\times 150}} = \frac{322}{342.93} = 0.939$$

Example – 8

Calculate Karl Pearson's coefficient of correlation for the following paired data :

X:	28	41	40	38	35	33	40	32	36	33
Y :	23	34	33	34	30	26	28	31	36	38

Solution :

Calculation of Karl Pearson's correlation coefficient

X	(x-35 dx	dx^2	y	(y-30) dy	dy^2	dxdy
28	-7	49	23	-7	49	49
41	6	36	24	4	16	24
40	5	25	33	3	9	15
38	3	9	34	4	16	12
35	0	0	30	0	0	0
33	-2	4	26	-4	16	8
40	5	25	28	-2	4	-10
32	-3	9	31	1	1	-3
36	1	1	36	6	36	6
33	-2	4	38	8	64	16
Σx=356	Σdx=6	Σdx^2=162	Σy=313	Σdy=13	Σdy^2=211	Σdxdy=85

$$r=\frac{\sum xy}{\sqrt{\sum x^2\sum y^2}}=\frac{123}{\sqrt{138\times164}}=\frac{123}{150.439}=0.818$$

Example – 6

Seven students have obtained the following ranks in two subjects, History and Geography. Find their rank correlation coefficient

Rank in History	7	1	4	6	5	3	2
Rank in Geography	5	1	2	3.5	3.5	7	6

Solution

Calculation of rank correlation coefficient

R_1	R_2	$(R_1- R_2)$ D	D^2
7	5	2	4
1	1	0	0
4	2	2	4
6	3.5	2.5	6.25
5	3.5	1.5	2.25
3	7	-4	16
2	6	-4	16
			$\Sigma D^2=48.5$

$$R=1-\frac{6\left[\sum D^2+\frac{1}{12}\left(M^3-M\right)\right]}{N^3-N}$$

$$=1-\frac{6\left[48.5+\frac{1}{12}\left(2^3-2\right)\right]}{7^3-7}=1-\frac{294}{336}=0.125$$

Example – 7

Calculation Karl Pearson's coefficient of correlation for the data given below, taking 66 and 63 as assumed means of x and y respectively

Height of Husband X	60	62	64	66	68	70	72
Height of Wives Y	61	63	63	63	64	65	62

Solution :

Calculation of coefficient of correlation

Example – 10

On the basis of concurrent deviation method find out coefficient of correlation from the following data :

Year	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969
Supply Index	114,	127,	128,	120,	120,	123,	127,	127,	133,	137
Price Index	108,	104,	104,	106,	100,	98,	99,	99,	97,	92,

Solution

Computation of co-efficient of correlation by concurrent deviation method.

Year	Supply Index (x)	Dx	Price Index (1)	Dy	C=DxDy
1960	114		108		
1961	127	+	104	-	-
1962	128	+	104	=	-
1963	120	-	106	+	-
1964	120	=	100	-	-
1965	123	+	98	-	-
1966	127	+	99	+	+
1967	127	=	99	=	+
1968	133	+	97	-	-
1969	137	-	92	-	-
N=9					C=2

$$\pi_c = \pm\sqrt{\pm(2C-N)} = \pm\sqrt{\pm\left(\frac{2\times2-9}{9}\right)}$$

$$= \pm\sqrt{\pm\left(\frac{4-9}{9}\right)} = -\sqrt{-\left(\frac{-5}{9}\right)}$$

$$= -\sqrt{0.5556}$$

$$\therefore \pi_c = -0.745$$

Example – 10

The ranking of 10 students in two subjects, i.e., Accountancy and Auditing, are given below.

Accountancy	3	5	8	4	7	10	2	1	6	9
Auditing	6	4	9	8	1	2	3	10	5	7

Compute the coefficient of Rank correlation

UNIT - IV

MEANING AND DEFINITION OF INDEX NUMBER

MEANING :

Certain questions in connection with the relative changes of costs production price etc. arise in our mind. The questions are like :

How much have general business conditions changed since a year ago and since five years ago on the average?

How much the general price level have risen compared with 2001-2002 ?

How has industrial production changed during the past one decade? Questions like these are common among those interested in business affairs and a valuable aid in answering them is the special statistical device known as the index number. The dictionary meaning of index number is 'a figure indicating the relative changes in cost, production, etc. of a given period of time, as compared with those of a specific period in the past represented by a number 100 and used as an arbitrary base'.

Index number is a specialized type of average. It is used extensively in all fields where statistical measures can be applied. Index number measures the central tendency of a time series. If the units in which two or more series are expressed are different, averages cannot be used to compare them. Where it is difficult to measure directly the variation in the effects of a group of factors, relative variations are measured. It suggests the need for some device by which the movements of complex groups can be portrayed accurately and it illustrates the difficulty of obtaining an accuratc picture of complex groups without the use of some summerising device such as index number. Without such device, no definite knowledge of the changed in a complex group of variables can be obtained. To obtain a definite picture, the individual price changed, wage changed, etc, must be summerised in a composite expression. This is both the nature and the value of an index number, as applied to problems of measuring business activity. It is an indicator of composite variation in a large number of elements. Index number is otherwise known as the economic baromerter which measures the relative change in the economic variables over a specified period of time.

DEFINITION

Index number has been defined by people differently. According to L.R. Connor, "An index number is a device for estimating the relative movements of a statistical variate in cases where measurements of its actual movements are inconvenients or impossible'.

Year	Production (x)	$(x-\bar{x})$ x	x^2	Unemployed	$(y-\bar{y})$ y	y^2	xy
1978	100	-4	16	15	0	0	0
1979	102	-2	4	12	-3	9	6
1980	104	0	0	13	-2	4	0
1981	107	3	9	11	-4	16	-12
1982	105	1	1	12	-3	9	-3
1983	112	8	64	12	-3	9	-24
1984	103	-1	1	19	4	16	-4
1985	99	-5	25	26	11	121	-55
	$\Sigma x=812$	$\Sigma x=0$	$\Sigma x^2=120$	$\Sigma y=120$	$\Sigma y=0$	$\Sigma y^2=184$	$\Sigma xy=-92$

$$\pi=\frac{\sum xy}{\sqrt{\sum x^2-\sum y^2}}$$

$$x=(x-\bar{x})y=(y-\bar{y})$$

$$\bar{x}=\frac{\sum x}{N}=\frac{832}{8}=104$$

$$\bar{y}=\frac{\sum y}{N}=\frac{120}{8}=15$$

$\Sigma xy=-92$, $\Sigma x^2=120$, $\Sigma y^2=184$

$$\pi=\frac{-92}{\sqrt{120\times184}}=-0.619$$

❖❖❖

Solution

Let us assume rank for Accounting be R_1 and rank for Auditing be R_2.

Computation of co-efficient of rank correlation

R_1	R_2	$D=(R_1-R_2)$	D^2
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	1
N=10			$\Sigma D^2=214$

$$R = 1 - \frac{6\sum D^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 214}{10^3 - 10}$$

$$= 1 - \frac{1284}{990}$$

$$= 1 - 1.297$$

$$\therefore R = -0.297$$

Example – 11

The following table gives indices of industrial production of registered unemployed (in hundred thousand) calculate the value of the coefficient so obtained.

Year	1978	1979	1980	1981	1982	1983	1984	1985
Index of production	100	102	104	107	105	112	103	99
Number unemployed	15	12	13	11	12	12	19	26

Solution :

Calculation of Karl Pearson's correlation coefficient

1. DEFINITION OF THE PURPOSE OF THE INDEX NUMBER

2. SELECTION OF THE BASE PERIOD

Index numbers measure the relative change in the level of a phenomenon as compared to the level of the same phenomenon on a previous date. The pervious date or the period on which the current variations are based is known as the base period of the index number. The base period should be a period of stable (from economic standpoint) average prices. The choice must be a compromise, since prices of all commodities will be seldom, if ever at an average level at the same time. If a period of high prices is chosen, the subsequent index numbers will be artificially low and vice-versa. So this period should be a period free from all abnormalities like war, famine or earthquake and should not be a period either of boom or of depression. The best way out of the difficulty is to take average rice over a number of years. Thus the base year should be normal or representative in same way. It should not be too distant in the past, so that a comparison of the price level as related to the base period will be of definite present or comparative value. There are two methods by which base period can be selected. They are :

a) *Fixed Base Method :*

In this method the base period is fixed. A particular year is generally chosed arbitrarily and the prices of subsequent years are expressed as relatives of the prices of the base year Sometimes instead of choosing a single year as the base, a period of a few years is chosen and the average prices of this period is taken as the base year's price causes surprising differences in the result. This is due to the fact that the chose of base is really a type of weighting. If we consider the individual commodities of an index number, if will be seen that the individual a prices will bear certain relations to each other in any given period, but these relations are not likely to be the same for any two period. Thus a change in base is in fact also a change in weights.

b) *Chain Base Method*

In this method there is no fixed base period. The year immediately preceding the one for which price relatives have to be calculated is assumed as the base year. Thus for the year 1956 the base year would be 1955 and for 1955 it would be 1989 and so on. The chief advantage of this method is that the price relatives of a year can be compareu with the price level of immediately preceding year. Now commodities can be included and old commodities can be omitted without destroying the comparative base. But in the method comparison cannot be made over a long period.

For calculating price relatives, the prices of the base year or base period is assumed at 100 and the prices of other years are expressed as percentages of the price of the base year. Price relative in fixed base can be calculated as follows :

the three percentage relatives of the year 1750 arithmetically, and the resulting relative was his index of the change in prices between the years 1500 and 1750.

Since the time of introduction of Carli's simple index number, the use of index numbers has increased considerably. The use and value of both general and special index numbers have increased very rapidly with the growing interest in statistics.

USES AND LIMITATIONS OF INDEX NUMBER

As a statistical device, the main purpose of present day index numbers in the business world is to indicate average changes, either generally as to wide range of business activities, or specifically as to one or few phases of business activities. The price level is one of the most important subjects of index number study and we have today many published index numbers, which indicate the changes in price. Index numbers tell us about the changes in the general price level in a country, and throw light on the value of the money. A study of purchasing power of money is very important from various points of view. Index numbers also help in a study of comparative purchasing power of money in different countries of the world and of stability in their price levels.

Other index numbers are published regularly to indicate changes in the value of trade and stock in hand prices, in production of commodities, in the cost of living and in many other business data. Cost of living index numbers tell us about he changes in the cost of living of different groups of people in society. With such index numbers it is possible to study the chances in the level of real wages of the laboures. Index of industrial production help in measuring the industrial progress made by a particular country. Similarly, index numbers of business activity threw light on the economic progress that various countries have made. Thus the index numbers of various types are useful in framing suitable economic and business policies, measuring trends and forecasting future business activities.

But it should always be kept in mind that index numbers have their own limitations. They are only approximate indicators of the relative level of a phenomenon. There can be errors not only in the collection of the data but also in the selection of the base selection of representative commodities and selection of appropriate weights. Each step in the construction of index numbers is full of possibilities of errors of all types but despite these dangers, it can safely be said that if an index number is not deliberately distort it will show correctly at least the trend of the phenomenon which it is measuring. But indices constructed for one purpose should not be used for other purposes were they may not be fully appropriate and may lead to fallacious conclusion. So it follows that there should be careful application of this useful tool.

STEPS IN THE CONSTRUCTION OF AN INDEX NUMBER

Although the conditions under which index numbers are constructed are not uniform, in general the steps are as follows.

In the words of P. Maslov 'in statistics, index numbers are numerical values characterizing the changes of complex economic phenomena over a period of time or space" Dr. J.C. Chaturvedi defines index number as "a statistical device used to express the averages in the magnitude of a group of related variable. It is a representative number".

In the words of H.Arkin and R.R. Coltoan. "The index number is a statistical device for measuring changes in groups of data.

Regarding the techniques of index number A.I. Bowley says that "Index numbers are meant to measure the change in some quantity which we cannot observe directly, which we know to have definite influence on many other quantities which we can so observe. Tending to increase all, or diminish all, which this influence, is concealed by the action of many causes affecting the separate quantities in separate ways.

As regards the purpose of index number Fisher says "purpose of the index number is that it shall be fairly represented, so far as one single figure can, the general trend of many diverging rations from which it is calculated. It should be a just compromise, among conflicting elements, the fair average of the golden mean".

