# ECONOMIC STATISTICS

## Syllabus for B.A.Economics, Non-Semester (DD & CE)

### UNIT - I

Definitions of Statistics –Functions - Importance - Sources of data – Methods of collecting primary and secondary data –Census and Sampling: Methods of sampling - Classification and Tabulation - Presentation of statistical data: Diagrams and Graphs

### UNIT – II

Measures of Central Tendency: Properties- Mean, Median, Mode, Geometric Mean and Harmonic Mean- Merits and Demerits

### UNIT – III

Measures of Dispersion: Meaning - Range, Quartile Deviation, Mean Deviation, Standard Deviation, Variance, Coefficient of variation, Lorenz Curve – Merits and Demerits

### UNIT – IV

Correlation and Regression- Meaning and Types of Correlation, Measurement: Karl Pearson Co-efficient of Correlation and Spearman's Rank correlation - Regression - Differences between Correlation and Regression - Regression equations

### UNIT – V

Analysis of Time Series and Index Number – Meaning of Time Series- Components of Time Series - Meaning of Index Number- Problems in the Construction of Index Numbers – Methods of construction of Index Number- Laspeyre's Method- Paasche's method and Fisher's Index number

### Reference Books:

S .P.Gupta, Elementary Statistical Methods –Sultan Chand & Sons, New Delhi, 2010

S. P.Gupta, Statistical Methods –Sultan Chand New Delhi, 2001.

K. Pazhani, Statistics, J.P.Publishers, Nagercoil, 2004

R.S.N. Pillai&Bagavathi, Statistics –S. Chand, New Delhi, 2006

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data. Statistics is a mathematical body of science that pertains to the collection, analysis, interpretation or explanation, and presentation of data, or as a branch of mathematics. Some consider statistics to be a distinct mathematical science rather than a branch of mathematics. While many scientific investigations make use of data, statistics is concerned with the use of data in the context of uncertainty and decision making in the face of uncertainty

In applying statistics to a problem, it is common practice to start with a population or process to be studied. Populations can be diverse topics such as "all persons living in a country" or "every atom composing a crystal". Ideally, statisticians compile data about the entire population (an operation called census). This may be organized by governmental statistical institutes. Descriptive statistics can be used to summarize the population data. Numerical descriptors include mean and standard deviation for continuous data types (like income), while frequency and percentage are more useful in terms of describing categorical data (like race).

Statistics is the science of learning from data, and of measuring, controlling, and communicating uncertainty; and it thereby provides the navigation essential for controlling the course of scientific and societal advances (Davidian, M. and Louis, T. A., 10.1126/science.1218685). Statisticians apply statistical thinking and methods to a wide variety of scientific, social, and business endeavors in such areas as astronomy, biology, education, economics, engineering, genetics, marketing, medicine, psychology, public health, sports, among many. "The best thing about being a statistician is that you get to play in everyone else's backyard." (John Tukey, Bell Labs, Princeton University). Many economic, social, political,

and military decisions cannot be made without statistical techniques, such as the design of experiments to gain federal approval of a newly manufactured drug.

Many have seen statistics as a device to achieve the degree of precision in the concept and theories of social sciences. In nut shell, if we analyze the way in which statistics is looked at, we broadly find two categories, one refers statistics as a set of figures and the other connotes it as a set of techniques.

**(1) Origin and growth of Statistics**

Statistics, as a subject, is as old a discipline as the human life is. The origin of statistics is revealed by the word itself which is said to have been derived either from the Latin Word 'Status' or the Italian word 'statista' or the German word 'Statistik' which means political state. Statistics originated from two dissimilar areas; political states and games of chance. These two areas are also called as two peculiar disciplines – oe primarily descriptive and the other essentially analytical. The former is concerned with the collection of data and the latter is associated with the concept of chance and probability. Statistics was used as a by-product of administrative activity. Govt. maintains records of various types of numerical data on population, births, deaths, literates, illiterates, employment, unemployment, Income, Taxes, Imports, exports etc. Statistics was used as a technique to collect periodical data to ascertain the manpower and material strength for military and fiscal purposes. In ancient Egypt, census of population and wealth was conducted for the erection of pyramids. In India statistics was an effective system of collecting data about 2000 years ago in ancient works of Manusmriti, Shukraniti, etc., 'Kautily's, Arthshastra' describes facts and figures of Chandragupta

Mauraya's regime. 'Din-i-akbari and Tuzuk-i-Babri also describe the system of data collection. The histories of the other countries of the world also clearly reveal the use of statistics in the administrative activities. The Old Testament contains several accounts of census taking. Ancient Babylonia and Rome gathered detailed records of population and resources. Governments began to register the ownership of land in the Middle Ages. William, the conqueror, ordered in 1086 the writing of the Domesday Book, a record of the ownership, extent, and value of the lands of England. This was England's first statistical abstract.

The theoretical development of statistics has its origin in the mid-seventeenth century when many gamblers and mathematicians of France, Germany and England are credited for its development. Pascal and P. Fermat, the two great French mathematicians made innovative efforts to solve the famous 'Problem of point' which was posed by the famous French gambler Chevalier De Mere. Their contribution became the foundation stone of the Science of Probability. James Bernoulli (1654-1705) developed the 'Normal Curve'. The use of 'Statistics' was popularized by Sir John Sinclair in his work Statistical Account of Scotland (1791-1799). Modern Theory of Statistics was gradually developed during the 18th, 19th and 20th centuries mathematicians. Laplace (1749-1827) gave the principles of 'Least squares' and established the 'Normal Law of Errors'. The famous statisticians Sir Francis Galton (1822-1911), Karl Pearson (1857-1936) and W.S. Gosset contributed to the study of Regression Analysis, Correlation Analysis as well as Chi-square test of Goodness of Fit, and t-test respectively. R.A. Fisher,

who is called “Father of Statistics”, has developed statistics for use in genetics, biometry, agriculture, psychology and education. He also contributed to the Estimation Theory, Sampling Distribution, Analysis of Variance (ANOVA) and design of experiments. Thus Prof. Ronald A. Fisher is the real exponent in the development of the ‘Theory of Statistics”.

The tremendous growth in the use of statistics is primarily due to two reasons; increased demand of statistics and decreasing cost of statistics.

**(2) Need or Importance of Statistics in Statistical Analysis**

Statistical techniques are of universal applicability. These techniques are used in almost all fields of knowledge e.g. social science, medical science, physical science, natural science and so on. So far as the field of economics is concerned, this technique has become so important that even the understanding of Elementary Economics requires knowledge of statistics. In this regard Marshall opined, “Statistics are the straw out of which I like every other economist have to make bricks.” Similarly C.E. Engeberg remarked, “No economist would attempt to arrive at a conclusion concerning production or distribution of wealth without an exhaustive study of statistics”.

However, the significance of statistical analysis is on the basis of the following grounds:

- It provides a tool for scientific analysis.
- It provides solution for various business problems.

- It enables proper allocation of resources.
- It helps in minimizing waiting and servicing cost.
- It enables the management to decide when to buy and how to buy.
- It helps in choosing an optimum strategy.
- It renders great help in the optimum allocation of resources.
- It facilitates the process of decision making.
- Management can know the reactions of the integrated business system through quantitative analysis.

**Meaning of Statistics**

According to Tate, "You can compute statistics by statistics from statistics." Therefore the word statistics has three aspects; (a) Statistics (b) Statistical science (c) Statistical measurement.

According to Oxford Dictionary, "Statistics has two meanings, as in plural sense and singular sense. In plural sense, it means a systematic collection of numerical facts and in singular sense, it is the science of collecting, classifying and using statistics."

**Definitions of Statistics**

It is not an easy job to define statistics. The reason being that no two statisticians agree on the limits and scope of statistics. According to dictionary meaning of Statistics – it refers to the singular sense where as numerical data refers to the plural sense. It is by this reason to the singular sense where as numerical data refers to the plural sense. It is by this reason that some authors have defined the

word in the plural form i.e., numerical data whereas others have defined it in the singular form i.e. statistical method. Thus the term 'statistics' is used in two senses.

**Statistics in the Plural Sense or Statistical Data**

---

In plural sense, it means a collection of numerical facts. It is in this sense that the public usually think of statistics, say figures concerned with population or production of wheat in India in different years or number of man-hours lost in industry in a specific year. Secondary Statistics e.g., percentages, averages and coefficients derived from numerical facts, are also included in the term statistics in this sense.

For a layman, Statistics are only mass of figures. To understand what statistics is, we should define it in a way different authors have defined it and then to conclude its definition.

A.L. Bowley has given as series of definitions. Some lay emphasis on one aspect and others on other aspect. At one place Bowley says "Statistics, may be called the science of counting". This view is not perfect and correct. Statistics is not concerned with counting only. It deals more with estimates. At another place, he says that "Statistics may rightly be called the science of averages". But calling statistics as a science of counting or averages, confines the scope of statistics. Bowley himself realized this drawback and stated that statistics cannot be confined to any one sense.

Webster defined Statistics as “The classified facts respecting 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”.

This definition has limited the scope of statistics. It relates statistics only to those facts which are concerned with the condition of the people in a state. This concept does not suit the modern world. Furthermore, this definition is not exhaustive because it does not take into account all aspects of human activity. The definition given by Secrist is regarded as the most exhaustive.

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

## Characteristics or Features of Statistics in the Plural sense or Numerical Data

The basic feature of statistics as a quantitative or numerical data run as follows:

### Aggregate of Facts

Statistics does not refer to a single figure but it refers to a series of figures. A single weigh of 50 kg is not statistics but a series relating to the weight of a group of persons is called statistics. It means, all those figures which relate to the totality of facts are called statistics. Such figures should also be comparable.

**Affected by Multiplicity of Causes**

Statistics are not affected by one factor only, rather they are affected by a large number of factors. It is because statistics are commonly used in social sciences. It is not an easy job to study the effects of any one factor on a phenomenon and effects of different sets of factors separately. In nutshell, we can say that statistics are affected considerably by multiply causes e.g. prices are affected by conditions of demand, supply, money supply, imports, exports and various other factors.

**Numerically expressed**

Another characteristic of statistics is that qualitative expressions like young, old, good, bad etc. are not statistics. To all statistics a numerical value must be attached. For example, the statements like "There are 932 females per 1,000 statements must contain figures so that they are called numerical statements of facts. Furthermore, such numerical expressions are precise, meaningful and convenient form of communication.

**Enumerated or Estimated according to Reasonable Standards of Accuracy**

In case the numerical statement are precise and accurate, then these can be enumerated. But in case the number of observations is very large, in that case the figures are estimated. It is obvious that the estimated figures cannot be absolutely accurate and precise. The accuracy, of course, depends on the purpose for which statistics are collected. There cannot be uniform standard of accuracy for all types of enquiries. Thus enumeration refers to exact count as there are ten students of statistics. It is 100% accurate statement. On the other hand, estimation refers to

round about figure e.g. we say that two lakhs people participated in the Rally. There can be a few hundreds more or less. Thus statistical results are true only on average.