Hence, an index number may be defined as a number, which is used to measure the level of a certain phenomena at any given data in comparison with the level of the same phenomena at some standard data.

From the various definition of index number, the following conclusions can be drawn about its nature:

i) Index number are numerically expressed.

ii) Index numbers are always comparative in nature. The main function of index numbers is to render comparison of economic phenomena possible. Hence, it is also called economic baromerter.

iii) Index numbers are closely related to averages.

HISTORY

About the middle of eighteenth century, an Italian name, Carli wished to know the effect of the discovery of America-upon the level of price in Italy. To measure the price change, which has taken place, he obtained a representative price of each three commodities in the year 1500 and in the year 1750. The three commodities chosen were grain, oil and wine, which were considered to be representative of all Italian commodities. These prices were then expressed as percentage relatives, that is prices in 1500 were each taken as100 and the prices in 1750 were expressed relatively to their corresponding prices for the year 1500. Carli then averaged

$$\text{Price relative for current year} = \frac{Currentyearsprices}{Baseyearsprice} \times 100$$

Price relative in chain base can be determined as follows :

$$\text{Price relative of current year} = \frac{Currentyearsprices}{Previousyearsprice} \times 100$$

3. SELECTION OF COMMODITIES

The next questions that arise is, what items should be selected for inclusion in the index? The answer to this will constitute the raw material of the index. Clearly, all items pertaining to an equity cannot be included. It is impossible. Their number has to be restricted. Care must be taken to see that items selected are representative of tastes, habits and customs of the people. Cosmetics and other luxury articles cannot fix a place in a working class schedule. Tangible commodities which can be easily recognized and which are not likely to vary in quality should find a place in an index. These restrictions narrow down the choice of items to a considerable extent.

Adequate number of articles should be selected might behave like the parent population. What is adequate depends upon individual circumstances and resources. (The accuracy of a sample, we know increases with the square not of its size). Consequently the larger the number that is included in the index the more accurate will be the result.

Sources of Data

After the commodities have been selected the next problem that arises is the collection of data regarding the commodities. The data include the price of the commodities and the quantities sold. The prices of a commodity vary from place and even at one place from shop to shop. Just as it is not possible to include all commodities in an index number. Similarly it is not possible to collect information's from all sources. Price quotations can be obtained from regularly published journals or from merchants, producers and other dealers who are in possession of the basic raw material. The former source should not be too much depended upon unless its reliability and accuracy have been established. Great caution is to be exercised in seeing that the quotations received are representative persons have to be done. Generally such places are chosen where a particular commodity is purchase or sold in large quantities. After the selection of the places, the next thing is to appoint some representatives who would supply the price quotations from time to time.

4. METHOD OF COLLECTION

The different source and places from where quotations are received must be comparable. Comparability is the essence of statistical data. Terms, units, qualities must be standardized. All persons must give quotations for the same quantity in the same units and with the same terms Prices may be quoted in two ways :

i) Money prices.

ii) Quantity prices.

Money prices are quoted per unit of commodity, e.g. what at Rs. 100 per quintal. Quantity prices, on the other hand, are the reverse of the former. They are quoted per unit of money e.g what one kilo per rupee. So far as possible money prices should be given in the quotations. They are easy to interpret and understand. Quantity prices cause unnecessary confusion. When money prices increase, quantity prices fall and vice versa. Thus the have to be differently interpreted.

When the actual collection of data begins two problems arise :

a) The frequency of obtaining quotations.

L) Which prices to collect opening, closing midway etc. So far as the first problem is concerned it will depend upon the nature of the index. For a monthly index two quotations per week may suffice. A decision on the second point should be taken with determination of the units.

Average Quotations

The daily, weekly or monthly quotations for each commodity which are received will have to be averaged. The average might be obtained by adding up the price of each commodity from different center and dividing the aggregate by the number centres.

The most popular method of constructing an index number of group of prices is to average the relatives for the individual prices. The aggregative method adopted by American statisticians, though simple and intelligible, is not very common. We have then to decide the problem of the most appropriate average to be used. He various statistical averages that we have seen in the chapter of "Measures of Central Tendency" have their own particular merits and uses. Theoretically any average may be used but from a practical point of view the selection has to be made from :

a) Arithmetic Average

b) Median, and

c) Geometric Mean.

It is most widely used in computation of index numbers. It is simple and well understood. But the demerits of this average are :

i) It is too much affected by extreme items.

ii) It gives too much weight to increasing items and too little to decreasing ones.

iii) It gives greater weight is given to rapidly rising items in comparison to those, which are rising slowly.

iv) Similarly more weight is given to rapidly falling items than to slowly falling ones

v) It is no reversible.

MEDIAN

Median, though superior to the arithmetic average in the sense that it is less affected by the presence of extreme values, in seldom used in index number construction. It is erratic when the number of items is small. At times it has to be interpolated, It does not remove the bias present in relatives due to rising or falling prices and, moreover, it is not reversible. However, it is easy to calculate and may be suitable when some of the data are of questionable accuracy and representativeness.

Geometric Mean

Forceful arguments in favour of the use of G.M may be advanced. Index numbers, we know, are concerned with ratios or relative changes and the superiority of this average lies in the fact hat it gives equal weight to equal ratios of change. Doubling of one price is compensated by having of another and G.M of such relatives will show no change while the A.M will show an increase. Moreover index numbers calculated by using this average are reversible and therefore, base shifting is easily possible. Another advantage claimed by this average is that it reduces the effect or upward movements and increases that of downward movements. This is in keeping with the economic principle of buying less in case of a rise in price and more when the price falls.

The geometric mean, on the other hand, is less understood, less popular and not easy to compute. It does not yield an objective quantity when, weights are used but in case of the arithmetic average weighting by quantity, gives the total expenditure.

Mathematically the geometric mean gives more accurate results than the arithmetic mean. The degree of inaccuracy in the latter however, is not very marked. Connor has evaluated the usefulness of various averages and he says, "On balance the advantages of the geometric mean preponderate". British statisticians are now more in favour of the use of this average.

6. METHOD OF WEIGHTING

Items very in importance. So far, we have not taken this fact into account in the computation of index numbers. In order to allow each commodity to have a reasonable influence on the final

3. Selection of The Base

The comparison of prices in any year has to be made with the prices of the base year, which in general, is an earlier, ideal period of time during which prices may have remained more or less stable. Generally the base period should be a normal one free from fluctuations and disturbances and should not be the distant in the post.

There are two methods by which base period can be selected. They are :

i) Fixed base method

ii) Chain base method

Fixed base method

In the fixed base method a definite year or the coverage of certain number of years is chosen and the price relatives are calculated with the fixed base.

Price Relative

A price relative is a pure number, free of any unit of measure. It is usually calculated as a price in any year as compared to the base year price.

We define it as :

$$\text{PR} = \frac{\text{Commodity price in current year}}{\text{Commodity price in base year}} \text{x } 100$$

$$\text{Current year's PR} = \frac{\text{Current year's price}}{\text{Base year's price}} \text{x } 100$$

$$\text{Link relatives of the current year} = \frac{\text{Current year's price}}{\text{Base year's price}} \text{x } 100$$

4. Selection of Averages

Since index numbers are specialized averages and decision has to be made as to which particular average (i.e. arithmetic mean, median, mode or harmonic mean) should be used for construction of particular index, Geometric mean is considered to be the best average in the construction of index numbers because of its less susceptibility and reversible nature. It also gives equal weights to equal ratio of change. The geometric mean always satieties the time reversal test.

5. Selections of Appropriate Weight

The items selected and construction of an index number being of unequal importance, a simple average will not be truly representative of the state of affairs. Suitable weights therefore

7. SELECTING ON FORMULA

Different formulae are used for the construction of different index numbers. The formulae are discussed in lesson 20. Each formula is used specific purpose and it has its advantages and disadvantages.

Construction or Wholesale Price Index Numbers

The objectives of construction of wholesale price index number is to measure the relative variation in the general price level. The construction of wholesale price index numbers involves certain problems, which are not easy to tackle. The problem are as follows :

1. *Selection of Commodities*

The wholesale price index number represents the price changes of all commodities in general. But it is not possible to include all commodities. So selection must be made to include a few representative item for the purpose. The representative items should be representative of the taste, habits, customs and necessities of the people for whom the index number relates. There is no hard and fast rule about the exact number of items that a good index number should have. But is can be said that the number of items from which wholesale price index should be constructed should be fairly large consistent with ease in handling them. The general purpose index number of price issued by the Economic Adviser's Office in India takes into account 78 commodities.

The economic Advisor's Index Number of wholesale prices in our country classified the commodities in five major groups, namely :

1. Food articles,
2. Industrial Raw-materials,
3. Semi-manufactures,
4. Manufactures and
5. Miscellaneous.

2. *Obtaining Price Quotations*

Items cannot be relevant unless their price quotations justify the purpose of the index numbers. A relation must be made of representative places and persons. Data must be as accurate as possible. Money prices (quantity of money per unit of commodity) and not commodity prices should be quoted. Uniformity, regularity and reliability of price quotations should be ensured before hand. Wholesale prices are more reliable than the retail prices.

index, weights should be applied. It may however, be noted that there is nothing like an unweighted index number. Every series is weighted in a arbitrary or haphazard fashion. Obviously in them the weights are unity. Every item is considered to be of equal importance.

These are two ways of assigning weights :

i) Implicit.
ii) Explicit.

Implicit weighting implies inclusion of a commodity or its variety in the index number of times, e.g. if wheat in an index has to be given thrice as much weight as rice then three varieties of wheat as against one of rice may be included in the series. By explicit weighting we understand the application of some outward evidence of importance to the items in an index.

What this evidence should be is a difficult problem to decide. It depends upon the object of the investigation. In a general wholesale price index number, the weights may be the quantities produced or sold at wholesale rates, but such weights will be inappropriate for a cost of living index in which case the retail price should be weighted by the proportionate expenditure upon the different articles as revealed by representative family budget information collected.

Fixed or Fluctuating Weights

Another problem connected with weights is that of deciding whether they should be fixed or fluctuating. Since the relative importance of a commodity changes over a period of time it is logical of weights are revised from time to time. One should be cautious in interpreting an index constructed on fluctuating weights as not only changes in prices but also changes in weights affecting index.

Methods

When weights are applied to methods of calculating, index numbers may be used

1. Weighted aggregative method : Under this method weights are applied to actual prices.
2. Weighted average of relatives method : Under this method weights are applied to price relatives and not to actual prices.

Weighted Aggregative Index

The basic formula under this method is :

$$\frac{\sum P_n q}{\sum P_n q} \times 100$$

Taking the physical quantities (q) to the appropriate weights, the index number may be calculated by a number of formulae given by different persons who are advocates of different base period for the weights.

should be attached to the items depending on their relative importance. There are two methods of assigning weights : 1. Implicit and 2. Explicit. In the implicit weighting, a commodity or its variety is included in the index number o times. In case of explicit weighting some outward evidence of importance o the various times in the index is given.

6. Selection of Appropriate Formula

The choice of the formula would depend not only the purpose of the index but also on the data available. No particular formula can be regarded as the base under all circumstances. The discriminating investigator will choose technical methods adapted to his data and appropriate to his purpose.

Steps in the Construction of index Numbers

The steps in the construction of wholesale index numbers are as follows :

1. Select a suitable list of commodities and make arrangements for obtaining their price quotations regularly.

2. Select a base year and construct current prices into price relatives based on the prices of the base year. There can be either a fixed base or chain base. A fixed base can be either a particular year or the average of number of years.

3. Select a measure of central tendency and obtain an average from amongst mean, median and geometric mean. Geometric mean has certain advantages over other averages in the construction of index numbers.

4. If weights have to be used they can be either implicit or explicit. Explicit weighting can be done either by weighted average of relatives method or weighted aggregate method.

(The method of construction of index numbers are described in detail in Chapter – 20)

Consumer Price Index Number

MEANING

A consumer price index number is a statistical measure of changes in prices of the goods and services brought by families of wage earners and electrical workers. It measures only changes in prices. It tells nothing about changes in the kind and amounts of good and services families buy or the total amount families spent for living, or the differences in the living costs in different places. Consumer price index cost of living index numbers work as economic indicators in the same way as wholesale price index numbers to measure the purchasing power of money.