**Collected in a Systematic Manner**

For accuracy or reliability of data, the figures should be collected in a systematic manner. If the figures are collected in a haphazard manner, the reliability of such data will decrease. Thus for reasonable standard of accuracy, the data should be collected in a systematic manner, otherwise the results would be erroneous.

**Collected for a Pre-determined Purpose**

The usefulness of the data collected would be negligible if the data are not collected with some pre-determined purpose. The figures are collected with some objective in mind. The efforts made without any set objective would render the collected figures useless. Thus the purpose of collecting data must be decided well in advance. Besides, the objective should be concrete and specific. For example, if we want to collect data on prices, then we must be clear whether we have to collect whole-sale or retail prices. If we want data on retail prices, then we have to see the number of goods required to serve the objective.

**Placed in Relation to each other**

The collection of data is generally done with the motive to compare. If the figures collected are not comparable in that case, they lose a large part of their significance. It means the figures collected should be homogeneous for comparison

and not heterogeneous. In case of heterogeneity, the figures cannot be placed in relation to each other.

**Singular Definitions or Statistical Methods**

---

In a singular sense, statistics implies statistics methods. Thus, it is a body or technique of methods relating to the collection, classification, presentation, analysis and interpretation of information. In a sense, statistics can be defined as:

According to Boddington: "Statistics is the science of estimates and probabilities."

According to Bowley, "Statistics may rightly be called the science of averages."

According to W.I. King, "Science of statistics is the method of judging collective natural or social phenomena from the results obtained by the analysis of an enumeration or collection of estimates."

According to P.H. Karmel, "The subject statistics is concerned with the collection, presentation, description and analysis of data which are measurable in numerical terms."

According to Seligman, "Statistics is the science which deals with the method of collecting presenting, comparing and interpreting numerical data-collected to throw some light on any sphere of inquiry."

## Characteristics or Stages in Statistical Investigations

### Collection

It is the first step in a statistical inquiry. The collection part is the backbone of the inquiry. If the collection of data is not in proper form, in that case the conclusions drawn can never be reliable. The source of data may be primary i.e., collected directly, or the data may be secondary i.e., available from existing published sources. The first hand collection of data is one of the most difficult and important tasks faced by the investigator.

### Organisation

The data collected from published sources are generally in organized form. But the figures which are collected from a survey, need organisation. The most important point in organizing a group of data is editing. This is done to correct omission, inconsistencies and wrong calculations in the survey. The classification is done to arrange the data according to some common characteristics possessed by the items. The last step in organisation is tabulation. The object of tabulation is to arrange the data in columns and rows for complete clarity as far as presentation of data is concerned. Thus organisation can be classified into three stages, (i) Editing (ii) Classification (iii) Tabulation.

### Presentation

After collection and organisation, the data should be presented. If the data are presented in an orderly manner, the statistical analysis gets facilitated. As far as the presentation of data is concerned. The classified data are to be presented in such a way that it becomes easily understandable.

**Analysis**

Once the data are collected, organized and presented, the next step is that of analysis. The main objective of analysis is to prepare data in such a fashion so as to arrive at certain definite conclusions. The methods most commonly used are measures of Central Tendency and are called measures of the first order. Measures of Dispersion are called measures of the second order. Skewness, Correlation, Regression, Interpolation etc. are called measures of the third order. The analysis of facts based on observation is termed as (a) Scientific analysis, (b) Numerical analysis, (c) empirical analysis.

**Interpretation**

The last stage of statistical investigation is to derive the results and give comments on the inquiry in question. Interpretation means to draw conclusions from the data collected and analyzed. The interpretation of data is not an easy job and requires a high degree of skill and experience. If the analyzed data are not properly interpreted, the whole object of the inquiry may be erroneous. It is only correct interpretation which may lead to reliable conclusions.

Thus it is clear that statistics is a science of taking decisions on the basis of numerical data properly collected, organized, presented, analyzed and interpreted.

**Functions / Applications of Statistics**

---

The various applications or functions performed by statistics in modern times are discussed as under:

**Simplification of Complex Facts:**

The foremost purpose of statistics is to simplify huge collection of numerical data. It is beyond the reach of human mind to remember and recollect the huge facts and figures. Statistical method makes it possible to understand the whole in the short span of time and in a better way.

**Comparison**

Comparison of data is yet another function of statistics. After simplifying the data, it can be correlated or compared by certain mathematical quantities like averages, ratios, co-efficient etc. In this regard Boddington opined that the object of statistics is to enable comparison to be made between past and present results with a view to ascertain the reasons for changes which have taken place and the effect of such changes in the future.

**Relationship between Facts**

Statistical methods are used to investigate the cause and effect relationship between two or more facts. The relationship between demand and supply, money-supply and price level can be best understood with the help of statistical methods.

**Formulation and Testing of Hypothesis**

The most theoretical function of statistics is to test the various types of hypothesis an discover a new theory. For instance, by using appropriate statistical tools we can test the hypothesis whether a particular coin is fair or not, whether Indian consumers are brand loyal etc.

**Measurement of Effects**

Statistical methods act as a guide to measure the effect of a policy. For example, the effect of a change in bank rate or a change in incomes tax etc. can best be judged by the statistical methods.

**Forecasting**

Statistical methods are of great use to predict the future course of action of the phenomenon. It is only on the basis of statistical techniques that the planners in India prepare future estimates for production, consumption, investment etc.

**Enlarges Individual Knowledge**

Statistical methods sharpen the faculty of rational thinking and reasoning of an individual. It is a master-key that solves the problems of mankind in every sphere of life. Thus, Whipple has rightly opined that statistics enables one to enlarge his horizon.

**Realization of Magnitude**

Statistics facilitates the realization of the magnitude of a problem. For instance, one may say that inflation in India has been increasing very rapidly. From this statement, one is unable to understand the gravity of the problem. But, if one says that inflation is increasing by 5.5% per annum, everyone will properly realize the gravity of the problem.

**To Indicate Trend Behavior**

Statistics helps to indicate trend behavior of certain fields of enquiry. The statistical techniques like Analysis of Time Series, Extrapolation etc. are highly used to know the trend behavior of the enquiry in question.

**Classification of Data:**

Classification refers to a process of splitting up the data into certain parts which helps in the matters of comparison and interpretation of the various features of the data. This is done by the various improved techniques of statistics.

**To measure Uncertainty**

In most of the social fields, comprising of business, commerce, economics, it becomes necessary to take decisions in the face of uncertainty and study the chance of occurrence of certain events and their effect on the policies adopted.

**Scope of Statistics**

The scope of statistics is studied under the following heads:

- Nature of statistics or Statistics as Science or an Art
- Subject matter of Statistics
- Relation of Statistics with other Sciences
- Limitations of Statistics.

**Nature of Statistics or Statistics as science or an Art**

Under nature, we study whether statistics is a science or an art. According to Tippett, “Statistics is both a science and an art”. As a science, it studies the statistics in a systematic manner. As an art, it uses statistics to solve the problems or real life.

It is a controversial subject that whether statistics is a science or an art. Science refers to a systematized body of knowledge. In general, it deals with the cause and effect relationship. It helps in drawing generalization in the form of principles or laws. The use of statistics in every science is universal. But statistics

cannot be known as science. Statistics refer to certain methods which help in arriving at some laws. Design of scientific experiments and the evaluation of their results makes use of principles and practices growing out of the science of statistics.

According to Dr. Bowley, "Statistics is the science of measurement of the social organism regarded as a whole in all its manifestations." This definition is defective. According to this definition, the scope of the science will be restricted to man and his activities.

Again according to Dr. Bowley, "Statistics is the science of counting." This is also not correct. Statistics is not only concerned with counting but also with tabulating, summarizing, drawing graphs etc. Again the word counting gives an idea of exactness. But this is not so in statistics.

Boddingtons defines Statistics as a "Science of estimates and probabilities." This definition is too vague and is concerned with certain methods by which conclusion can be derived in this science but the scope of statistics is not merely confined to these things.

According to Dr. Bowley, "Statistics may rightly be called the science of averages." This is also not correct since it takes into account only one step in the process of statistical operations.

But a noteworthy point is that statistics is not an exact science like Physics, Chemistry etc. It is because statistical phenomena are generally affected by multiplicity of causes which cannot be measured accurately. It means statistics is a science in a limited sense. It is a specialized branch of knowledge. Wallis and

Roberts in their book- "Statistics -A New Approach" have stated that "Statistics is not a body of substantive knowledge but a body of methods for obtaining knowledge."

It is a known fact that if science is a systematic knowledge then art is an action. From this point in view, statistics is also an art. In statistics, we apply various methods to obtain facts, derive conclusions and finally suggest, policy measures.

We can conclude with Tippet's Words, "It is both a science and an art. It is a science in that its methods are basically systematic and have general application, and an art in that their successful application depends to a considerable degree on the skill and special experience of the statistician and his knowledge of the field of application, e.g., economics."

**Limitations of Statistics**

Although, statistics is a very useful science yet it suffers from certain limitation. According to Newsholme, "It must be regarded as an instrument of research of great value but having several limitations which are not possible to overcome and as such they need a careful attention."

**Statistics does not deal with individuals**

Statistics deals only with the aggregates rather than individual items. An individual item like height of a student in a class is 5′ 6′-is not called statistics. In statistical methods, we deal with aggregates and not with a single figure. When we say that average height of a class is 5′ 8′, this individual figure refers to the aggregate of individuals. Statistics cannot be of much help for making a study of

the changes which may have taken place in individual cases. Thus it is clear that statistics is concerned with aggregates and not with individual items.

**Qualitative Aspect Ignored**

The statistical method cannot study the nature of phenomena which cannot be expressed in quantitative form. The phenomena which cannot be expressed quantitatively, cannot be a part of the study of statistics. These characteristics include health, intelligence etc. There is no doubt that the data which cannot be quantitatively expressed, needs conversion of qualitative data into quantitative data. Thus experiments are being conducted to measure the reactions of human mind statistically. Presently there is hardly any field where statistics does not apply. One of the branch of statistics, the "Theory of Attributes" deals with qualitative data.

**Statistics deals with Average**

Statistical findings are true only on an average. According to W.I. King, "Statistics largely deals with averages and these averages may be made up of individual items radically different from each other." For instance, if we may say that the average production of wheat in the last ten years is 250 quintals, it does not mean that the production of every year is equal. Production of a particular year may be less or more than the other years. Thus, statistical information is true only on an average.

**Statistics can be Misused**

Statistics can be misused by ignorant or wrongly motivated persons. The data used by untrained people can lead to misleading results. The statistics can be

handled correctly only by those who have sufficient knowledge in statistics. It is correctly W.I. King points out, “One of the short-comings of statistics is that they do not bear on their face the label of their quality.”

**Method of Studying Problem**

There are many methods to solve the problem. Statistics is one of them. According to Croxten and Cowden, “It must not be assumed that the statistical method is the only method to use in research, neither should this method be considered the best attack for every problem.”

**Results true only on an Average**

We know that statistics is not as accurate science as other science. Similarly the statistical methods are not very precise and correct. In the same fashion, the laws of statistics are not universal like the laws of physics, Chemistry or Astronomy. The statistical laws are true only on average. Statistics are concerned with those phenomena which are affected by multiplicity of causes. In this way, statistics is less exact science as compared to other natural sciences.

**Statistics is only a means**

Statistics is only one of the methods of studying a problem. There are other methods of studying a problem like culture, religion, philosophy etc. Statistics is only a means and not an end. It analysis the facts and throws light on the real situation.

**Statistical Survey**

A survey or inquiry means search for knowledge. Statistical survey or statistical inquiry means search for knowledge with the help of statistical methods. Statistical survey is a technical job which requires specialized knowledge and skill. According to Giffin, "Statistical enquiries have always required considerable skill on the part of the statistician, rooted in a broad knowledge of the subject-matter area and combined with considerable ingenuity in overcoming practical difficulties."

**Planning the Statistical Survey**

A proper planning is essential before a statistical survey is conducted. Planning must precede execution. Careful planning of statistical survey is essential to get the best results at the minimum cost and time. It is very essential to consider the following points while planning a statistical survey.

- Nature of information to be collected should be decided.
- Objective of the survey should be fully known.
- Scope of the survey should be determined.
- Source of data collection or types of data to be used i.e., primary data or secondary data should be decided.
- Type of enquiry i.e., census method or sampling method, should be decided before hand.
- Unit of data collection should be defined.
- Reasonable standard of accuracy should be fixed.
- Choice of frame should be made.

**Statistical Units**

Statistical unit is the basis of collection of statistics in a statistical inquiry. These are the units in terms of which data are collected, such as, for production of sugar 'tones' for weight of persons 'kilograms' etc.

**Statistical Data:**

A sequence of observation, made on a set of objects included in the sample drawn from population is known as statistical data

1. **Ungrouped Data**: Data which have been arranged in a systematic order are called raw data or ungrouped data.
2. **Grouped Data**: Data presented in the form of frequency distribution is called grouped data

**Sources of Data**

There are many ways of classifying data. A common classification is **based upon *who collected the data***.

**Primary data**: Data collected by the investigator himself/ herself for a specific purpose. ***Examples***: Data collected by a student for his/her thesis or research project.

**Secondary data**: Data collected by someone else for some other purpose (but being utilized by the investigator for another purpose). ***Examples***: Census data being used to analyze the impact of education on career choice and earning.

Broadly speaking, there are two sources of statistical data-**internal and external**. If data collected from outside, are called external data. External data can be collected either from the primary (original) so or from secondary sources. Such data are termed as primary and secondary data respective.

**Primary data:**

Primary data are first hand information. This information is collected directly from the source by means of field studies. Primary data are original and

are like raw materials. It is the most crude form of information. The investigator himself collects primary data or supervises its collection. It may be collected on a sample or census basis or from case studies.

**Some Advantages of using Primary data**:

1. The investigator collects data specific to the problem under study.
2. There is no doubt about the quality of the data collected (for the investigator).
3. If required, it may be possible to obtain additional data during the study period.

**Some Disadvantages of using Primary data (for reluctant/ uninterested investigators)**:

1. The investigator has to contend with all the hassles of data collection-

- deciding why, what, how, when to collect
- getting the data collected (personally or through others)
- getting funding and dealing with funding agencies
- Ethical considerations (consent, permissions, etc.)
- all desired data is obtained accurately, and in the format it is required in
- there is no fake/ cooked up data
- unnecessary/ useless data has not been included

**Secondary data:**

Secondary data are the Second hand information. The data which have already been collected and processed by some agency or persons and are not used for the first time are termed as secondary data. According to M. M. Blair, "Secondary data are those already in existence and which have been collected for some other purpose." Secondary data may be abstracted from existing records, published sources or unpublished sources.

**Some Advantages of using Secondary data:**

1. The data's already there- no hassles of data collection
2. It is less expensive

3. The investigator is not personally responsible for the quality of data ("I didn't do it")

**Some disadvantages of using Secondary data:**

1. The investigator cannot decide what is collected (if specific data about something is required, for instance).
2. One can only hope that the data is of good quality
3. Obtaining additional data (or even clarification) about something is not possible (most often) standard connotations".

The distinction between primary and secondary data is a matter of degree only. The data which are primary in the hands of one become secondary for all others. Generally the data are primary to the source who collects and processes them for the first time. It becomes secondary for all other sources, who use them later. For example, the population census report is primary for the Registrar General of India and the information from the report are secondary for all of us.

Both the primary and secondary data have their respective merits and demerits. Primary data are original as they are collected from the source. So they are more accurate than the secondary data. But primary data involves more money, time and energy than the secondary data. In an enquiry, a proper choice between the two forms of information should be made. The choice to a large extent depends on the "preliminaries to data collection".

**METHODS OF COLLECTION OF PRIMARY DATA**

The primary data are collected by the following methods.

1. Direct personal investigation. 2. Indirect personal investigation 3. Investigation thought questionnaire. 4. Investigation through questionnaire in the charge if enumerator 5. Investigation through local's reports.

**1. Direct Personal Investigation:**

According to this methods the investigator has to collect his information himself personally form the source concerned. It means the investigator should be are the spot where the enquiry concerned. It means the investigator should be at the spot where the enquiry is being conducted, it is also expected that the investigator should be very polite and courteous. Further he should acquaint himself with the surrounding situation and must know their local customs and tradition.

**Advantages**: 1. The information collected by this methods is reliable and accurate 2. It is a good method for intensive investigation 3. This method gives a satisfactory result provided the scope of inquiry is narrow.

**Disadvantages**: 1.This method is not suitable for extensive inquiry 2. It's required a lot of expenses and time 3. The bias on the part of investigator can damage the whole inquiry 4. Sometimes the informant may be reluctant to answered the question

**2. Indirect Personal Investigation:**

This method is used when the informants are reluctant to give the definite information. e.g., if a government servant is asked to give the information regarding his income. He will not be willing to give the information for the additional income which he earned by doing part time worked. In such cases what is done? The investigators puts the informant some suitable indirect question which provides him some suitable information. Thus the only difference between the first and the second methods is that in the first methods he investigator puts directs question and collect the information while in the second methods no direct question is put to the informant but only indirect questions are asked. Even then, if it is not possible for the investigator to collect the information by the above

methods then the information is collected through indirect sources, i.e. from the persons who have full knowledge of the problem under study. The persons from whom the desired information is collected are known as witnesses.

Usually a list of question is prepared which is put before the collected by this method largely depends upon the persons who are selected collected by this method largely depends upon the persons who are selected to give information. Hence it is necessary to take the following precautions for the selection of the informant.

**3. Investigation through questionnaire:**

According to this method a standard list of questions relating to the particular investigation is prepared. This list of questions is called a questionnaire. The data are collected "By sending the questionnaire to the informants and requesting them to return the questionnaire after answering the questions. " This method is an important one and is usually used by research workers, non-official bodies and private individuals.

**Choice of Questionnaire**:

The success of the investigation largely depends upon the proper choice the questions to be put to the informants. While preparing a questionnaire the following points should be kept in mind.

i. **Short and clear**: -

The questions should be short and clear so as to be easily intelligible to every man. There should be no ambiguity in the questions. If some technical terms are used in the questionnaire, their definitions should be given.

**ii. Few in number and easy**: The questions should be few in number. A large number of the questions would harm the informants because they take much time

to answer, with the result they would not pay much attention to ever question and would try to save their akin by giving vague answers.

**iii. Moreover the questions should be easy to answer**.

**a. Definiteness**: The questions should be such the answers of which are definite and exact. Preferably the questions should be such the replies of which are in the form of "Yes" or "No" Such questions should not be framed the replies of which are vague in nature because such replies are of no use to a statistician.

**b. Corroborating in nature**: The questions should be such that their replies check the value replies and truth can be easily verified from them.

**c. Non-confidential information**: The questions framed should not be such which call the confidential information of the informants. This will injure the feelings with the result that they would not give proper answer.

**d. Logical sequence**: The questions framed should be put in some logical order; their replies should also be put in the same order because this would facilities the work.

**4. Investigation through questionnaire in charge of enumerators**:

According into this method enumerators are appointed who go to the informants with the questionnaire and help them in recording the answer. Here the enumerators explain the background, aim and object of the problem under investigation and emphasize the necessity of giving correct answer. They also help the informants in understanding some t4echnical terms of question the concept of which is not clear to the informants. Thus the questionnaire is filled by the informants in the presence and help of the enumerators.

**5. Investigation through local reports**:

According to this method the collection of data is neither through the questionnaire nor through the enumerators but through local correspondents. This

method of collecting the data is not reliable and it should be used only at those places where the purpose the investigation is served by rough estimates.

## COLLECTION OF SECONDARY DATA

The secondary data are those which have already been collected by someone other than the investigator himself, and as such the problems associated with the original collection of data do not arise here. The secondary data can be collected directly either form published or unpublished sources. The following are the sources of published at from which secondary data can be collected.

**1. Official publications**, i.e. the publication of the central statistical office, Karachi , Ministry of Finance , Ministry of Food, Agriculture, Lahore, Industry, etc… the provincial statistical Bureau, etc.

**2. Semi-Official publications** , etc., the publication issued by the state Bank of Pakistan Railway Board , Board of Economic Enquiry , District councils, Municipalities, Central Cotton Committee, etc

**3. Publication** of trade-association, chambers of commerce, co-operative societies, and unions.

**4. Research publication**, submitted by research workers, economists, University bureaus, and other institutions.

**5. Technical or trade journals**. Sources of Unpublished Data: The secondary data are also available from the unpublished data. Type of material can be obtained from the chamber of commerce, trade associations, labor bureaus and research workers.

**Scrutiny of Secondary Data:**

In the words of Bowley, " It is never safe to take published statistics at their face value without knowing their meaning and limitations and it is always

necessary to criticize arguments that can be based on them , " Thus the data collected by some other person should not be fully depended as they might have pitfalls. Thus it becomes necessary to find out the inconsistencies probable errors and omissions in the data. This necessities the scrutiny of secondary data because it is just possible that the data might be inaccurate, inadequate or even unsuitable for the purposes of investigation. Hence the secondary data should possess the following qualities: **1. Reliability 2. Suitability 3. Adequacy.**

**CENSUS AND SAMPLING:**

It is obvious that when whole population is taken into account, data collection is called Census Method, whereas when a small group that is representative of the entire population is used, it is called a Sample Method.