NEED

The need for constructing consumer price index arises because the general index number fail to give an exact idea of the effect of the change in the general price level on the cost of different classes of people since a given change in the level of prices affect different classes of people in different manners. This index studies the effect of changes of certain chosen combinations of goods on the standard of living of people considered as consumers

UTILITY

The consumer price index are of great significance because of following :

1. The general use of this index is mostly in wage negotiations and wage contacts. Automatic adjustments of wage or dearness allowance component of wages are governed in many countries by such index.

2. At. Government level index numbers are used for wage policy, price policy, rent control, taxation and general economic policies.

3. The index numbers are also used to measure changing purchasing power of the currency, real income etc.

4. Index numbers are also used for analyzing markets for particular kind of goods and services.

LIMITATIONS

1. Possibility of selection of wrong or non-representative items.

2. No measure of changes in the manner of living.

3. Inadequate reflection of inflationary effect block market prices changes in quality etc.

4. Arbitrary assignment of weights.

5. No representation of experience of any particular family.

6. Frequent need for recalculation.

Construction of a Consumer Price Index

These indices can be computed separately for :

i) Different classes of people,

ii) Different regions or homogeneous groups.

iii) Different time periods.

iv) Different occupation.

Each class or region has its own choice of commodities whose qualities and quantities consumed differ from time to time and place to place. Moreover retail prices, which are the basic cost of living index numbers differs from time to time and place to place.

Steps :

The following are the steps in construction the consumer price index :

1. The class of people for whom this index number is to be constructed is decided first. Along with the class of people the geographical area covered by the index is also necessary.

2. The next step is to conduct a family budget enquiry covering the population group for whom the index is to be designed.

3. The items covered are those, which are in common use by the class concerned. These items are classified into certain well accepted groups :

Namely :

i) Food

ii) Clothing

iii) Fuel and lighting

iv) House rent and

v) Miscellaneous

4. Price quotations for these items are to be obtained from the locality in which this class of people reside or from where they make their purchases.

5. Suitable weighιs are attached to different commodities depending on their importance as revealed by family budget inquiries or otherwise. Cost of index numbers are then calculated by the weighted aggregate of actual prices or the weight and average of price relatives. The methods has already been considered in Chapter – 20.

Method of Constructing the Index

The index for cost of living of consumers may be constructed by applying any of the following methods :

1. Aggregate Expenditure method or Aggregative method.

2. Family budget Method or the Method of Weighted Relatives.

What changes in the cost of living figures of 2005 are compared with 2004 are seen?

Solution of cost of living index number for 2005 with 2004 as the base.

Items of Expenditure	2004 P_0	2005 P_1	$\underline{P_1} \times 100$ P_0=PR=P	W	PW
Food	150	145	96.7	35	3383.3
Rend	30	30	100.0	15	1500.00
Clothing	75	65	86.7	20	1734.0
Fuel	25	23	92.0	10	920.0
Miscellaneous	40	45	112.5	20	2250.0
				W=100P	W=9787.3

Cost of living index $\frac{\sum PW}{\sum W} = \frac{9787.3}{100} = 97.87$. Thus a fall of 2.13% has taken place in the cost of living of middle class families in the given city of India in 2005 as compared it 2004.

SELF - TEST – 14

1. Explain the significance of Index number. Also mention their limitations.
2. Discuss the steps involved in the construction of an index number.
3. Index numbers are economic barometres "Explain this statement.
4. Discuss various problems in constructing index numbers.

METHOD OF CONSTRUCTING INDEX NUMBER

A large number of methods have been devised for constructing index numbers. They can be broadly, divided into two groups.

a) Unweighted indices, and

b) Weighted indices.

In the unweighted indices weights are not assigned whereas in the weighted indices weights are assigned to the various times. Each of these types may be further divided under two groups

i) Simple Aggregative, and

ii) Simple Average of Relatives.

(A) UNWEIGHTED INDEX NUMBERS

(1) Simple Aggregative Method

This is the simplest method of constructing index numbers when this method is used to

Article	Quantity 1960	1960	Price 1968
Food	3 mds	12.00	18.00
Cloth	12 yds	1.00	0.90
Electricity	40 units	0.20	0.25
Rent	3 rooms	25.00	23.00
Miscellaneous	16 units	0.40	0.50

(i) Aggregate Expenditure Method

Article	Q_0	P_0	P_1	P_0Q_0	P_1Q_0
F	3	12.00	18.00	36.00	54.00
C	12	1.00	0.90	12.00	10.80
E	40	0.20	0.25	8.00	10.00
R	3	25.00	23.00	75.00	69.00
M	30	0.40	0.50	12.00	15.00
				143.00	158.00

$$I = \frac{\sum P_1Q_0}{\sum P_0Q_0} \times 100 \frac{158.80}{143.00} \times 100 = 111.5$$

(ii) Family Budget Method

Article	Q_0	P_0	P_1	$PR=P_1P_0$	$V=P_0 X Q_0$	PRV
F	3	12.00	18.00	150.00	36	5400
C	12	1.00	0.90	90.00	12	1080
E	40	0.20	0.25	125.00	8	1000
R	3	25.00	23.00	12.00	75	6900
M	30	0.40	0.50	125.00	12	15880
					V=143	RV=15880

$$I = \frac{\sum RV}{\sum V} = \frac{15880}{143} = 111.5$$

The result of both the methods are identical.

Illustration – 2

A enquiry into the budgets of middle class families in a city in India gave the following information :

Expense on	Food	Rent	Clothing	Fuel	Miscellaneous
	35%	15%	20%	10%	20%
Price 2004	Rs.150	Rs.30	Rs.75	Rs.25	Rs.40
Prices 2005	Rs.145	Rs.30	Rs.65	Rs.23	Rs.45

1. *Aggregate Expenditure Method*

When this method is used, the quantities of commodities consumed by the particular group in the base year are estimated and those figures or their proportions are used as weights. The prices of commodities for various groups for the current year are multiplied by the quantitatives consumed in the base year and the aggregate expenditure incurred in buying their commodities is obtained. In a similar manner the prices of the base year and aggregate expenditure for the base period is obtained. The aggregate expenditure of the current year is divided by the aggregate expenditure of the base year and the quantity is multiplied by 100 to get the symbolically is

$$\text{Consumer price index} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

When P_1 and P_0 stand for the price of the current year and base year and q_0 for quantities in the base year.

2. *Family Budget Method*

In this method the family budget of a large number of people and for whom the index number is meant, are carefully studied and the aggregate expenditure of the average family on various items is estimated. These constitute the weights, Current year's prices are converted into price relatives on the basis of the base years prices and these price relatives of the commodities, to the base year. The total of these products is divided by the sum of the values (or weights) and the resulting figure is the desired index numbers.

Symbolically

$$\text{Consumer price index} = \frac{\sum PRV}{\sum V}$$

Where PR = P_1 x 100 for each item

V = Value weights i.e. P_0Q_0

This method is the sum as the weight and average of price relatives discusses lesson 20.

It should be noted that the answer obtained by applying the aggregate expenditure method of the family budget method shall be same.

Illustration – 1

Construct the cost of living index number from the following data for 1968 with 1960 as base using (i) the aggregate expenditure method, and (ii) the family budget method.

Illustration – 3

Find out the index number of each year from the following data.

Year	Price	Year	Price
1996	78	2001	99
1997	88	2002	102
1998	70	2003	112
1999	78	2004	99
2000	94		75

Solution

This problem can be solved by taking the first year, i.e. 1990 as the base or by taking the average price of the prices as base. The first alternative is preferable.

Base 1955 = 100

Year	Price	Index No.	Year	Price	Index No.
1999	778	100	2001	99	127
1997	88	113	2002	102	131
1998	70	90	2003	112	144
1991	78	100	2004	99	127
1997	94	121	2005	75	96

Index for 1956 = $\frac{P_1}{P_0}\times100=\frac{88}{78}=113$ approx.

Similarly others will also be calculated.

Note : Figures have been approximated to a whole number.

2. *Simple Average of relative Method*

The most popular method of constructing an index number of group of prices is to average the individual prices. The aggregative method through simple and intelligible, is not very common. We have then to decide the problem of the most appropriate average to be used. The various statistical averages that we have seen in the chapter on "Measures of Central Tendency", have their own particular merits and uses. Theoretically any average i.e. arithmetic mean, median, mode, geometric mean or harmonic mean may be used for averaging the relatives. When arithmetic mean is used for averaging the relatives, the formula for computing the index is :

$$\sum(\frac{P_1}{P_0}\times100)=\frac{88}{78}=113$$

P_{01} = N

Year	Rs.	Year	Rs.	Year	Rs.
1999	78	1995	98	2001	99
1991	54	1996	94	2002	75
1992	56	1998	78	2004	71
1993	72	1999	76	2005	50
1994	102	2000	112		

Solution

Index number of jute price base 1994 = 100

Year	Price in Rs.	Index No. x 100	Year	Price in Rs.	Index No. x 100
1990	78	77	1999	78	77
1991	54	53	2000	76	75
1992	67	66	2001	112	110
1993	56	55	2002	99	97
1994	72	71	2003	76	75
1995	102	100	2004	75	74
1996	98	96	2005	71	70
1997	94	92	2006	50	49
1998	88	86			

Index number for 1989 with 1994 as base

$$\frac{P_1}{P_0} \times 100 = \frac{78}{102} \times 100 = 77 \text{ (approx)}$$

Index number for 1990 with 1994 as base

$$\frac{P_1}{P_0} \times 100 = \frac{54}{102} = 53$$

Note ; The figures are approximated to a whole number.

This problem can be solved by taking the first year, i.e. 1955 as the base or by taking the average price of the prices as base. The first alternative is preferable.

Base 1955 = 100

Similarly others will also be calculated.

Note : Figures have been approximated to a whole number.

construct a price index. The total of current year price for the various commodities in question is divided by that total of base year prices and the quotient is multiplied by 100. Symbolically

$$P_{01} = \frac{\sum P_1}{\sum P_0} 1 \times 100$$

ΣP_1 = Total of current year prices of various commodities.

ΣP_0 = Total of base year prices of various commodities

Illustration – 1

From the following data construct an index for 2005 taking 2004 as base.

Commodities	Prices in 2004 (Rs.)	Prices in 2005 (Rs.)
A	100	140
B	80	120
C	160	180
D	220	240
E	40	40

Solution :

Construction of price index.

Commodities	Prices in 2004 p_0	Prices in 2005 P_1
A	100	140
B	80	120
C	160	180
D	220	240
E	40	40
	P_0=600	P_1=720

Index number for 2004 with 2005 as base = $\frac{720}{600} \times 120$

This means that as compared to 2004 in 2005 there is net increase in the prices of commodities are given effect in the price index. We construct index numbers by this method on the assumption that the various times and their prices are quoted in the same unit.

Illustration - 2

In the following table the wholesale prices of jute from 1989 to 2005 are given. Construct numbers taking 1991 as base year.

Where No refers to the number of items (commodities whose price relatives are thus averaged when geometric mean is used for averaging the price relatives the formula for obtaining the index becomes.

$$\text{Long } P_{01} = \frac{\sum \log \frac{(P_1 \times 100)}{P_0}}{N} \times 100 = \frac{88}{78} = 113 \quad \text{or } \sum \log \frac{P}{P} \, who \, \frac{P'}{P_0} = 1 \times 100$$

$$P_{01} = \text{Anti log } \frac{(\sum \log \frac{P_1}{P_0} \times 100)}{N} = \text{Anti log } \Sigma \log P$$

Other measures of central value are not in common use for averaging relatives.

Illustration – 4

From the following data construct for 2005 taking 2004 as base by the average of relatives method using (a) arithmetic mean and (b) geometric mean for average relatives.

Commodities	Price in 2004 Rs.	Price in 2005 Rs.
A	100	140
B	80	120
C	160	280
D	220	240
E	40	40

Solution : (a) Index number using arithmetic means of price relatives.

Commodities	Price in 2004 Rs.	Price in 2005 Rs.	Price relatives x 100
A	100	140	140.0
B	80	120	150.0
C	110	180	112.5
D	220	240	109.1
E	40	40	100.0

$$\frac{\sum P_1}{P_0} \times 100$$

$$P_{01} = \quad = \frac{616.6}{5} = 123.32 \frac{\sum P_1}{P_0} \times 611.6$$

(b) Index number using geometric mean or price relatives.