A **census** is the procedure of systematically acquiring and recording information about the members of a given population. It is a regularly occurring and official count of a particular population. The term is used mostly in connection with national population and housing censuses; other common censuses include agriculture, business, and traffic censuses. A census is a study of every unit, everyone or everything, in a population. It is known as a complete enumeration, which means a complete count.

It is obvious that when whole population is taken into account, data collection is called Census Method

**Let us make an in-depth study of the two methods for collecting statistical data.**

## 1. Census Method:

The data which are collected by the investigator himself is called primary data. Census data can be thought of as primary data.

When the data collector or investigator collects data or information about each and every item in the population and other related areas, it is known as census method. As this method deals with the investigation of the entire population, it is also called complete enumeration method.

If a survey covers 100 p.c. population, it is called a census method. In other words, here each and every item or unit constituting the 'entire population' or 'the universe' is selected for statistical enquiry. If a statistical enquiry is conducted to study the nature and pattern of urbanisation, then the universe consists only of the urban population of India. This method is called complete enumeration method because information from each and every unit belonging to India's urban population is collected.

### Merits of Census Method:

First, for an extensive study (however expensive it may be) this method is considered to be an ideal one. For example, in the population census, we obtain quite a large number of key information, such as birth rate, death rate, infant mortality rate, literacy rate, ratio of urban-rural population, trend on urbanisation and so on.

Secondly, accuracy in the results is obtained. The data collected are more accurate and reliable under this census method since information are gathered from various angles. However, reliability in data and their accuracy in results are surely obtained provided enumerators do their work honestly and sincerely.

**Demerits of Census Method:**

First, being extensive in nature, the complete enumeration method is much expensive since a considerable amount of money, time, and labour are demanded and involved.

Secondly, this method is often not feasible or practicable because the concept of the 'universe' is hypothetical. Since "universe" is the basis of data collection, its applicability becomes very much limited. This method cannot be met with urgency.

This means that if an urgent statistical information for the entire population for policy purposes is needed, this method will surely be less helpful. In other words, it is too cumbersome and inefficient to obtain a complete picture of the target population.

Thirdly, in the census method, often large number of non-sampling errors creeps in. This means that results obtained may not be uniform.

**2. Sample Survey Method:**

Instead of the census method, data analysts often consider a portion or a sample of the population. If a sample population—instead of full survey of a population—is investigated then we have sample survey method or portion enumeration method.

A sample is anything less than a full survey of a population. For example, liquor consumption among college and university students is to be investigated. For this purpose, a small number of students following a particular prescribed technique will be picked up (sample drawn) and their habits for consumption of liquor will be investigated. The primary objective of such sample enquiry is to estimate some characteristics of the population from which the sample is selected.

**Merits of Sample Survey Method:**

Sample survey method has many advantages over census method. That is why this method is more popular than the latter method. In the words of A. C. Rosander, **"**If carefully designed, the sample is not only considerably cheaper but may give results which are just accurate and sometimes more accurate than those of a census technique".

**This method is preferable for the following reasons:**

First, since only a part of the population is investigated under this method, it takes less time, less money and less labour. There is saving in time also since sampling enquiry requires less fieldwork, tabulation and data processing than a full survey method. Sample survey method can be conducted when the investigators face the problem of budget constraint. It is cheaper to collect information from a sample group. However, for conducting the entire enquiry, small group of investigators or specialist investigators are paid huge amount and employed but the output is much more. That is why it is also said that this method results in reduced unit cost of enquiry.

Secondly, conclusions and results obtained from this method are more accurate and reliable as fewer or chosen sample units are surveyed. Trained personnel are usually employed to collect data and investigate the problem. Above all, these people use sophisticated and latest designed techniques so that results become more accurate and reliable. Further, it is true that sampling errors cannot be avoided, but such errors are easier to estimate and control.

Thirdly, the small sample data provide a good benchmark for the entire population.

Finally, from the administrative point of view, the sample method is considered as an ideal one as the organisation as well as administration considers

the process practically more convenient. Administrative network does not usually require being considerably elaborate or extensive.

**Sampling*:***

A *sample* is the group of people who take part in the investigation. The people who take part are referred to as "participants". *Sampling* is the process of selecting participants from the population.

Samples are parts of a population. For example, you might have a list of information on 100 people out of 10,000 people. You can use that list to make some assumptions about the entire population's behavior. Unfortunately, it's not *quite* that simple. When you do stats, your sample size must be optimal — not too large or too small. Then once you've decided on a sample size you must use a sound technique for actually drawing the sample from the population.

Sampling Methods can be classified into one of two categories:

I. **Probability Sampling**: Sample has a known probability of being selected

**II.Non-probability Sampling**: Sample does not have known probability of being selected as in convenience or voluntary response survey.

**I. Probability Sampling**

In probability sampling it is possible to both determine which sampling units belong to which sample and the probability that each sample will be selected. The following sampling methods are examples of **probability sampling**:

- Simple Random Sampling (SRS)
- Stratified Sampling
- Cluster Sampling
- Systematic Sampling

**II. Non-probability Sampling**

The following sampling methods that are listed in your text are types of **non-probability sampling that should be avoided**:

- **Quota Sampling**
- **Purposive Sampling**
- **Convenience sample**

**Probability Sampling:**

**1. Simple Random Sampling (SRS):**

It is a technique in which sample is drawn that each and every unit in the population has an equal and independent chance of being included in the sample.

**2. Stratified Sampling:**

A stratified sample is a mini-reproduction of the population. Before sampling, the population is divided into characteristics of importance for the research. For example, by gender, social class, education level, religion, etc. Then the population is randomly sampled *within* each category or **stratum**. If 38% of the population is college-educated, then 38% of the sample is randomly selected from the college-educated population. Stratified samples are as good as or better than random samples, but they require fairly detailed advance knowledge of the population characteristics, and therefore are more difficult to construct.

**3. Cluster Sampling**:

It is very different from Stratified Sampling. With **cluster sampling** one should divide the population into groups (clusters) and obtain a simple random sample of so many clusters from all possible clusters and also obtain data on every sampling unit in each of the randomly selected clusters.

It is important to note that, unlike with the strata in stratified sampling, the clusters should be microcosms, rather than subsections, of the population. Each cluster should be heterogeneous. Additionally, the statistical analysis used with cluster sampling is not only different, but also more complicated than that used with stratified sampling.

**4. Systematic Sampling:**

It is also known as quasi-random sampling. A systematic sampling is selected at random sampling. When a computer list of the population is available, this method is used. We arrange the items in numerical, alphabetical, geographical or any other order.

## Non-probability Sampling:

### 1. Quota Sampling:

Quota sampling is designed to overcome the most obvious flaw of availability sampling. Rather than taking just anyone, you set quotas to ensure that the sample you get represents certain characteristics in proportion to their prevalence in the population. Note that for this method, you have to know something about the characteristics of the population ahead of time. Say you want to make sure you have a sample proportional to the population in terms of gender - you have to know what percentage of the population is male and female, then collect sample until yours matches. Marketing studies are particularly fond of this form of research design.

### 2. Purposive Sampling:

Purposive sampling is a sampling method in which elements are chosen based on purpose of the study. Purposive sampling may involve studying the entire population of some limited group (sociology faculty at Columbia) or a subset of a population (Columbia faculty who have won Nobel Prizes). As with other non-probability sampling methods, purposive sampling does not produce a sample that is representative of a larger population, but it can be exactly what is needed in some cases - study of organization, community, or some other clearly defined and relatively limited group.

### 3. Convenience sample:

A **convenience sample** is a matter of taking what you can get. It is an **accidental** sample. Although selection may be unguided, it probably is not random, using the correct definition of everyone in the population having an equal chance of being selected. Volunteers would constitute a convenience sample.

## CLASSIFICATION

"Classified and arranged facts speak of themselves, and narrated they are as dead as mutton" This quote is given by J.R. Hicks.

The process of dividing the data into different groups ( viz. classes) which are homogeneous within but heterogeneous between themselves, is called a classification. It helps in understanding the salient features of the data and also the comparison with similar data. For a final analysis it is the best friend of a statistician.

### Methods of Classification:

The data is classified in the following ways:

1. According to attributes or qualities this is divided into two parts:

   (A) Simple classification

   (B) Multiple classifications.

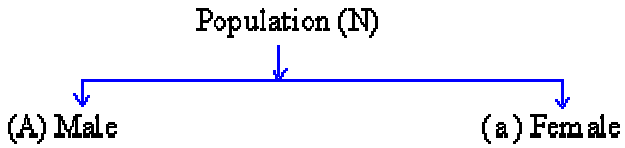
2. According to variable or quantity or classification according to class intervals.

### Qualitative Classification:

When facts are grouped according to the qualities (attributes) like religion, literacy, business etc., the classification is called as qualitative classification.

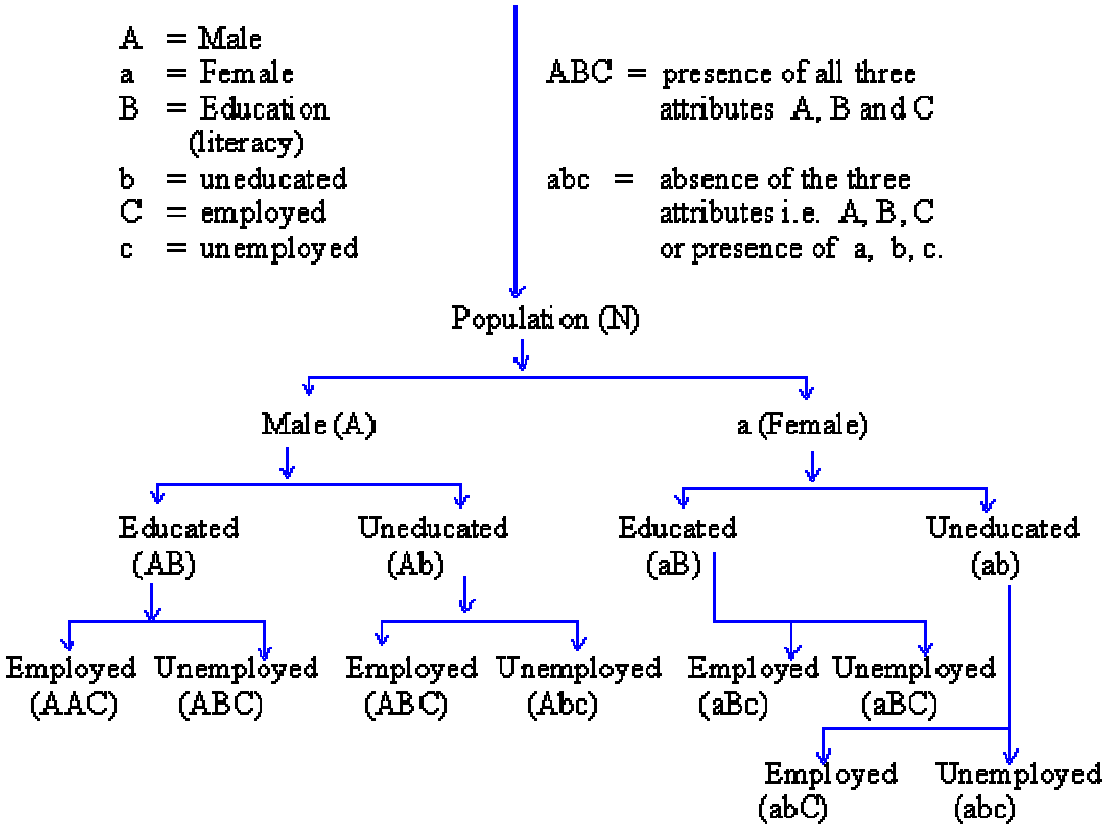
**(A) Simple Classification:**

It is also known as classification according to Dichotomy. When data (facts) are divided into groups according to their qualities, the classification is called as 'Simple Classification'. Qualities are denoted by capital letters (A, B, C, D .....) while the absence of these qualities are denoted by lower case letters (a, b, c, d, .... etc.) For example,