Commodities	Price in 2004 Rs.	Price in 2005 Rs.	Price relatives	Log P
A	100	140	140.0	2.1461
B	80	120	150.0	2.1761
C	160	280	112.5	2.0512
D	220	240	109.1	2.0376
E	40	40	100.0	2.0000
				? log P = 10.4112

P_{01} = Anti log $(\frac{\log P}{N})$ = Anti log $(\frac{10.4112}{N})$ = Anti log 2.0822 = 120.9

Weighted index Numbers

Items (commodities) vary in importance. So far, we have not taken this fact into account in the computation of index numbers. In order to allow each commodity to have a reasonable influence on the final index, weights should be applied. It may however, be noted that there is nothing like an index number. Every series is weighted in an arbitrary or haphazard fashion. Obviously in them the weights are unity. Every item is considered to be of equal importance.

There are two ways of assigning weights.

i) Implicit

ii) Explicit

Implicit weighting implies inclusion of a commodity or its variety in the index number of times, e.g. if what in an index has to be given thrice as much weights as rice than three varieties of wheat as against one of rice may be included in the series. By explicit weighting we understand the application of some outward evidence of importance to the items in the index.

Implicit weighting (or the unweighted index) is far from realistic in most of the cases. Construction of useful index numbers requires a conscious effort to assign to each commodity a weight in accordance with importance in the total phenomenon that the index is supposed to describe. Weighted index numbers are of two types :

1. Weighted Aggregative indices, and
2. Weighted Average of Relatives indices.

1) *Weighted Aggregative Method*

It applied to actual prices. There are various method of assigning weights and consequently a large number of formulae for constructing index numbers have been devised of which some of the more important ones are :

1. Laspeyres, Method : Using base period quantities as weights (q_0)

$$P_{01} = \frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100$$

Steps :

(i) Multiply the current years prices of various commodities with base year weights and obtain $\Sigma P_1 q_0$.

(ii) Multiply the base year prices of various commodities with base year weights and obtain $\Sigma P_0 q_0$.

(iii) Divide $\Sigma P_1 q_0$ by $P_0 q_0$ and multiply the quotient by 100. This gives us the price index.

2. Paasche's Method : Using current years quantities (q_1 weights)

$$I_{01} = \frac{\sum P_1 q_1}{\sum P_0 q_1} \times 100$$

Steps :

(i) Multiply the base year prices of various commodities with current year weights and obtain $\Sigma P_0 q_1$

(ii) Multiply the base year prices of various commodities with current year weights and obtain $\Sigma P_0 q_1$.

(iii) Divide $\Sigma P_1 q_1$ by $\Sigma P_0 q_1$ and multiply the quotient by 100.

This gives us the prices index.

3. Marshall – Edge Worth's Method : Using the average (or total) quantities of base and current years as weights. This is also known as the "Hybrid-Weights" formula :

$$P_{01} = \frac{\sum (q_0 + q_1) P_1}{\sum (q_0 + q_1) P_1} \times 100$$

or, opening the brackets :

$P_{01} = \Sigma P_1 q_0 + \Sigma P_1 q_1 \times 100$

Solution :

Calculation of Various Indices

Commodities	1995 Price	1995 Qty	2005 Price	2005 Qty.	P_1q_0	P_0q_0	P_1q_1	P_0q_1
A	4	16	8	12	128	64	96	48
B	10	20	12	10	240	200	120	100
C	8	28	10	20	280	224	200	160
D	4	38	4	26	152	152	104	140
				$\Sigma P_1q_0=$ 800	$\Sigma P_0q_0=$ 640	$\Sigma P_1q_1=$ 520	$\Sigma P_0q_1=$ 412	

1. Laspaye's Method

$$P_{01}=\frac{\sum P_1q_0}{\sum P_0q_0}\times 100$$

$$=\frac{180}{640}\times 100=125$$

2. Passche's Method

$$P_{01}=\frac{\sum P_1q_1}{\sum P_0q_1}\times 100$$

$$=\frac{520}{612}\times 100=126.2$$

3. Bowley's Method

$$P_{01}=\frac{\frac{\sum P_1q_0}{\sum P_0q}+\frac{P_1q_1}{P_0q_1}}{2}\times 100\frac{\frac{800}{640}+\frac{520}{412}}{2}\times 100$$

$$=\frac{1.25+1.262}{2}\times 100\frac{2.512}{2}\times 100=125.6$$

$$or P_{01}=\frac{L+P}{2}=\frac{125+262}{2}=125.6$$

4. Fisher's Ideal method

$$P_{01}=\sqrt{\frac{P_1q_0}{P_0q_0}\times\frac{P_1q_1}{P_0q_1}}=\sqrt{\frac{800}{640}\times\frac{520}{412}}\times 100$$

$$=1.25\times 1.262\times 100$$

$$=1.256\times 100$$

$$=125.6$$

$$P_{01} = \frac{\sum P_1 q}{\sum P_0 q} \times 100 \text{ (Where q} = \frac{q_0 + q_1}{2} = \text{average quantity)}$$

8. Fisher's 'ideal' index Prof. Irving – Fisher has given a number of formulae for constructing index numbers and of those he calls one as the 'ideal' index. The Fisher's Ideal Index is given by the formula-

$$P_{01} = \sqrt{\frac{P_0 q_0}{P_0 q_0} \times \frac{P_1 q_1}{P_0 q_1}} \times 100 or P01 = \sqrt{L \times p} \text{ (W = weights)}$$

This method uses base and current year weights for preparing two index numbers (each with different weights), which are averaged geometrically. This is also called crossed Weight formula. In other words it is the geometric mean of the Laspayre's and Passche's indices.

This above formula is known as ideal because of the following reasons :

i) It is based on the geometric mean which is theoretically considered to be the best average for constructing index numbers.

ii) I takes into account both current and base year prices and quantities.

iii) It satisfies both the time reversal test as well as factor reversal test as suggested by Fisher

iv) It is free from bias.

Illustration - 4

Construction index numbers of price from the following data by applying.

1. Laspeyer's Method
2. Paasche's Method
3. Fisher's ideal Method, and
4. Marshall-Edge worth Mathod.

Commodities	1995 Price	Quantity	Price	2005 Quantity
A	4	16	8	12
B	10	20	12	10
C	8	28	10	20
D	4	38	4	26

4. Drobisch and Bowley's Method : They have suggested that instead of finding out the geometric mean of the index numbers computed from Laspeyres' and Paasche's method their arithmetic mean would give a better result, i.e.

$$P_{01} = \frac{L-I-P}{2}\times 100 \quad \text{L = Laspeyres' Index}$$

P = Paasche's Index

$$\frac{\left(\frac{\sum P_1Q_0}{\sum P_0Q_0}\right)+\left(\frac{\sum P_1Q_1}{P_0Q_1}\right)}{2}\times 100$$

5. Walsh's Method : Wash has suggested a modification in Marshal-Edge Wroth's Method. He takes into account the geometric mean and not the arithmetic mean in the construction of index number, this formula is

$$P_{01} = \sum\sqrt{q_0(q_1p_1})\times 100$$
$$\sum\sqrt{q_0(q_1p_0)}$$

6. Keynes's Method : Keynes takes into account the quantities of each commodity common to each year as weights and then suggests the preparation of an index number, i.e.

$$P_{01}\frac{\sum P_1qH.C.F}{\sum P_0qH.C.F}\times 100 \quad \text{(H.C.F=Highest Common Factor)}$$

7. Kelly's Method : Truman L Kelly has suggested the use of selected weights which need not necessarily be equal to the quantities marked or consumed. Here weights are the quantities, which may refer to some period, not necessarily the base year or current year. Thus, the average quantity of the two or more years may be used as weights.

This formula is :

$$P_{01} = \frac{\sum P_1W}{\sum P_0W}\times 100 \quad \text{(W = weights)}$$

or

$$P_{01} = \frac{\sum P_1q}{\sum P_0q}\times 100 \quad \text{(= quantities)}$$

QUANTITY INDEX NUMBERS

A quantity index number is the counterpart of price index number. Unlike the latter it measures the changing value of a varying aggregate of goods at fixed prices. It is indicative of changed in physical quantities at constant prices, computational procedure for such an index is the same as in the case of price index but for the fact that the words "Quantity" and "Price" are interchanged. Symbolically. When Laspeyres' method is used.

$$Q_{01} = \frac{\sum Q_1P_1}{\sum Q_0P_0} \times 100$$

When Paasche's formula is used

$$P_{01} = \frac{\sum P_1q_0}{\sum P_0q_0} \times 100$$

When Fisher's formula is used $Q_0 = \sqrt{\frac{\sum Q_1P_1}{\sum q_0P_0} \times \frac{\sum q_1p_1}{\sum q_0P_1} \times 100}$

The index number of price measures the changing value of a fixed quantity of goods while the number of quantity measures the changing value of varying aggregated of goods at fixed prices.

Illustration - 7

Form the following data compute a quantity index :

Commodity	Quantity		Price in 177 Rs.
	2004	2005	
Rice	60	50	60
Wheat	40	60	80
Jawar	20	30	40

Solution

Computation of quantity index.

Commodities	Q_0	Q_1	P_0	Q_1P_0	Q_0P_0
Rice	60	50	60	3000	3600
Wheat	40	60	80	4800	3200
Jawar	20	30	40	1200	800
				$\Sigma Q_1P_0 = 9000$	$\Sigma Q_0P_0 = 7600$

$$Q_{01} = \frac{Q_1P_0}{Q_0P_0} \times 100 \ \frac{90000}{7600} \times 100 = 118.4$$

Thus compared to 2004 the quantity index has gone up by 18.4 percent in 2005.

1. Express the price of each item of the current year as a relative of the price of the same item in the base year by the formula :

$$\frac{P_1}{P_0}\times 100$$

2. Assign weights – which may be taken as the total quantity on value (Quantity % price)
3. Multiply each price relative by its corresponding weight.
4. Sum up the price relatives thus obtained.
5. Divide this sum by the total of weights. The result is the required number.

Using the total expenditure in base year (q_0 x P_0) as weight, the formula for index number is :

$$P_{01} = \frac{\sum P_0 q_0 = \frac{P_1}{P_0}\times 100}{\sum P_0 q_0}$$

Illustration – 6

From the following data compute price index by applying weighted avera e of price relatives method.

Solution

Computation of index number of prices by the method of weighted average of Relatives.

Commodities	Price P_0	2004 Qty. q_0	Base year Weights (P_0Q_0)	Price Relative P_1	2005 Price	Current Weights x Price
				$\frac{P_1}{P_0}\times 100$		$(P_0\, q_0)\frac{P_1}{P_0}\times 100$
E	4	10	40	8	200	8000
F	6	20	120	9	150	18000
G	10	5	50	5		2500
Total			210			28500
Index No.						135.7

From the above Table, we get,

$$\sum[(P_0 q_0 \times \frac{P1}{P0}\times 100$$

and $\Sigma\,(P_0Q_0) = 210$

Thus Price Index Number = P 2005 = $\frac{28500}{210} = 135.7$

QUANTITY INDEX NUMBERS

A quantity index number is the counterpart of price index number. Unlike the latter it measures the changing value of a varying aggregate of goods at fixed prices. It is indicative of changed in physical quantities at constant prices, computational procedure for such an index is the same as in the case of price index but for the fact that the words "Quantity" and "Price" are interchanged. Symbolically. When Laspeyres' method is used.

$$Q_{01} = \frac{\sum Q_1P_1}{\sum Q_0P_0} \times 100$$

When Paasche's formula is used

$$P_{01} = \frac{\sum P_1q_0}{\sum P_0q_0} \times 100$$

When Fisher's formula is used $Q_0 = \sqrt{\frac{\sum Q_1P_1}{\sum q_0P_0} \times \frac{\sum q_1p_1}{\sum q_0P_1} \times 100}$

The index number of price measures the changing value of a fixed quantity of goods while the number of quantity measures the changing value of varying aggregated of goods at fixed prices.

Illustration - 7

Form the following data compute a quantity index :

Commodity	Quantity		Price in 177 Rs.
	2004	2005	
Rice	60	50	60
Wheat	40	60	80
Jawar	20	30	40

Solution

Computation of quantity index.