**(B) Manifold or multiple classifications:**

In this method data is classified using one or more qualities. First, the data is divided into two groups (classes) using one of the qualities. Then using the remaining qualities, the data is divided into different subgroups. For example, the population of a country is classified using three attributes: sex, literacy and business as,



**Classification according to class intervals or variables:**

The data which is expressed in numbers (quantitative data), is classified according to class-intervals. While forming class-intervals one should bear in mind that each and every item must be covered. After finding the least value of an item and the highest value of an item, classify these items into different class-intervals. For example if in any data the age of 100 persons ranging from 2 years to 47 years, is given, then the classification of this data can be done in this way:.

**Table - 1**

| Age in years | No. of persons |
|---|---|
| 0 - 10 | 5 |
| 10 - 20 | 9 |
| 20 - 30 | 32 |
| 30 - 40 | 34 |
| 40 - 50 | 20 |

**According to the class-intervals in classification the following terms are used:**

**i) Class-limits:** A class is formed within the two values. These values are known as the class-limits of that class. The lower value is called the lower limit and is denoted by $l1$ while the higher value is called the upper limit of the class and is denoted by $l2$. In the example given above, the first class-interval has $l1 = 0$ and $l2 = 10$.

**ii) Mid-value or class-mark:** The arithmetical average of the two class limits (i.e. the lower limit and the upper limit) is called the mid-value or the class mark of that class-interval. For example, the mid-value of the class-interval ( 0 - 10 ) is

$\frac{\ell_1+\ell_2}{2} = \frac{0+10}{2} = 5$, of $(10-20)$ it is $\frac{10+20}{2} = 15$ and so on.

**iii) Class frequency:**

The units of the data belong to any one of the groups or classes. The total number of these units is known as the frequency of that class and is denoted by fi or simply f. In the above example, the frequencies of the classes in the given order are 5, 9, 32, 34 and 40 respectively.

**iv) Class boundaries:**

The class limits are the largest or the highest and the smallest or the lowest values of the class. The two boundaries of the class are the lower limit and upper limit of the class. Class limit is also known as class boundaries. For example, take the class 10-20. The lowest value is 10 and the highest value is 20.

Classification is of two types according to the class-intervals - (i) Exclusive Method (ii) Inclusive Method.

**i) Exclusive Method:**

In this method the upper limit of a class becomes the lower limit of the next class. It is called ' Exclusive ' as we do not put any item that is equal to the upper limit of a class in the same class; we put it in the next class, i.e. the upper limits of classes are excluded from them. For example, a person of age 20 years will not be included in the class-interval ( 10 - 20 ) but taken in the next class ( 20 - 30 ), since in the class interval ( 10 - 20 ) only units ranging from 10 - 19 are included. The exclusive-types of class-intervals can also be expressed as :

0 and below 10 or 0 - 9.9
10 and below 20 or 10 - 19.9
20 and below 30 or 20 - 29.9 and so on

**ii) Inclusive Method:**

In this method the upper limit of any class interval is kept in the same class-interval. In this method the upper limit of a previous class is less by 1 from the lower limit of the next class interval. In short this method allows a class-interval to include both its lower and upper limits within it. For example :

**Table - 2**

| Inclusive method | | Inclusive method | |
|---|---|---|---|
| Class | Frequency | Class | Frequency |
| 0 - 4 | 5 | 0 - 4.9 | 5 |
| 5 - 9 | 7 | 5 - 9.9 | 7 |
| 10 - 14 | 9 | 10 - 14.9 | 9 |
| 15 - 19 | 12 | 15 - 19.9 | 12 |
| 20 - 24 | 11 | 20 - 24.9 | 11 |
| 25 - 29 | 14 | 25 - 29.9 | 14 |

**Open-end Class Intervals:**

In any question when the lower limit of the first class-interval or the upper limit of the last class-interval, are not given then subtract the class length of the next immediate class-interval from the upper limit. This will give us the lower limit of the first class-interval. Similarly add the same class length to the lower limit of the last class-interval. But always notice that the lower limit of the first class ( i.e. the lowest class) must not be negative or less than 0. For example:

**Table - 3**

| With open ends | Completed classes | With open ends | Completed classes |
|---|---|---|---|
| Below 10 | 0 - 10 | Below 10 | 0 - 10 |
| 10 - 20 | 10 - 20 | 10 - 25 | 10 - 25 |
| 20 - 30 | 20 - 30 | 25 - 40 | 25 - 40 |
| 30 - 40 | 30 - 40 | 40 - 70 | 40 - 70 |
| 40 - 50 | 40 - 50 | above 70 | 70 - 100 |
| above 50 | 50 - 60 | | |

**Relative Frequency Distribution:**

The relative frequency of a class is the frequency of the class divided by the total number of frequencies of the class and is generally expresses as a percentage.

**Example** The weight of 100 persons was given as under:

| Weights (in Kgs) | 60-62 | 63-63 | 66-68 | 69-71 | 72-74 |
|---|---|---|---|---|---|
| No. of persons ($f_i$) | 5 | 18 | 42 | 27 | 8 |

**Solution:**

**Table - 4**

| Relative frequency table | | |
|---|---|---|
| Weight (in Kg.) | No. of persons ($f_i$) | Relative frequency |
| 60 - 62 | 5 | $\frac{5}{100} \times 5\%$ or 0.05 |
| 63 - 65 | 18 | $\frac{18}{100} \times 18\%$ or 0.18 |
| 66 - 68 | 42 | $\frac{42}{100} \times 42\%$ or 0.42 |
| 69 - 71 | 27 | $\frac{27}{100} \times 27\%$ or 0.27 |
| 72 - 74 | 8 | $\frac{8}{100} \times 8\%$ or 0.08 |
| Total | $\Sigma f_i = 100$ | |

**Note :** The word frequency of a class means, the number of times the class is repeated in the data or the total number of items or observations of the data belongs to that class.

**Cumulative Frequency:**

Many a times the frequencies of different classes are not given. Only their cumulative frequencies are given. The total frequency of all values less than or equal to the upper class boundary of a given class-interval is called the cumulative frequency up to and including that class interval. In this situation both the limits of a class-interval are not written; either lower or upper limit is written. These cumulative frequencies are called less than or more than cumulative frequencies. For example,

| Class-interval | 0 - 10 | 10 - 20 | 20 - 30 | 30 - 40 | 40 - 50 |
|---|---|---|---|---|---|
| Frequency | 4 | 9 | 5 | 12 | 15 |

**Table - 5**

| Less than cumulative frequency | | More than cumulative frequency | |
|---|---|---|---|
| ( Upper limits ) | ( cum. freq. ) | (Lower limits ) | ( cum. freq. ) |
| Less than 10 | 4 | More than 0 | 45 = 41 + 4 |
| Less than 20 | 4 + 9 = 13 | More than 10 | 41 = 32 + 9 |
| Less than 30 | 13 + 5 = 18 | More than 20 | 32 = 27 + 5 |
| Less than 40 | 18 + 12 = 30 | More than 30 | 27 = 15 + 12 |
| Less than 50 | 30 + 15 = 45 | More than 40 | 15 |

**Preparation of Frequency Distribution:**

We shall now study how to classify the raw data in a tabular form. Consider the data collected by one of the surveyors, interviewing about 50 people. This is as follows:

Size of the shoes: 2, 5, 6, 8, 2, 5, 6, 7, 6, 8, 7, 4, 3, .. This is called the raw data. Here some values repeat themselves. For instance the size 5 is repeated 10 times in 50 people. We say that the value of 5 of the variate has the frequency of 10. Frequency means the number of times a value of the variate or an attribute, as the case may be, is repeated in the data.

A table which shows each value of the characteristic with its corresponding frequency is known as a **Frequency Distribution**. The procedure of preparing such a table is explained as below:

**Discrete variate:** Consider the raw data which gives the size of shoes of 30 persons

2, 5, 6, 4, 5, 7, 4, 4, 6, 2
3, 5, 5, 4, 5, 6, 5, 4, 3, 2
4, 4, 5, 4, 5, 5, 3, 2, 4, 4

The least value is 2 and the highest is 7. All sizes are integers between 2 and 7 ( both inclusive ). We can prepare a frequency distribution table as follows :

**Table - 6**

| Sizes of shoes | Tally Marks | Frequency |
|---|---|---|
| 2 | \|\|\|\| | 4 |
| 3 | \|\|\| | 3 |
| 4 | 𝍸 𝍸 | 10 |
| 5 | 𝍸 𝍸 | 9 |
| 6 | \|\|\| | 3 |
| 7 | \| | 1 |
| Total | | 30 |

In this example the size difference from 2 to 7 is very small. If the range of a variate is very large, it is inconvenient to prepare a frequency distribution for each value of the variate. In such a case we divide the variate into convenient groups and prepare a table showing the groups and their corresponding frequencies. Such a table is called a **grouped frequency distribution**.