Commodities	Q_0	Q_1	P_0	Q_1P_0	Q_0P_0
Rice	60	50	60	3000	3600
Wheat	40	60	80	4800	3200
Jawar	20	30	40	1200	800
				ΣQ_1P_0= 9000	ΣQ_0P_0 = 7600

$$Q_{01} = \frac{Q_1P_0}{Q_0P_0} \times 100 \ \frac{90000}{7600} \times 100 = 118.4$$

Thus compared to 2004 the quantity index has gone up by 18.4 percent in 2005.

Factor Reversal Test

Fisher says "Just as our formula should permit the interchange of the two times without giving inconsistent results, so to ought to permit inter changing the prices and quantities without given in consistent result, i.e. the two results multiplied together should give the true value ratio that is :

$$P_{01} \times Q_{01} = \frac{\sum P_1 q_0}{\sum P_0 q_0}$$

The product of current year price index (on base year) and current year quantity index (on base year) should be equal to the ratio of the total current year value 9P$_1$q$_1$) to the total base year value (P$_0$q$_0$) value being equal to price x quantity.

$$\text{Price index } P_{01}\sqrt{\frac{\sum P_1 Q_0}{\sum P_0 q_0} \times \frac{\sum P_1 Q_1}{\sum P_0 q_1}}$$

Test is satisfied by the Fisher's ideal index. Its practical application is however, limited as it requires current weight for the year for which the index is to be constructed' these are generally not available.

Circularity Test

Another important test which a good index number should satisfy is the circularity test. This may be explained with the help of a simple example i.e.

If suppose the 2005 price of a certain commodities is thrice of what it was in 1999 and its 1999 price is again twice of what it was in 1991, then its 2005, then its 2005 price must be nine times its price of 1991. In other words, this means that we should be able to get an index for 2005 based on 1991 by multiplying the index for 2005 relative to 1999 by the corresponding index for 1999 based on the year 1991 an index which complies with this test has the advantage of reducing the computation every time a change has to be made. Such index numbers can be adjusted from year to year without referring each time to the original bases. Simple aggregate index represented by $\Sigma P_1 / \Sigma P_0$ (without the factor hundred) where P_1 represents prices of current years and P_0 prices of base year satisfies this test. This test therefore, is :

$$P_{01} \times P_{..} \times P_{20} = 1$$

It is an extension of the Time reversal Test and applies ot indices of more than two years. In general terms :

$$P_{12} \times P_{23} \times P_{34} \times \ldots \ldots\ldots P_{n-1} \times P_{n1} = 1$$

Mathematical Tests

Professor Irving Fisher in his famous work "The Making of Index Numbers" has done commendable work for developing a scientific theory of index Numbers. His formula known as Fisher's "ideal" formula is mathematically the most satisfactory one. The test its dependability and accuracy two tests are applied.

i) The time (or Base0 Reversal Test.

ii) The factor Reversal Test.

Time Reversal Test

Regarding this test, Fisher says : "The test is hat the formula for calculating an index number should be such that it will give the same ratio between one point of comparison and the other, no matter which of the two is taken as the base". In other words this implies that by changing the base of an index number from one year to the other, no alternation should take place in the relative magnitude of the two indices.

Symbolically : $P_{01} \times P_{10} = 1$

When P_0 is the index for time "1" no time "0" as base and P_0 is the index for time "0" on time "1" as base. If the product is not unity, there is said to be a time bias in the method. Thus if from 2004 to 2005 the price of wheat increased from Rs. 120 to Rs. 140 per quintal the price in 2005 the price of wheat increased from Rs. 120 to Rs. 140 per quintal the price in 2005 should be 133 1.3 percent of the price in 2004 should be 75 percent of the price in 2005. One figure is the reciprocal of the other, their product (1.333 x 75) is unit. This is obviously true for each individual price relative and according to the time reversal test it should be true for the index number.

Let us see how Fishr's Ideal formula satisfies the test.

$$P_{01} = \sqrt{\frac{\sum P_1Q_1}{\sum P_0q_0} \times \frac{\sum P_1Q_1}{\sum P_0q_1}}$$

Since $P_{01} \times P_{10} = 1$ the Fisher's ideal index satisfies the test.

$$Q_{01} = \frac{\sum P_1q_0}{\sum P_0q_0} \times \frac{\sum P_0q_0}{\sum P_1q_0}$$

$$P_{01} \times q_{01} = \sqrt{\frac{\sum P_1q_0}{\sum q_0P_0} \times \frac{\sum P_1q_1}{\sum PqP_1} \times \frac{\sum P_0q_1}{\sum P_0q_0} \times \frac{P_0q_0}{P_1q_1} = \sqrt{1 = 1}}$$

Since $P_{01} \times P_{20} = 1$ the Fisher's ideal index stratifies the test.

ustration - 8

Compute by suitable method the index number o quantities from the data given below

Commodities	2004 Price	Total value	2005 Price	Total value
X	16	160	20	220
Y	20	180	24	216
Z	32	512	40	680

Solution

Since we are given the value and price we can obtain quantity figure by dividing value figures by price for each commodity. We can then apply fisher's for finding out quantity index

Computation of quantity index by Fisher's Method

Commodities	2004	2005						
	P_0	Q_0	P_1	Q_1	Q_1P_0	Q_0P_0	Q_1P_1	Q_0P_1
X	16	10	20	11	176	160	220	200
Y	20	9	24	9	180	180	216	216
Z	32	16	40	17	544	512	680	640
					$\Sigma Q_1P_0 =$ 900	$\Sigma Q_0P_0 =$ 852	$\Sigma Q_1P_1 = 116$	$\Sigma Q_0P_1 = 1056$

Quantity Index or $Q_0 = \sqrt{\frac{\sum Q_1P_1}{\sum q_0P_0} \times \frac{\sum q_1P_1}{\sum q_0P_1} \times 100}$

$$= \sqrt{\frac{900}{852} \times \frac{1116}{1056} \times 100}$$

$$= \sqrt{1.116 \times 100}$$

$$= 1.056 \times 100$$

Value index number :

The value of a single commodity is the product of its price quantity. Thus a value index V is the sum of the values of given year divided by the sum of he values of the year.

The formula, therefore, is :

$$V = \frac{\sum Q_1P_1}{\sum P_0Q_0} \times 100 \text{ (V = Value Index)}$$

Where P_1Q_1 = Total value of all commodities in the given period and P_0Q_0 = Total value of all commodities in the base period. Since in most cases the value figures are given, the formula can be stated more simply.

$V \frac{\sum V_1}{\sum V_0}$ in which V stands for value.

Illustration – 9

The following figures relate to the prices and quantities of certain commodities, construct an appropriate index number and show if it satisfies the time reversal test.

Commodities	2004		2005	
	Price	Quantities	Price	Quantities
Wheat	60	100	64	100
Gram	50	80	60	70
Barley	36	100	32	110

Solution Index number by Fisher's ideal method

Commodities	2004		2005					
	P_0	Q_0	P_1	Q_1	P_1Q_0	P_0Q_0	P_1Q_1	P_0Q_1
Wheat	60	100	64	100	6400	6000	6400	6000
Gram	50	80	60	70	4800	4000	4200	3500
Barley	36	100	32	110	3200	3600	3500	3960
					P_0Q_0	P_0Q_0	P_1Q_1	P_0Q_1

$$P_{01}\sqrt{\frac{\sum P_1Q_0}{\sum P_0q_0}\times\frac{\sum P_1Q_1}{\sum P_0Q_1}}$$

$$=\times P_{01}\sqrt{\frac{14400}{13600}\times\frac{14120}{13460}}\times 100$$

= 1.111 x 100 = 1.054 x 100 = 105.4

Time reversal test is satisfied when P_{10} x P_{10} = 1

Substituting the values of Σ P_1Q_0 Σ P_0Q_0 etc.

$$P_{01}\frac{\sum P_1Q_0}{\sum P_1Q_1}\times\frac{\sum P_0Q_0}{\sum P_1Q_0}$$

$$=\sqrt{\frac{13.450}{14.120}\times\frac{13600}{14400}}$$

$$=P_{01}\times P_{01}=\sqrt{\frac{14400\times14120\times13460\times13600}{13600\times13460\times14120\times14400}}=\sqrt{1}=1$$

Hence time reversal test is satisfied by the above formula.

Illustration – 10

Prove using the following data that the Factor Reversal Test is satisfied by the Fisher's ideal formula for index number.

Commodity	Base Year		Current Year	
	Price	Quantity	Price	Quantity
A	6	50	10	56
B	2	100 ·	2	120
C	4	60	6	60
D	10	30	12	24
E	8	40	12	30

Solution Fisher's ideal index numbers for the current eyar.

Commodity	Base Year		Current Year					
	Price	Qnty.	Price		Quantity			
	P₁	Q₁	P₁	Q₁	P₁Q₁	P₁Q₁	P₁Q₁	P₁Q₁
A	6	50	10	56	300	500	336	560
B	2	100	2	120	200	200	240	240
C	4	60	6	60	240	360	240	260
D	10	30	12	24	300	360	240	288
E	8	40	12	36	320	480	288	432
					1360	1900	1344	1880

Fisher's Ideal Index No. $P_{01}\sqrt{\frac{\sum P_1Q_0}{\sum P_0Q_0}\times\frac{\sum P_1Q_1}{\sum P_0Q_1}}\times 100$

$$=\frac{1900}{1360}\times\frac{1880}{1344}\times 100$$

$$=\sqrt{1395\times 1.395\times 100}$$

$$=139.5 \text{ x } 100$$

$$= 139.5$$

Factor reversal test.

$$P_{01}\times Q_{01}=\frac{\sum P_1Q_1}{\sum P_0Q_0}$$

The formula

According to Fisher $P_{01}\frac{\sum P_1Q_1}{\sum P_0Q_0}\times\frac{\sum P_1Q_1}{\sum P_0Q_1}$

And $Q_{01}\frac{\sum Q_1Q_0}{\sum Q_0P_1}\times\frac{\sum P_1Q_1}{\sum Q_0Q_1}$

Substituting the values, $P_{01} \times Q_{01}$ $= \frac{1900}{1360} \times \frac{1880}{1344} \times \frac{1344}{1360} \times \frac{1880}{1990}$

$$= \frac{1880 \times 1880}{1360 \times 1360} = \frac{1880}{1360}$$

The factor reversal test is satisfied.

THE CHAIN INDEX NUMBERS

Under the fixed base method the year or the period of years to which all other prices are related is constant for all times. This is a serious defect. With the passage of time the base becomes distant and outmoded. It is practically useless for comparison, Tastes. habits and customs changes, and scientific progress demands dropping of old and obsolete articles and uncles and inclusion of new ones in the index. All this requires alternation in the base period list. A better comparison is obtained by using the chain base under which prices of a year linked with the proceeding year and not with any fixed year. Thus the index for 2005 will be based on 2003 that of 2005 on 2004 and so no. Symbolically :

$$P_{01} = \frac{1}{n}\left(\sum \frac{P_1}{P_0} \times 100\right)$$

$$P_{12} = \frac{1}{n}\sum\left[\frac{P_2}{P_1} \times 100\right]$$

$$P_{23} = \frac{1}{n}\sum\left[\frac{P_3}{P_2} \times 100\right] \text{ and so on}$$

Thus in its simplest from, the chain indexes are in which the figures for each year (or sub-period thereof) are first expressed as percentages of the preceding year. These percentages are then chained together by successive multiplication to form a chain index. Business – men and others are often interested in making comparisons with the previous year and not with any distant past. Link relatives obtained by the chain base method serve this purpose. If, however, it is desired to convert these relatives to a common base the results may be chained to obtain chin relatives.

STEPS IN CONSTRUCTING A CHAIN INDEX

i) Express the figures for each year as percentages of the preceding year. The results so obtained are called link relatives.

Solution

Year	Fixed Base Indices	Converted to chain base indices.	
			= 100
2000	376	$\frac{392\times100}{376}=104.3$	
2001	392	$\frac{408\times100}{392}=104.1$	
2002	408	$\frac{380\times100}{408}=93.1$	
2004	392	$\frac{392\times100}{3800}=103.2$	
2005	400	$\frac{400\times100}{392}=102.0$	

Formula used = $\frac{P_1}{P_0}\times100$

The 0 is not fixed but changing. It is just the previous eyar every time

Illustration - 3

Form the chain base index numbers.

2000	2001	2002	2003	2004	2005
92	102	104	98	103	101

Solution

Year	Chain base indices Converted to fixed base indices base 1958		
2000	92		92.0
2001	102	$\frac{102\times92}{100}$	=93.84
2002	104	$\frac{104\times93.84}{100}$	=97.59
2003	98	$\frac{98\times97.59}{100}$	=98.51
2004	103	$\frac{103\times95.64}{100}$	=98.51
2005	101	$\frac{101\times98.51}{100}$	=99.50
Formula used	Current year's chain base indices x Previous year's fixed base indices.		

(b) Construction of Chain indices.

Year	Price of wheat	Link Relations	Chain Indices 1963 = 100
1996	100	100.0	100.0
1997	120	120 x 100 = 120.00 100	120 x 100 = 120 100
1998	124	124 x 100 = 103.33 120	103.33 x 120 = 124 100
1999	130	130 x 100 = 104.84 124	104.84 x 124 = 130 124
2000	140	140 x 100 = 107.69 130	107.69 x 130 = 140 100
2001	156	156 x 100 = 111.43 140	11.43 x 140 = 156 100
2002	164	164 x 100 = 105.13 156	105.13 x 156 = 164 100
2003	168	168 x 100 = 102.44 164	102.44 x 164 = 168 100
2004	176	176 x 100 = 104.76 163	104.76 x 163 = 176 100
2005	180	180 x 100 = 102.27 176	102.27 x 176 = 180 100

Illustration - 2

From the fixed base index numbers given below, prepare chain base index numbers :

2000	2001	2002	2003	2004	2005
376	392	408	380	392	400

$$\text{Link relative} = \frac{\text{Current year's Price}}{\text{Previous Year Price}} \times 100$$

ii) Chain together these percentages by successive multiplication to form a chain index of any year is the average link relative of that year multiplied by chain index of previous year divided by 100. in the form of formula.

Chain index for Current year

= Average link relative Chain index of current
year x previous year 100

Illustration - 1

From the following data of the wholesale prices of wheat for the years construct index numbers taking (a) 1996 as base, and (b) by chain base method.

Year	Price of wheat year (Rs. per quintal)	Year	Price of wheat (Rs. per quintal)
1996	100	2001	156
1997	120	2002	164
1998	124	2003	168
1999	130	2004	176
2000	140	2005	180

Solution (a) Construction of index numbers taking 1996 as base.

Year	Price of wheat	Index Nos. (1963 = 100)	Year	Price of wheat	Index Nos. (1963 = 100)
1996	100	100	2001	156	$\frac{156}{156} \times 100 = 100$
1997	120	$\frac{120}{100} \times 100 = 124$	2002	164	$\frac{164}{100} \times 100 = 164$
1998	124	$\frac{124}{100} \times 100 = 124$	2003	168	$\frac{168}{100} \times 100 = 168$
1999	130	$\frac{130}{100} \times 100 = 130$	2004	176	$\frac{176}{100} \times 100 = 176$
2000	140	$\frac{140}{100} \times 100 = 140$	2005	180	$\frac{180}{100} \times 100 = 180$

100

For example for 2002 $\frac{104 \times 93.84}{100} = 97.59$

Use the following data in industrial production of India to compare annual fluctuations in the India industnal activity by the chain base method.

Year	1991-92	1992-93	1993-94	1994-95	1995-96	1996-97
Index	120	122	116	120	120	137
Year	1997-98	1998-99	1999-00	2000-01	2001-02	2002-03
Index	136	146	156	137	162	149
Year	1993-01	2004-05				
Index	160	160				

Solution

The given index numbers are fixed base index numbers. Their base year is a year previous to 1988-1992. We have to convert it to chain base index, i.e. link relatives have to be obtained.

Year	Index No.	Link relatives	Year	Index No.	Link relatives
188-92	120	100.0	1997-99	149	$\frac{149}{136} \times 100 = 109.6$
1992-93	122	$\frac{122}{120} \times 100 = 101.7$	1999-00	156	$\frac{156}{149} \times 100 = 104.7$
1993-94	116	$\frac{116}{122} \times 100 = 95.1$	2000-01	137	$\frac{137}{156} \times 100 = 87.8$
1994-95	120	$\frac{120}{116} \times 100 = 103.4$	2001-02	162	$\frac{162}{137} \times 100 = 118.3$
1995-96	120	$\frac{120}{120} \times 100 = 100.0$	2002-03	149	$\frac{149}{162} \times 100 = 92.0$
1996-97	137	$\frac{127}{120} \times 100 = 114.2$	2003-04	160	$\frac{160}{149} \times 100 = 107.4$
1997-98	136	$\frac{136}{137} \times 100 = 99.3$	2004-05	160	$\frac{160}{160} \times 100 = 160.0$

Illustration - 5

Given below are the prices of wheat for six years. Calculate the price relatives (a) taking 2000 as base, and 9b) taking averae price for six years as base :

Year price of wheat in Rs. per Quintal	2000	2001	2002	2003	2004	2005
	40	50	45	55	65	105

Solution

Year	Price per quintal	Price relatives 1963=100	Price Relatives Average = 100 (60=100)
2000	40	100.0	$\frac{40}{60}\times 100 = 66.7$
2001	50	$\frac{50}{40}\times 100 = 115.0$	$\frac{50}{60}\times 100 = 83.3$
2002	45	$\frac{45}{60}\times 100 = 111.5$	$\frac{45}{60}\times 100 = 75.0$
2003	55	$\frac{55}{40}\times 100 = 137.5$	$\frac{55}{60}\times 100 = 91.7$
2004	65	$\frac{65}{40}\times 100 = 162.5$	$\frac{65}{60}\times 100 = 108.3$
2005	105	$\frac{105}{40}\times 100 = 262.5$	$\frac{105}{60}\times 100 = 175.0$

Illustration - 6

Calculate the fixed base index numbers and chain base index numbers and chain base index numbers from the following data. Are the two results some? If not, why?

Commodity	Price in Rupees				
	2001	2002	2003	2004	2005
A	2	3	5	7	8
B	8	10	12	4	18
C	4	5	7	9	12

Solution

Since base year is not specified the first year in order of time i.e 2001 is taken as base. As no weights are given the appropriate method for calculating fixed base index numbers is the price relative method.

So for fixed base index numbers we have :

Commodity	2001	2002	2003	2004	2005
A	100	150	250	350	400
B	100	125	150	50	225
C	100	125	175	225	300
Total	300	400	575	625	925
Average, i.e fixed base Index no.	100	133.3	191.7	208.3	308.3

Chain base Index number chained to 2001

Group	Percentage based on preceding year				
	2001	2002	2003	2004	2005
A	100	150	166.7	140.00	114.3
B	100	125	120.0	33.33	450.0
C	100	125	140.0	128.60	133.3
Total of link relatives	300	400	426.7	301.63	697.6
Average	100	133.33	142.23	100.64	232.53
Chain indices 1974 = 100	100	133.33	189.64	190.85	443.78

On comparison we find that except for first two years, the two series of index numbers obtained by fixed base and chain base index numbers are computer by combining two or more series chain index numbers will be usually different from fixed base index numbers excepted for the first two given years.

Conversion of Chain Index to fixed Base Index

The following procedure is followed to convert thé chain base index numbers into fixed base index numbers :

(i) For the first year the fixed base index will be taken the same as the chain base index, However, If the index numbers are to be constructed by taking first years as the base then in that case in index for the first year is taken as 100.

(ii) For calculating the indices for other years the following formula is used:

$$\frac{\text{Current years chain base indices x previous years fixed base indices}}{100}$$

= Current years fixed base index. Or

$$\frac{\text{Current years C.B.I x Previous years F.B.I}}{100}$$

= Current years F.B.I

Illustration - 8

The following index numbers of prices (1996 = 100)

Year	1996	1997	1998	1999	2001	2001	2003
Index no.	100	110	120	200	400	401	408

Year	2003	2004	2005
Index no.	380	370	340

Shift the base from 1996 and recast the index numbers.

Solution

Index numbers with 2002 as base (2002 = 100)

Year	Index. Nos. (1959=100)	Index Nos. (1965=100)	Year	Index. Nos. (1959=100)	Index Nos. (1965=100)
1996	100	$\frac{100}{400}\times 100 = 25.0$	2001	410	$\frac{410}{400}\times 100 = 102.5$
1997	110	$\frac{110}{400}\times 100 = 27.5$	2002	400	$\frac{400}{400}\times i00 = 100.0$
1998	120	$\frac{120}{140}\times 100 = 30.0$	2003	380	$\frac{380}{400}\times 100 = 95.0$
1999	200	$\frac{200}{400}\times 100 = 50.0$	2004	370	$\frac{370}{400}\times 100 = 92.5$
2000	400	$\frac{400}{400}\times 100 = 100.0$	2005	340	$\frac{340}{400}\times 100 = 85.0$

The new series with 2002 as base is obtained very easily by dividing each entry of the first column by 400, i.e. the values of the index for 2002 and multiplying the ratio by 100. Thus under number for 1993.

$$= \frac{\text{Index number for 1996}}{\text{Index number for 2002}}\times 100 = \frac{100}{400}\times 100 = 25.0$$

$$\text{Index number for 1997} = \frac{\text{Index number for 1997}}{\text{Index number for 2002}}\times 100 = \frac{100}{400} = 27.5$$

In the similar other indices can be obtained.

possible to apply in all cases. Another methods may be followed which gives nearly the same results when arithmetic mean is used for averaging and gives exactly the same result as the first when geometric mean is used for averaging. The method is under :

Divide each index number or the series by the index number of the time period selected as new base and multiply the result so obtained by 100. The figure thus obtained will give required series with the new base.

Let us explain it further with the help of an example.

Illustration - 7

The index numbers for various years with 1984 as base for a certain commodity are as follows :

Year	1984	1985	1990	1995	2000	2005
Index no.	10	111	126	150	162	180

Shift the base to the year 1995

Solution

Let us make the calculation for 2000. If 1995 is to be the new base, its index number must be 100. But in the old series it is 150 and index number for 2000 is 162. The problem stated in simple terms is to determine the index for 2000 where the index for 2000 is changed from 150 to 100 i.e. If the figure for 1995 is 150, the figure for 2000 = 162.

If the figure for 1995 is 1, the figure for 2000 = $\frac{162}{150}$

$\therefore$

If the figure for 1995 is 100, the figure for 2000 = $\frac{162}{150}\times100=108$

Therefore, 108 is the index number for 2000 with 1995 as base.

Similarly the index number for 2005 with 1995 as base will be $\frac{180\times100}{100}\times120$

Index number for 2990 will be $\frac{126\times100}{150}$ = 84 and so on.

The new series with its base shifted to 1995 is thus –

Year	1984	1985	1990	1995	2000	1995
Index no.	66.7	74.0	84.00	100.0	108.0	120.0

Illustration – 7

(SEE ILLUSTRATION 3) PAGE NO. 25 MERITS AND DEMERITS OF THE CHAIN BASE METHOD.

The advantages of the chain base method are :

(i) Under this method the index for the current year is related to the year immediately preceding it. This enables us to know the extent of the changes that has come in the current year as compared to the previous year. This is certainly more useful to business than a fixed index which is related to a year of the distance past.

(ii) Under this method it is possible to introduce new items or drop out old ones without having to recalculate the whole services. This is because of the fact that the index of any one year is related only to the year just preceding it and the changes occurring in neghbouring period, periods are never so great as to impair comparability. Thus, if the list of commodities needs frequent change the chain base method is preferable to the method fixed base.

(iii) Weights can be adjusted as frequently as possible. This flexibility is of great significance in may types of index numbers.

(iv) Index number calculated by the chain base method are free to a greater extents from seasonal variations that those obtained by the other method.

This method, however, involves lengthy calculations and if an error is committed it tends to be perpetuated in changing process.