Consider the marks (out of 100) of 50 students as below:

40, 39, 43, 62, 30, 47, 33, 31, 17, 28
36, 29, 40, 32, 39, 24, 57, 42, 15, 30
50, 52, 47, 65, 31, 07, 37, 47, 17, 20

25, 53, 65, 85, 89, 56, 55, 41, 43, 10
44, 40, 69, 22, 40, 65, 39, 36, 71, 12

The range of the variate (marks) is very large. Also we are eager to know the performance of the students. The passing limit is 35 and above. Marks between 35 and 44 form the third class (or grade). Marks ranging between 45 - 59 are considered as second class and 60 - 100 form the first class. Thus we have a grouped frequency distribution as:

**Table - 7**

| Marks | Tally Marks | Frequency |
|---|---|---|
| 0 - 34 | 𝍸 𝍸 𝍸 I | 16 |
| 35 - 44 | 𝍸 𝍸 𝍸 III | 18 |
| 45 - 59 | 𝍸 IIII | 9 |
| 60 - 100 | 𝍸 II | 7 |
| Total | | 50 |

**TABULATION:**

A table is a systematic arrangement of data into vertical column and horizontal rows. The process of arranging data into rows and column is called tabulation. The purpose of tabulation is to present the data in such a way that they become more meaning full and can easily understood by a common man.

It is the process of condensation of the data for convenience, in statistical processing, presentation and interpretation of the information.

A good table is one which has the following requirements:

1. It should present the data clearly, highlighting important details.
2. It should save space but attractively designed.
3. The table number and title of the table should be given
4. Row and column headings must explain the figures therein.
5. Averages or percentages should be close to the data.
6. Units of the measurement should be clearly stated along the titles or headings.
7. Abbreviations and symbols should be avoided as far as possible.
8. Sources of the data should be given at the bottom of the data.
9. In case irregularities creep in table or any feature is not sufficiently explained, references and foot notes must be given.
10. The rounding of figures should be unbiased.

Tabulation is the systematic arrangement of the statistical data in columns or rows. It involves the orderly and systematic presentation of numerical data in a form designed to explain the problem under consideration. Tabulation helps in drawing the inference from the statistical figures. Tabulation prepares the ground for analysis and interpretation. Therefore a suitable method must be decided carefully taking into account the scope and objects of the investigation, because it is very important part of the statistical methods.

**TYPES OF TABULATION:**

In general, the tabulation is classified in two parts, that is a simple tabulation, and a complex tabulation.

Simple tabulation, gives information regarding one or more independent questions. Complex tabulation gives information regarding two mutually dependent questions.

**One-Way Table:**

| ONE-WAY TABLE | |
|---|---|
| DIVISION | POPULATION (Millions) |
| Karachi | 10.875968 |
| Hyderabad | 14.186954 |
| Sukkur | 12.994401 |

This table gives us information regarding one characteristic information about the population in different divisions of Sindh. All questions that can be answered in ONE WAY TABLE are independent of each other. It is therefore an example of a simple tabulation, since the information obtained in it is regarding one independent question, that is the number of persons in various divisions of Sindh in millions.

**Two-Way Table:**

These types of table give information regarding two mutually dependent questions. For example, question is, how many millions of the persons are in the Divisions; the One-Way Table will give the answer. But if we want to know that in the population number, who are in the majority, male, or female. The Two-Way Tables will answer the question by giving the column for female and male. Thus the table showing the real picture of divisions sex wise is as under:

| TWO-WAY TABLE | | | |
|---|---|---|---|
| DIVISION | POPULATION (Millions) | | |
| | Male | Female | Total |
| Karachi<br>Hyderabad<br>Sukkur | | | |

**Three-Way Table:**

Three-Way Table gives information regarding three mutually dependent and inter-related questions. For example, from one-way table, we get information about population, and from two-way table, we get information about the number of male and female available in various divisions. Now we can extend the same table to a three way table, by putting a question, "How many male and female are literate?" Thus the collected statistical data will show the following, three mutually dependent and inter-related questions:

1. Population in various divisions.
2. Their sex-wise distribution.
3. Their position of literacy.

| | THREE-WAY TABLE | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| DIVISION | POPULATION (Millions) | | | | | | | | |
| | Male | | | Female | | | Total | | |
| | Literate | Illiterate | Total | Literate | Illiterate | Total | Literate | Illiterate | Total |
| Karachi | | | | | | | | | |
| Hyderabad | | | | | | | | | |
| Sukkur | | | | | | | | | |

This table gives information concerning the literacy of both male and female in various divisions of Sindh. From the table we can explain the sex which has more education in relation to division, and also, we can say whether literacy is low in rural areas than in urban areas.

**Essential parts of a table:**

A statistical table is divided into 8 parts, which are explained below.

**i. Title of the table:**

A title is a heading at the top of the table describing its contents. A title usually tells us, what is the nature of the data, where the data are, what time period do the data cover, how are the data classified.

**ii.** Caption**:**

The headings for various column and rows are called columns caption and row caption.

**iii.** Box **head:**

The portion of the table containing Column caption is called box head.

**iv.** Stub**:**

The portion of the table containing row caption is called stub.

**v.** Body **of the table:**

The body of the table contains the statistical data which have to be presented in different rows and column.

**vi. Prefatory notes or head notes:**

Prefatory note appears between title and body of the table and enclosed in brackets. It is used to throw some light about the units of measurements e.g. in lakhs, in thousand, in tones etc.

**vii. Foot note:**

A foot note an always given at the bottom of the table but above the source note. Afoot note is a statement about something which is not clear from headings, title, stubs and captions etc. suppose when the profit earned by a company is shown in a table footnote should define whether it is profit before tax or profit after tax.

**viii. Source note:**

A source note is placed immediately below the table but after the footnote. It refers to the source from where information has been taken.

**Question:**

According to 1972 census the population of Punjab was 37508 thousands, of which 19942 thousands male. During the same census the population of Baluchistan was 2405 thousands, of which 1272 thousands were male. During 1961 census, the population of bunjab was 25581 thousands, of which 13643 were male. During the same census the population of Baluchistan was 1161 thousands, of which 640 thousands were male. Arrange the above information's in a table and clearly indicate different parts of the table.

**Solution:**

**population of Punjab and Baluchistan provinces for 1961 and 1972 censuses** Prefatory note: (figure in thousands)

| | **Punjab** | | | **Baluchistan** | | |
|---|---|---|---|---|---|---|
| census | male | female | total | Male | female | Total |
| 1961 | 13643 | 11938 | 25581 | 640 | 521 | 1161 |
| 1972 | 19942 | 17566 | 37508 | 1272 | 1133 | 2405 |

**Foot note:** All areas including gawadar.

**Source note:** Population census report 1961 and 1972.

Tables and graphs are visual representations. They are used to organise information to show patterns and relationships. A graph shows this information by representing it as a shape. Researchers and scientists often use tables and graphs to report findings from their research. In newspapers, magazine articles, and on television they are often used to support an argument or point of view

## DIAGRAMMATIC AND GRAPHIC DISPLAYS

In the last chapter we have seen how to condense the mass of data by the method of classification and tabulation. It is not always easy for a layman to understand figures, nor is it is interesting for him. Apart from that too many figures are often confusing. One of the most convincing and appealing ways in which statistical results may be represented is through graphs and diagrams. It is for this reason that diagrams are often used by businessmen, newspapers, magazines, journals, government agencies and also for advertising and educating people.

**Presentation of data:**

The presentation of data refers to how mathematicians and scientists summarize and present data related to scientific studies and research. In order to present their points, they use various techniques and tools to condense and summarize their findings. These tools include the use of tables, graphs and subsets to provide an overview of their calculations and the data they have mined

Ways on how to present data:

1. **Textual Method**: The reader acquires information through reading the gathered data.

2. **Tabular Method**: Provides a more precise, systematic and orderly presentation of data in rows or columns

3. **Graphical Method**: The utilization of graphs is most effective method of visually presenting statistical results or findings.

We have discussed the techniques of classification and tabulation that help us in organizing the collected data in a meaningful fashion. However, this way of presentation of statistical data does not always prove to be interesting to a layman. Too many figures are often confusing and fail to convey the massage effectively.

One of the most effective and interesting alternative way in which a statistical data may be presented is through diagrams and graphs. There are several ways in which statistical data may be displayed pictorially such as different types of graphs and diagrams. The commonly used diagrams and graphs to be discussed in subsequent paragraphs are given as under:

**Types of Diagrams:**

1. Simple Bar
2. Multiple Bar
3. Staked Bar or Sub-Divided Bar or Component Bar
    - Simple Component Bar
    - Percentage Component Bar
    - Sub-Divided Rectangular Bar
    - Pie diagram

**Types of Graphs:**

1. Histogram
2. Frequency Curve and Polygon
3. Ogive curves

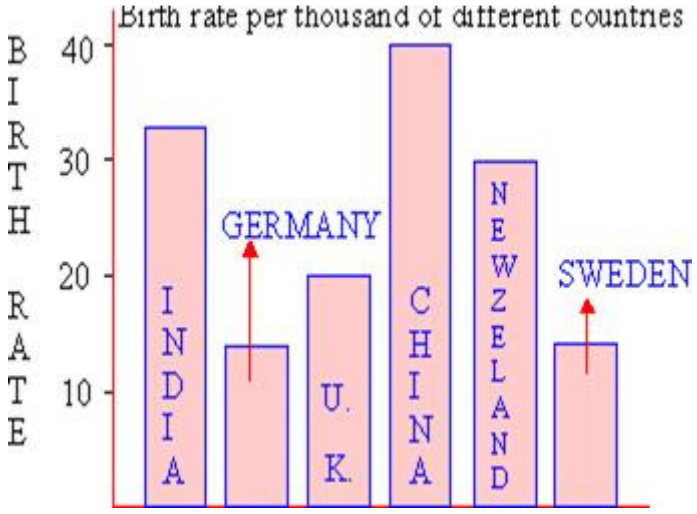
**Bar Diagrams:**

**1) Simple 'Bar diagram':-** It represents only one variable. For example sales, production, population figures etc. for various years may be shown by simple bar charts. Since these are of the same width and vary only in heights ( or lengths ), it becomes very easy for readers to study the relationship. Simple bar diagrams are very popular in practice. A bar chart can be either vertical or horizontal; vertical bars are more popular.

**Illustration :-** The following table gives the birth rate per thousand of different countries over a certain period of time.

| Country | Birth rate | Country | Birth rate |
|---|---|---|---|
| India<br>Germany<br>U. K. | 33<br>15<br>20 | China<br>New Zealand<br>Sweden | 40<br>30<br>15 |

Represent the above data by a suitable diagram.