BASE SHIFTING, SPLICING AND DEFLATING

Base shifting : many times it becomes necessary to shift the base of a series of index numbers from one period to another. It is needed either because the base has become too old and, consequently practically useless, of comparison is to be made with a series with a different base and its value in 1996 and 2005 be 150 and 300 respectively. Let another series of indices, say, of production, have a base 1952 and its value in 1960 be 200. From these figures one may conclude that as the change in cost of living is of 150 points (300-150), in the latter series is greater. But this conclusion is not correct as the two series have different base periods. To have valid comparisons it will be necessary to correct the cost of living series into a new series with 1996 as the base year, i.e. the base of this series should be shifted to 1996.

The best method of base shifting which will give correct results is to reconstruct the series with the new base. This means that for each year relatives corresponding to each commodity included in that index number are recomputed one the new base and hen averaged out. This new average will give appropriate index number. But this process is very lengthy and may not be

Splicing

The problem of combining two or more overlapping series of index numbers into one continuous series is called splicing.

It is usually found that in course of time some articles included in an index number series may go out of the market. New ones may come int. Their relative importance may also change when these changes becomes sufficiently important their inclusion in the index number becomes necessary. As consequence the old series of index number is discontinued and a new series is constructed with the year of discontinuation of the first as base. This means that we now have two series of index numbers for the same phenomenon one of them coming up to the year from which the other begins. Thus the index numbers contained in two series are not directly comparable for the simple reason that they are prepared on different basis. In order to facilitate the comparison these two series are put together in one continuous series i.e. the two series are spliced together The method for doing this is :

Multiply the various indices of the new series by the index number of the last year in the old series and divided he result so obtained by 100 i.e.

$$\text{Spliced index no} = \frac{\text{Index no of current year x Old index of base year}}{100}$$

Illustration – 9

Year	1989	1990	1995	2000	2005
Series A	1000	120	150	....	....
Series B	....	....	100	112	136

Here series A was discontinued in 1995 and that year a new series was started. It is desired to splice the two series.

Solution

Let us make calculations for 2000

When index number for 1995 is 100 index number 2000 = 112 (given by series B).

When index number for 1995 is 150, index number 2000 = $\frac{112 \times 150}{100} \times 18$

168 becomes the index number for 1950 in the spliced series.

The two series spliced in this way gives the result as follows

Year	1989	1990	1995	2000	2005
Sliced series	100	120.0	150.0	160.0	204.0

Instead of carrying series A forward, series B may be brought back wards.

In this case every figure of series A is divided by the index number of the year in which change takes place and the result so obtained is multiplied by 100. In the present Illustration the two series spliced in the way give the result as follows :

Year	1989	1990	1995	2000	2005
Spliced series	66.7	80.0	100.0	112.0	136.0

Illustration - 10

The index A given was started in 1991 and continued up to 2001 in which year another index B was started. Splice the index B to index A so that a continuous series of index numbers from 11991 up-to-data may be available.

Year	Index A	Index B	Year	Index A	Index B
1991	100		2002		120
1992	110		2003		110
1993	112		2004		130
2000	138		2005		150
2001	150	100	2006		

Solution

Index B Spliced to Index A.

Year	Index A	Index B	Index B Spliced to Index A 1991 as base.
1991	100		
1992	110		
1993	112		
....			
....			
2000	138		
2001	100	$\frac{150}{100}\times 100 = 150$	
2002	120	$\frac{150}{100}\times 120 = 180$	
2003	140	$\frac{150}{100}\times 140 = 210$	
2004	130	$\frac{150}{100}\times 130 = 195$	
2005	150	$\frac{150}{100}\times 150 = 225$	

The spliced index now refers to 1991 as base and we can make a continuous comparison of index numbers from 100 on wards.

In the above case it is also possible to splice the new index in such a manner that a comparison could be made with 2001 as base. Thus would be done by multiplying the old index by the ratio $\frac{100}{150}$. Thus the spliced index for 1991 would be $\frac{100}{150} \times 100 = 66.7 = 66.7$, for 1992

$\frac{100}{150} \times 100 = 73.3$

For 1993, $\frac{100}{150} \times 002 = 74.7 = 74.7$ etc. This process appears to be more useful because a recent year can be kept as a however, much would depend upon the object.

Illustration

Year	Index X series	Index Y series	Year	Index X series	Index Y series
1966	10		2001	116	
1967	105		2002	118	
1997	300	100	2003	120	
1998	108		2004	108	108
1999	110		2005	110	106
2000	115			115	

Here index series X was discontinued in 1998 and that year a new series Y was started. It is desired to splice the two series.

Solution

Splicing of index numbers

Year	Index Series	Index X	Splicing technique series Y	Y Spliced to X
1966	100			
1967	105			
……	……	(Started)		
……	……			
1997	300	100	100 x 300/100	300
1998	Stopped	108	108 x 300/ 100	324
1999		110	110 x 300/ 100	330
2000		115	115 x 300 / 100	345
2001		116	116 x 300 / 100	348
2002		118	118 x 300 / 100	354
2003		120	120 x 300 / 100	360
2004		108	108 x 300 / 100	324
2005		106	106 x 300 / 100	318

For the other years are also found out in the same process. In the above example, the index numbers of real wages has fallen from 100 in 1997 to 56.6 in 2005. In other words in spite the fact that the money wage has increased from Rs 200 in 1997 to Rs 375 in 2005, the worker is not better off.

1. What is consumer prise index ?
Wow does it differ from wholesale price index?

2. Distinguish between fixed and chain base indices. Give a suitable illustration to show the difference. Point out the merits and demerits of the two methods.

3. Write short notes on :
   - a. Base shifting of index number
   - b. Splicing of index numbers.
   - c. Deflating of index numbers.

4. Calculate the price index numbers by
   - a. Laspeyre's method
   - b. Pasche's method
   - c. Fisher's ideal method
   - d. Bowly's method

Commodity	Price	2004 Quantity	2005 Price	Quantity
A	20	8	40	6
B	50	10	60	5
C	40	15	50	10
D	20	20	20	15

Ans. (a) 124.69, (b) 125.23, (c) 124.97 (d) 124.96

PROBLEMS RELATING TO INDEX NUMBER AND SOLUTION

PROBLEM – 1

Construct the cost of living index number for the year 2005 based on 2000 from the following table by assigning the given weights

Group	Group index no. per 2005 with 2000	Weights
Food 152	48	
Food and Lighting	110	5
Clothing	130	15
House Rent	100	12
Misc.	80	20

If index of cost of living was 100, wages $\frac{150}{130} \times 100 = 667$ =Rs. 115.4 approx.

Similarly, the deflated income for 2005 $\frac{215 \times 100}{208} = 103.4$

The deflated incomes for various years are thus :

Year	1993	1994	1999	2001	2005
Deflated Income (Rs)	120	1190	115.4	125.3	103.4

Illustration

The annual wages of a worker in rupees along with cost of living index numbers are given below. Deflate the wages series and prepare index number of real wages.

Year	1997	2000	2001	2002	2003	2004	2005
Wages (Rs)	200	240	350	360	360	370	375
Cost of living Index no.	100	160	280	290	300	300	330

Solution

Deflation of wage-series and construction of Real wages index number.

Year	Wage Rs.	Cost of living index	Real wages (col 2 col.3)	Rs. Real wage 1997=100 x 100
1997	200	100	200	100
2000	240	160	150	75
2001	350	280	125	62.5
2002	360	290	124.1	62
2003	360	300	120	60
2004	370	320	115.6	57.8
2005	375	330	113.6	56.8

The deflated income for 1997

Or real Wage $\frac{\text{Money wage}}{\text{cost of living index}} \times 100$

$\frac{200}{100} \times 100 = 200$

Similarly Real Wage for 2000

STATISTICAL DEFLATION

Economists and businessmen often find statistical series presented in terms of value (money). Imports and exports, construction contracts, industrial production etc. are expressed in terms of rupees. Fluctuations in such series are caused by (i) changes in physical volume and (ii) changes in the price – level.

They are however interested more in studying changed in physical quantities. Consequently fluctuations due to changes in price level have to be eliminated from such series. This is known as "deflation". In other words deflation means making allowance for the effect of changing price levels. A rise in price level means a reduction in the purchasing power of money. His can be explained well by an example.

Suppose the price of wheat rises from Rs.50 per quintal in 1995 to Rs.100 per quintal in 2005. it means that in 2005 one can try only 50 kgs. Of wheat for Rs.50 which he was spending on wheat in 2005 or in other words, the value of rupee is only 50 paise in 2005 as compared to 1995. Thus the value (or purchasing power) of a rupee is simply the reciprocal of an appropriate price index written as a proportion. If prices increase by 60% the price index is 1.60 and what a rupee will buy is only 1/1.60 or on 5/8 of what it was or approximately 62 paise. Similarly, if prices increase by 25% the price index is 1.25 (125 per cent) and the purchasing power of the rupee is 1/1.25=0.80=paise.

The method of deflating a series of figures to the base year level of suitable number series is to divide the figure corresponding to various time periods of the given series by the corresponding figure of the index number series and multiply the result to obtain by 100.

Illustration

Following table gives the monthly wages of a worker together with the cost of living index numbers. Deflate the monthly wages by the cost of living index number.

Year	1993	1994	1999	2001	2005
Wage per month (Rs)	120	125	150	178	215
Cost of living Index no.	100	105	130	142	208

Solution

Let us calculate the figure for 1999. In this year the wages are Rs.150 P.M and cost of living index number is 130. To get the deflated income one has to proceed as follows :

When index of cost of living is 130 wages = Rs.150

Solution :

When the weights are given and it is not mentioned that weighted geometric mean is to be used or weighted arithmetic mean Moreover, in cost of living index numbers, weighted arithmetic mean is used in actual construction. So tho indeed number of cost of living i.e. weighted arithmetic of given figures is

$$\text{Index No} = \frac{252\times 48+110\times 5+130\times 12+100+8020}{48+5+15+12+20} = \frac{12596}{100} = 125.96$$

PROBLEM – 2

Construct the cost of living index number from the table given below :

Group	Index for 2005	Expenditure
1. Food	550	46%
2. Clothing	215	10%
3. Food & Lighting	220	7%
4. House rent	150	12%
5. Miscellaneous	275	25%

Solution :

Construction of cost of living index numbers

Group	Index no. I	Expenditure V	IV
1. Food	550	46	25.300
2. Clothing	215	10	2.150
3. Food & Lighting	220	7	1.510
4. House rent	150	12	1.800
5. Miscellaneous	278	25	6.875
		ΣV=100	Σ IV = 37.665

$$\text{Cost of living index} \quad = \frac{\sum IV}{\sum V} = \frac{37.665}{100} = 376.65$$

PROBLEM - 3

The price quotations of four different commodities for 1991 and 2005 are given below. Calculate the index number of 2005 with 1991 as base by using (i) simple average of price relatives and (ii) weighted average of price relatives.

Commodity	Weights	Price in Rupees	
		2005	1991
A	5	4.5	2.0
B	7	2.2	2.5
C	6	4.5	3.0
D	2	1.8	1.0

Solution :

Calculation of index number

Commodity	Price in Rs.		Price Relative Weights		
	1991	2005	2005	R.W	R.W
A	2.0	4.5	225	5	1125
B	2.5	2.2	128	7	896
C	3.0	4.5	150	6	900
D	1.0	1.8	180	2	360
Total			683	20	3281

Index no. of 2005 using

(i) Simple Average of elatives $= \frac{\sum R}{n} = \frac{683}{4} = 170.75$

(ii) Weighted Average of Relatives $= \frac{\sum RW}{\sum W} = \frac{3281}{20} = 164.05$

Note:

When weights are given, commodities weighted arithmetic mean is used. As against weighted average; simple average ordinarily indicates simple arithmetic mean and not geometric mean. So here, under (i) Simple arithmetic mean is calculated even though it is not a good method of averaging index number using in weights and not geometric mean. So here, under (i) Simple arithmetic mean is calculated even though it is not a good method of averaging index number using in weights :

PROBLEM - 4

An enquiry into the budgets of middle class families in a city in India gave the following information :

Expenses on	Food	Rent	Clothing	Food	Miscellaneous
	35%	15%	20%	10%	20%
Price in Rs.	150	30	15	25	40
Price in Rs.	154	30	65	23	45

What changes in the cost of living figures of 1969 as compared with 1968 are seen ?

Solution :

Construction of cost of living index number for 1969 with 1968 as the base.

Items of expenditure	Price in Rs. 1968 1969		Price Relatives (P)	W	P.W.
	P_0	P_1	P_1/P_0 Taking 1968=100		
Food	150	145	96.67	35	3383.3
Rent	30	30	100.00	15	1500.0
Clothing	75	65	86.7	20	1734.0
Fuel	25	23	92.0	10	920.0
Misc.	40	45	112.5	20	2250.0
				ΣW = 100	Σ PW = 9787.3

$$\text{Cost of Living Index} = \frac{\sum PW}{\sum W} = \frac{9787.3}{100} = 97.87$$

Thus a fall of 2.13% (100 – 97.87) has taken place in the cost of living of middle class families in the given city of India in 1969 as compared with 1968.

PROBLEM - 5

Calculate the price index for 1966 with 1962 as base i.e. 100 from the following data using unweighted arithmetic mean.

Article	Unit	Price 1962 (Rs)	Price 1966 (Rs.)
Wheat	Per Md.	10.0	25.0
Ghee	" Seer	4.0	6.0
Wood	" Md.	2.0	2.0
Sugar	" Seer	.075	0.50
Cloth	" Yard	2.50	1.0

Calculate also the index for 1962 with 1966 as 100 and comment upon the result.

Solution :

Year base-1914	First series base - 1929	Second series base - 1935		Third series index spliced with base 1935
1981	100			$\frac{100 \times 67}{2100} = 33$
1987	120			$\frac{120 \times 67}{200} = 40$
1996	200	100		$\frac{100 \times 100}{150} = 67$
2001		150	100	100
2005			120	120

PROBLEM - 9

Compute the cost of living index number for 1985 on the basis of 1946 level of prices from the following data, using from the following data, using the Family Budget Method.

Articles	Qty. consumed in 1946	Unit	Price 1946	Price 1955
Rice	5 mounds	Per mound	12	16
Wheat	1....	....	10	20
Barley	5.....	....	8	10
Gram	1.....	...	6	12
Athr	5.....	....	8	12
Other Pulses	2.....	....	6	8
Gur	2......	....	5	10
Salt	12.5 seers	Per seer	8	10
Oil	24....	.....	40	50
Ghee	4	.....	2.5	4
Cloth	40 yards	Per yard	.5	1
Fire wood	10 mounds	Per md.	1	1.6
Kerosene	1 tin	Per tin	4	7
House	1 unit	Per unit	24	70

PROBLEM - 7

Calculate the weighted cost of living index number form the following indices.

Group	Index No. (I)	Weights
Food	105	70
Fuel & Lighting	108	10
Clothing	112	12
Rent	106	5
Miscellaneous	102	3

Solution :

Cost of living index number

Group	Index No. (I)	Weights (W)	(IW)
Food	105	70	7,350
Fuel & Lighting	108	10	1,080
Clothing	112	12	1,344
Rent	106	5	530
Miscellaneous	102	3	306
		Σ W+100	Σ IW=10,610

Index Number $\frac{\sum IW}{\sum W} - \frac{10.610}{100} = 106.1$

PROBLEM - 8

In 1920 a statistical Bureau started an index of production based on 1914 with the following results :

Year	1914 (Base)	1920	1926
Index	100	120	200

In 1930 the Bureau reconstructed thee index on yet another with base 1929.

Year	1929 (base)	1939
Index	100	120

In 1939, the Bureau again reconstructed the index number on yet another plan with base 1935

Year	1935 (base)	1939
Index	100	120

It is required to splice these three series together so as to give a continuous series with base 1935. Draw up a working table in paralles columns and show the results for 1914, 1920, 1935 and 1939

Solution :

Article	Unit	Price in Rs		Price 1966 with 1962 as base	Relatives for 1962 with 1966 as base
		1962	1966		
Wheat	Per Md.	10.00	25.00	$\frac{25}{10}\times 100 = 250$	$\frac{10}{25}\times 100 = 40$
Ghee	Per Seer	4.00	6.00	$\frac{6}{4}\times 100 = 150$	$\frac{4}{6}\times 100 = 67$
Wood	Per Md.	200	2.00	$\frac{2}{2}\times 100 = 100$	$\frac{2}{2}\times 100 = 100$
Sugar	Per Seer	075	0.50	$\frac{50}{75}\times 100 = 67$	$\frac{75}{50}\times 100 = 150$
Cloth	Per Yard	2.50	1.10	$\frac{1}{2.5}\times 100 = 40$	$\frac{2.5}{1}\times 100 = 250$
			Total	607	607
			Average	121.4	121.4

It shows that both the cases give the same results, but it is not proper. This is due to the defect of arithmetic mean. The time reversal test is not satisfied here.

The following are the group index number and the group weights of an average working class family budget. Construct the cost of living index number by assigning the given weights.

	Index No.	Weights
Group	352	48
Food	220	10
Fuel & Lighting	200	10
Clothing	150	10
Rent	180	12
Miscellaneous		

Solution :

Cost of living index numbers

Group	Index No. (I)	Weights (W)	Weighted Relatives (IW)
Food	352	48	16.896
Fuel & Lighting	220	10	2.200
Clothing	200	10	2.00
Rent	150	10	1.500
Miscellaneous	180	12	2.160
Total		90	24.756

Cost of living index number $\frac{\sum IW}{\sum W} = \frac{24.756}{90} = 270$ (approx)

Commodities & Unit	Price in Rs. 1976	Price in Rs. 1977
Butter (kg)	14.00	15.00
Cheese (kg)	10.00	12.00
Milk (Ltre)	1.50	1.80
Bread(1)	0.70	0.75
Eggs (dozen)	3.50	4.00
Ghee (tin)	100.00	110.00

Solution :

(a) Price index based on simple average of price relative

Commodities & Unit	Price in Rs. 1976	Price in Rs. 1977	$\frac{P_1}{P_0} \times 100$
Butter (kg)	14.00	15.00	107.14
Cheese (kg)	10.00	12.00	120.00
Milk (Ltre)	1.50	1.80	120.00
Bread(1)	0.70	0.75	107.14
Eggs (dozen)	3.50	4.00	114.29
Ghee (tin)	100.00	110.00	110.00

N = 6 $\frac{\sum P_1}{\sum P_0} \times 100 = 678.57$

Price Index $\frac{\sum P_1}{\frac{P_0}{N}} \times 100 = \frac{678.57}{6} = 113.09$

(b) Price index based on geometric mean of price relatives.

Commodities & Unit	Price in 1976 (P_0)	Price 1977 (P_1)	Price Relatives (P)	Log P
Butter (kg)	14.00	15.00	107.14	2.0300
Cheese (kg)	10.00	12.00	120.00	0.0792
Milk (Ltre)	1.50	1.80	120.00	2.0792
Bread(1)	0.70	0.75	107.14	2.0300
Eggs (dozen)	3.50	4.00	114.29	2.0580
Ghee (tin)	100.00	110.00	110.00	2.0414

N = 6 Σlog p = 12.3178

$P_{01} = AL \frac{\left[\sum \log p\right]}{N} = AL \frac{\left[\sum 12.3178\right]}{6}$ = AL 2.053 - 113

Solution :

Articles	Qty.	Unit	Price 1946	In Rs. 1955	Price relative for 1955 $\frac{P_1}{P_0} \cdot 100$	Value of goods consumed in the base year	IV
Rice	5 mounds	Per yards	12	16	133.3	60	8000
Wheat	1....	Per yards	10	20	200	10	2000
Barley	5.....	" "	8	10	125.40		5000
Gram	1.....	" "	6	12	200	6	1200
Athr	0.5.....	" "	8	12	150	4	600
Other Pulses	2.....	" "	6	8	133.3	12	1600
Gur	2.......	Per seer	5	10	200	10	2000
Salt	12.5 seers	Per seer	8	10	125	2.5	312.5
Oil	24....	Per seer	40	50	125	24	3000
Ghee	4	Pr seer	2.5	4	160	10	1600
Cloth	40 yards	Per yard	.5	1	200	20	4000
Fire wood	10 mounds	Per md.	1	1.6	160	10	1600
Kerosene	1 tin	Per tin	4	7	175	4	700
House	1 unit	Per unit	24	70	125	24	3000
						236.5	34612.5

Cost of living by family Budget Method = $\frac{\sum IV}{\sum V} = \frac{34612.5}{236.5}$ = 146.4 (approx)

Solution :

Year base-1914	First series base - 1929	Second series base - 1935		Third series index spliced with base 1935
1981	100			$\frac{100 \times 67}{200} = 33$
1987	120			$\frac{120 \times 67}{200} = 40$
1996	200	100		$\frac{100 \times 100}{150} = 67$
2001		150	100	100
2005			120	120

PROBLEM - 9

Compute the cost of living index number for 1985 on the basis of 1946 level of prices from the following data, using from the following data, using the Family Budget Method.

Articles	Qty. consumed in 1946	Unit	Price 1946	Price 1955
Rice	5 mounds	Per mound	12	16
Wheat	1....	....	10	20
Barley	5.....	....	8	10
Gram	1.....	...	6	12
Athr	5.....	....	8	12
Other Pulses	2.....	....	6	8
Gur	2......	....	5	10
Salt	12.5 seers	Per seer	8	10
Oil	24....	.....	40	50
Ghee	4	.....	2.5	4
Cloth	40 yards	Per yard	.5	1
Fire wood	10 mounds	Per md.	1	1.6
Kerosene	1 tin	Per tin	4	7
House	1 unit	Per unit	24	70

Solution :

Groups	Index for Jan. 1960 (1967)=100) (1)	Weight (W)	(WI)
Food	125	384	43500
Rent	101	88	8888
Clothing	122	97	11834
Fuel & Light	118	65	7670
Household Expenses	114	71	8094
Miscellaneous	113	35	3955
Services	A	79	97A
Entertainment etc.	104	217	22568
Total		1000	106509 + 79A

The index $= \frac{106509 + 79A}{100} = 115.278$

Or 1106.509 + 79A = 1.15,278

Or 79A = 1,15,278 = 106,509 = 8769

Or A $= \frac{8769}{79} = 111$

Hence the increase in service group III.

Solution :

Index number of Business Activity

Item	Weightage (W)	Index (1)	(Wi)	Log of index (L)	Weight x log (L xW)
Industrial production	36	250	9000	2.3979	86.3244
Mineral	7	135	945	2.3.3	14.9121
Production internal	20	200	4800	2.3010	55.2240
Trade financial	20	135	2700	2.1303	42.6060
Activity exports & imports	7	325	2275	2.5119	17.5833
Shipping activity	6	300	1800	2.4771	14.8626
Total	ΣW=100	ΣWI = 21520		Σ L x W = 231.5124	

Index number (Simple mean)

$$= \frac{WI}{W} = \frac{21520}{100} = 215.20$$

Index number (Geometric mean)

= Anti log $\frac{Weight \times \log}{Weight}$

= Anti log = $\frac{231.5124}{100}$

= Anti log 2.3151

=20.5

The second answer is better

PROBLEM – 14

The different groups of family expenditure are given weights as follows:

Food	348	Household expenses	71
Rent	88	Miscellaneous	35
Clothing	97	Services	79
Fuel & Light	65	Entertainment etc	217

The corresponding increase in price for January 1960 based on 1957 was 25, 1, 22, 18, 14, 13, 7 and 4.

If the average percentage increase in price for the whole group os 15, 278 find the increase in service group.

PROPBLEM – 12

Given the data

Commodities		
	A	B
	1	1
P_0	10	5
Q_0	2	X
P_1	5	2
Q_1		

Where p q respectively stand for price and quantity, and subscripts (O & 1) stand for the period. Find x if the ratio between Laspayres L and Paasche's index number is

L: P = 28 : 27

Solution :

Calculate Laspeyres and Paasche's indices and equate then to the given ratio in order determine the value of x.

Commodities	P_0	Q_0	p_q	q_1	P_1q_0	P_0q_0	P_1Q_1	P_0q_1
A	1	10	2	5	20	10	10	5
B	1	5	x	2	5x	5	2x	2

Note: In order to work out the ratios 100 has been omitted from the formula

PROBLEM – 13

Construct the index number of business activity in India from the following data.

Item	Weightage	Index
Industrial Production	36	250
Mineral Production	7	135
Internal Trade	24	200
Financial Activity	20	135
Exports & imports	7	325
Shipping activity	6	300