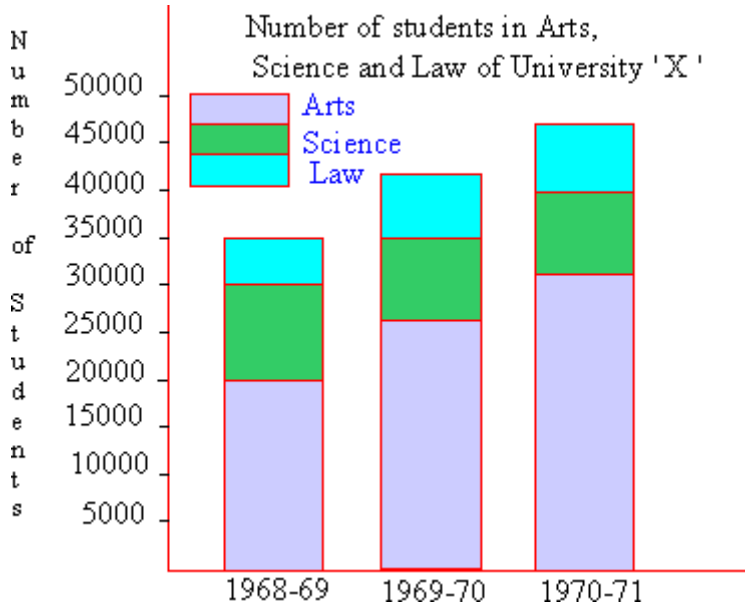

Comparing the size of bars, you can easily see that China's birth rate is the highest while Germany and Sweden equal in the lowest positions. Such diagrams are also known as component bar diagrams.

**2) Sub - divided Bar Diagram:-**

While constructing such a diagram, the various components in each bar should be kept in the same order. A common and helpful arrangement is that of presenting each bar in the order of magnitude with the largest component at the bottom and the smallest at the top. The components are shown with different shades or colors with a proper index.

**Illustration:-** During 1968 - 71, the number of students in University ' X ' are as follows. Represent the data by a similar diagram.

| Year | Arts | Science | Law | Total |
|---|---|---|---|---|
| 1968 - 69 | 20,000 | 10,000 | 5,000 | 35,000 |
| 1969 - 70 | 26,000 | 9,000 | 7,000 | 42,000 |
| 1970 - 71 | 31,000 | 9,500 | 7,500 | 48,000 |



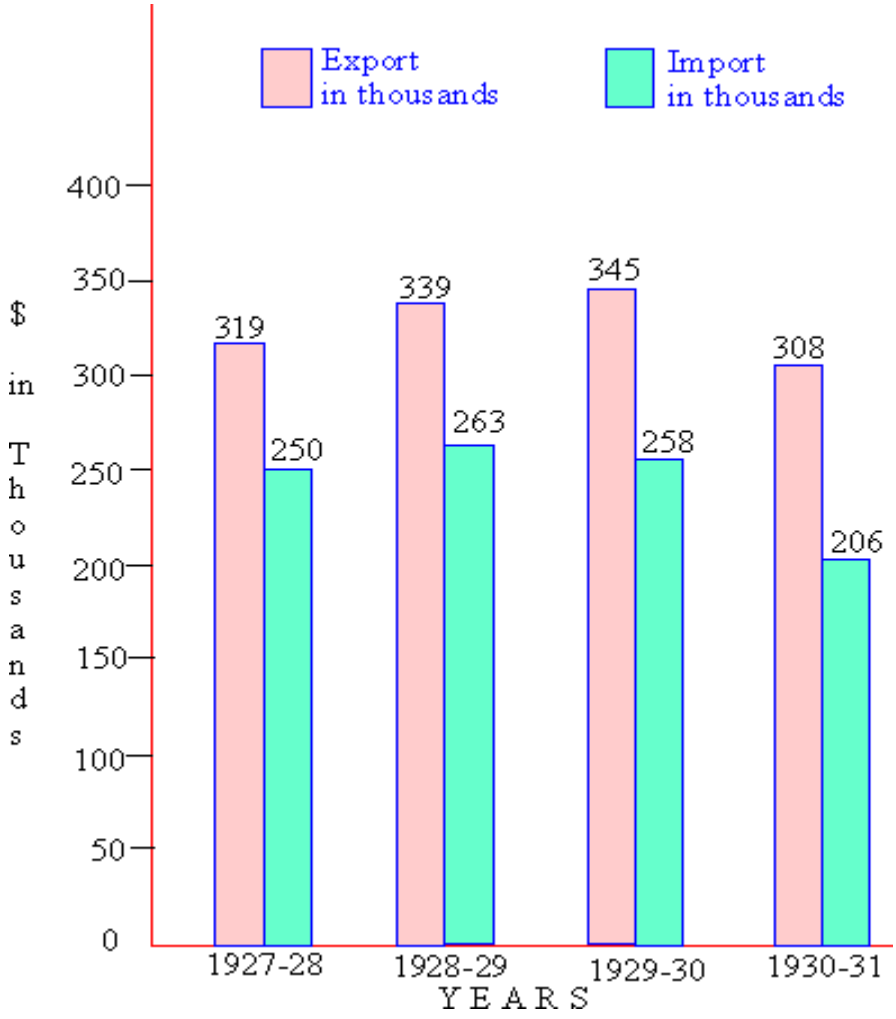

### 3) Multiple Bar Diagram:-

This method can be used for data which is made up of two or more components. In this method the components are shown as separate adjoining bars. The height of each bar represents the actual value of the component. The components are shown by different shades or colors. Where changes in actual values of component figures only are required, multiple bar charts are used.

**Illustration:-** The table below gives data relating to the exports and imports of a certain country X ( in thousands of dollars ) during the four years ending in 1930 - 31.

| Year | Export | Import |
|---|---|---|
| 1927 - 28 | 319 | 250 |
| 1928 - 29 | 339 | 263 |
| 1929 - 30 | 345 | 258 |
| 1930 - 31 | 308 | 206 |

Represent the data by a suitable diagram



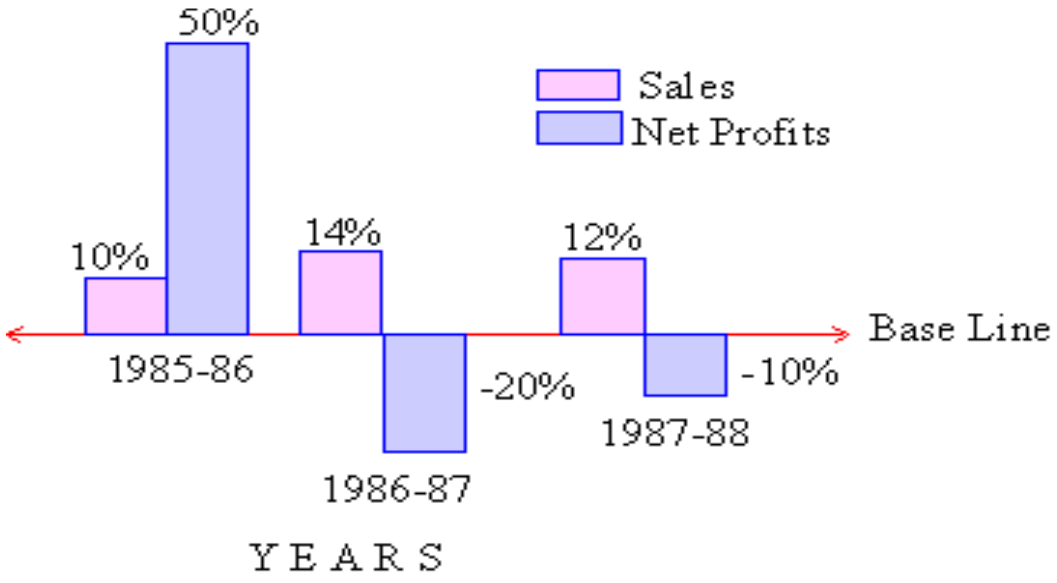
**4) Deviation Bar Charts:-**

Deviation bars are used to represent net quantities - excess or deficit i.e. net profit, net loss, net exports or imports, swings in voting etc. Such bars have both positive and negative values. Positive values lie above the base line and negative values lie below it.

**Illustration:-**

| Years | Sales | Net profits |
|---|---|---|
| 1985 - 86 | 10% | 50% |
| 1986 - 87 | 14% | -20 |
| 1987 - 88 | 12% | -10% |

Present the above data by a suitable diagram showing the sales and net profits of private industrial companies.



**Pie Chart:**

i) Geometrically it can be seen that the area of a sector of a circle taken radially, is proportional to the angle at its center. It is therefore sufficient to draw angles at the center, proportional to the original figures. This will make the areas of the sector proportional to the basic figures.

For example, let the total be 1000 and one of the component be 200, then the angle will be

$$\left(\frac{200}{1000}\right) \times 360^0 = 72^0$$

In general, angle of sector at the center corresponding to a component

$$= \left(\frac{\text{Component}}{\text{Total}}\right) \times 360^0$$

ii) When a statistical phenomenon is composed of different components which are numerous (say four or more components), bar charts are not suitable to represent them because, under this situation, they become very complex and their visual impressions are questioned. A pie diagram is suitable for such situations. It is a circular diagram which is a circle (pie) divided by the radii, into sectors ( like slices of a cake or pie ). The area of a sector is proportional to the size of each component.

iii) As an example consider the yearly expenditure of a Mr. Ted, a college undergraduate.

| | |
|---|---|
| Tuition fees | $ 6000 |
| Books and lab. | $ 2000 |
| Clothes / cleaning | $ 2000 |
| Room and boarding | $12000 |
| Transportation | $ 3000 |
| Insurance | $ 1000 |
| Sundry expenses | $ 4000 |
| Total expenditure | = $ 30000 |

Now as explained above, we calculate the angles corresponding to various items (components).

Tuition fees = $\frac{6000}{30000} \times 360^0 = 72^0$

Book and lab = $\frac{2000}{30000} \times 360^0 = 24^0$

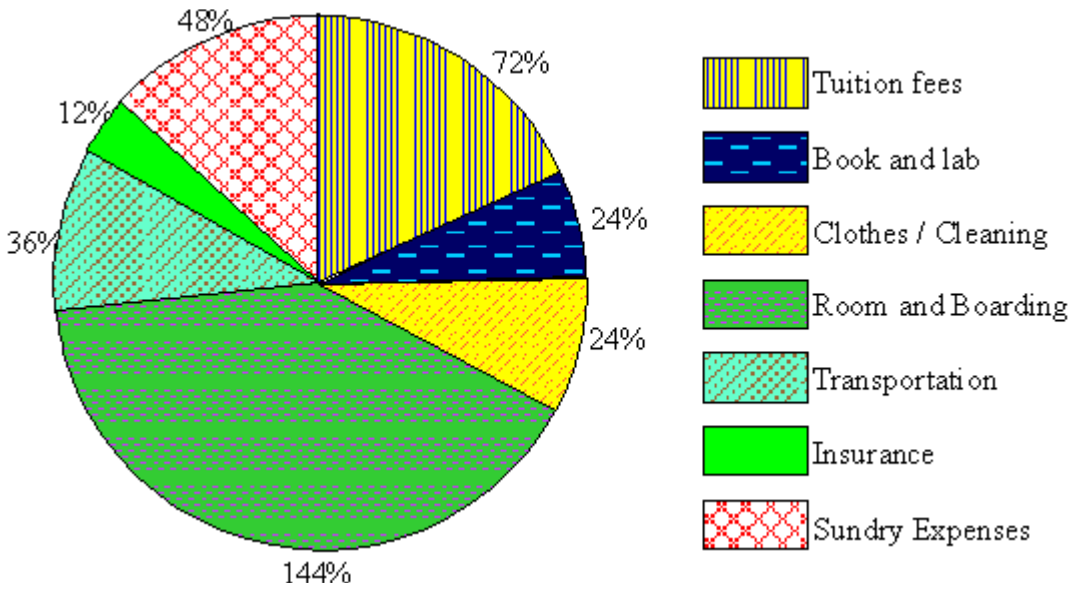
Clothes / cleaning = $\frac{2000}{30000} \times 360^0 = 24^0$

Room and boarding = $\frac{12000}{30000} \times 360^0 = 144^0$

Transportation = $\frac{3000}{30000} \times 360^0 = 36^0$

Insurance = $\frac{1000}{30000} \times 360^0 = 12^0$

Sundry expenses = $\frac{4000}{30000} \times 360^0 = 48^0$




**Uses:-**

A pie diagram is useful when we want to show relative positions ( proportions ) of the figures which make the total. It is also useful when the components are many in number.

Note:- The sectors of the circle ( i.e. of a pie diagram) are ordered from largest to the smallest for easier interpretation of the data and they must be drawn in the counter-clockwise direction.

**Graphs:**

A graph is a visual representation of data by a continuous curve on a squared ( graph ) paper. Like diagrams, graphs are also attractive, and eye-catching, giving a bird's eye-view of data and revealing their inner pattern.

**Graphs of Frequency Distributions:-**

The methods used to represent a grouped data are :-

1.Histogram
2.FrequencyPolygon
3.FrequencyCurve
4. Ogive or Cumulative Frequency Curve

1.**Histogram: -** It is defined as a pictorial representation of a grouped frequency distribution by means of adjacent rectangles, whose areas are proportional to the frequencies.

To construct a Histogram, the class intervals are plotted along the x-axis and corresponding frequencies are plotted along the y - axis. The rectangles are constructed such that the height of each rectangle is proportional to the frequency of the that class and width is equal to the length of the class. If all the classes have equal width, then all the rectangles stand on the equal width. In case of classes having unequal widths, rectangles too stand on unequal widths (bases). For open-classes, Histogram is constructed after making certain assumptions. As the rectangles are adjacent leaving no gaps, the class-intervals become of the inclusive type, adjustment is necessary for end points only.

For example, in a book sale, you want to determine which books were most popular, the high priced books, the low priced books, books most neglected etc. Let us say you sold total 31 books at this book-fair at the following prices.

$ ....2, $ 1, $ 2, $ 2, $ 3, $ 5, $ 6, $ 17, $ 17, $ 7, $ 15, $ 7, $ 7, $ 18, $ 8, $ 10, $ 10, $ 9, $ 13, $ 11, $ 12, $ 12, $ 12, $ 14, $ 16, $ 18, $ 20, $ 24, $ 21, $ 22, $ 25.

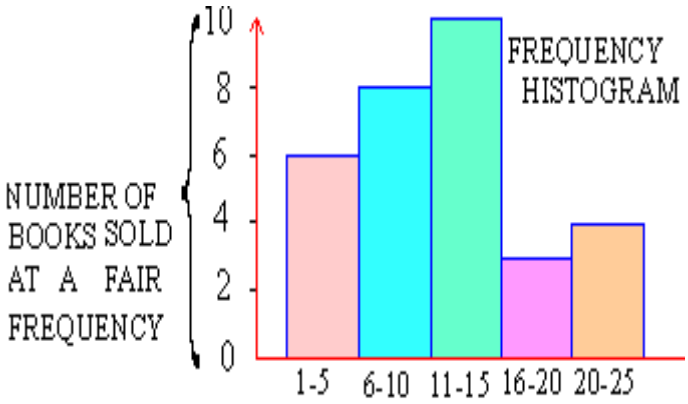
The books are ranging from $1 to $25. Divide this range into number of groups, class intervals. Typically, there should not be fewer than 5 and more than 20 class-intervals are best for a frequency Histogram.

Our first class-interval includes the lowest price of the data and, the last-interval of course includes, the highest price. Also make sure that overlapping is avoided, so that, no one price falls into two class-intervals. For example you have class intervals as 0-5, 5-10, 10-15 and so on, then the price $10 falls in both 5-10 and 10-15. Instead if we use $1 - $5, $6=$10, the class-intervals will be mutually exclusive.

Therefore now we have distribution of books at a book-fair

| Class-interval | Frequency |
|---|---|
| $ 1- $ 5 | 6 |
| $6 - $10 | 8 |
| $11 - $15 | 10 |
| $16 - $20 | 3 |
| $21 - $25 | 4 |
| Total | $n = \Sigma fi = 31$ |

Note that each class-interval is of equal width i.e. $5 inclusive. Now we draw the frequency Histogram as under.




**Relative Frequency Histogram:-**

It uses the same data. The only difference is that it compares each class-interval with the total number of items i.e. instead of the frequency of each class-interval, their relative frequencies are used. Naturally the vertical axis (i.e. y-axis) uses the relative frequencies in places of frequencies.

In the above case we have

| Class-interval | Frequency | Relative frequency |
|---|---|---|
| $ 1 - $ 5 | 6 | 6/31 |
| $ 6 - $10 | 8 | 8/31 |
| $11 - $15 | 10 | 10/31 |
| $16 - $20 | 3 | 3/31 |
| $21 - $25 | 4 | 4/31 |

The Histogram is same as in above case.

**Construction of Histogram when class-intervals are unequal:-**

In a Histogram, a rectangle is proportional to the frequency of the concern class-interval. Naturally, if the class-intervals are of unequal widths, we have to adjust the heights of the rectangle accordingly. We know that the area of a

rectangle = l. h. Now suppose the width ( l ) of a class is double that of a normal class interval, its height and thus the corresponding frequency must be halved. After this precaution has been taken, the construction of the Histogram of classes of unequal intervals is the same as before.

**Note :- The smallest class-interval should be assumed to be " NORMAL "**

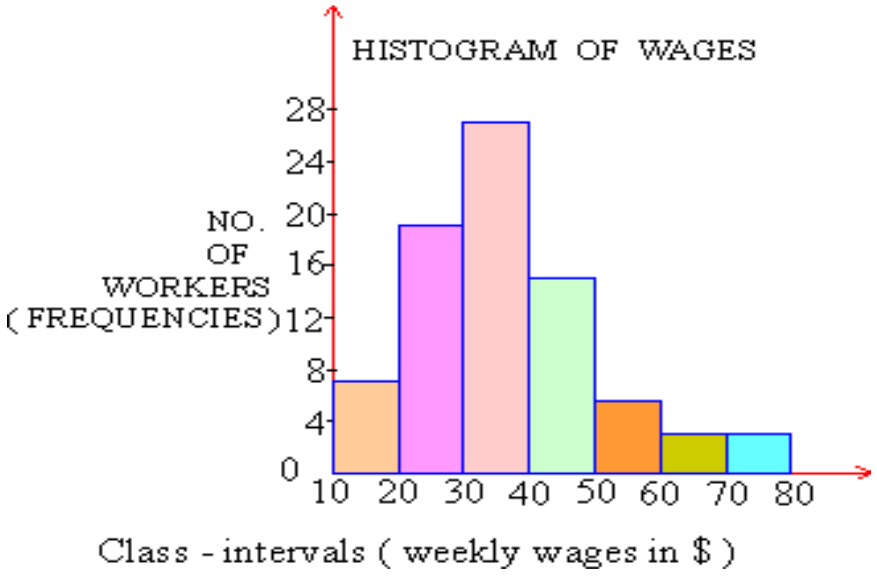
**Illustration:-** Represent the following data by means of Histogram.

| Classes: | 11-14 | 16-19 | 21-24 | 26-29 | 31-39 | 41-59 | 61-79 |
|---|---|---|---|---|---|---|---|
| Frequencies: | 7 | 19 | 27 | 15 | 12 | 12 | 8 |

**Solution:** Note that class-intervals are unequal and also they are of inclusive type. We have to make them equal and of the exclusive type.

Correct factor = ( 16 - 14 ) / 2 = 1. Using it we have

| Classes: | 10-15 | 15-20 | 20-25 | 25-30 | 30-40 | 40-60 | 60-80 |
|---|---|---|---|---|---|---|---|
| Frequencies: | 7 | 19 | 27 | 15 | 12 | 12 | 8 |
| Adjusted Heights: (Frequencies) | 7 | 19 | 27 | 15 | 12/2 = 6 | 12/4 = 3 | 12/4 = 3 |



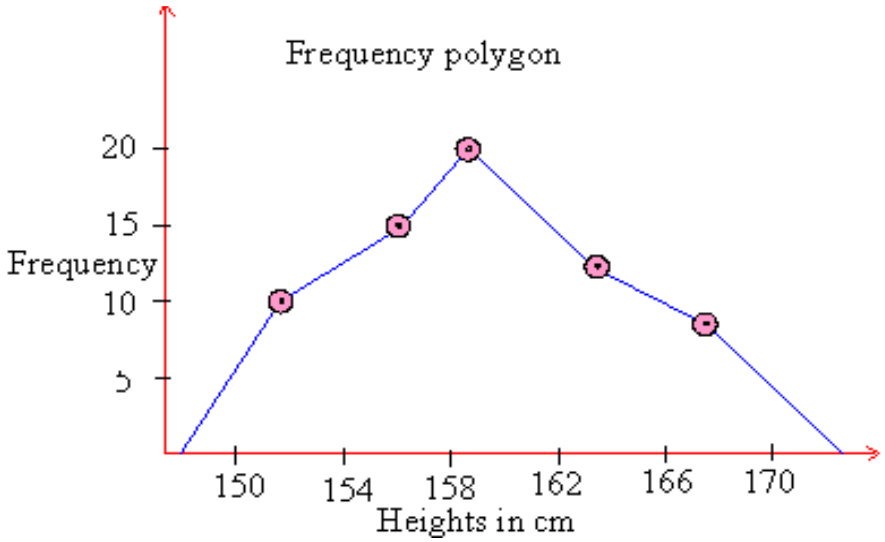
**2) Frequency Polygon:-**

Here the frequencies are plotted against the mid-points of the class-intervals and the points thus obtained are joined by line segments.

Example

| Height in cm:. | 150 - 154 | 154 - 158 | 158 - 162 | 162 - 166 | 166 - 170 |
|---|---|---|---|---|---|
| No. of children: | 10 | 15 | 20 | 12 | 8 |

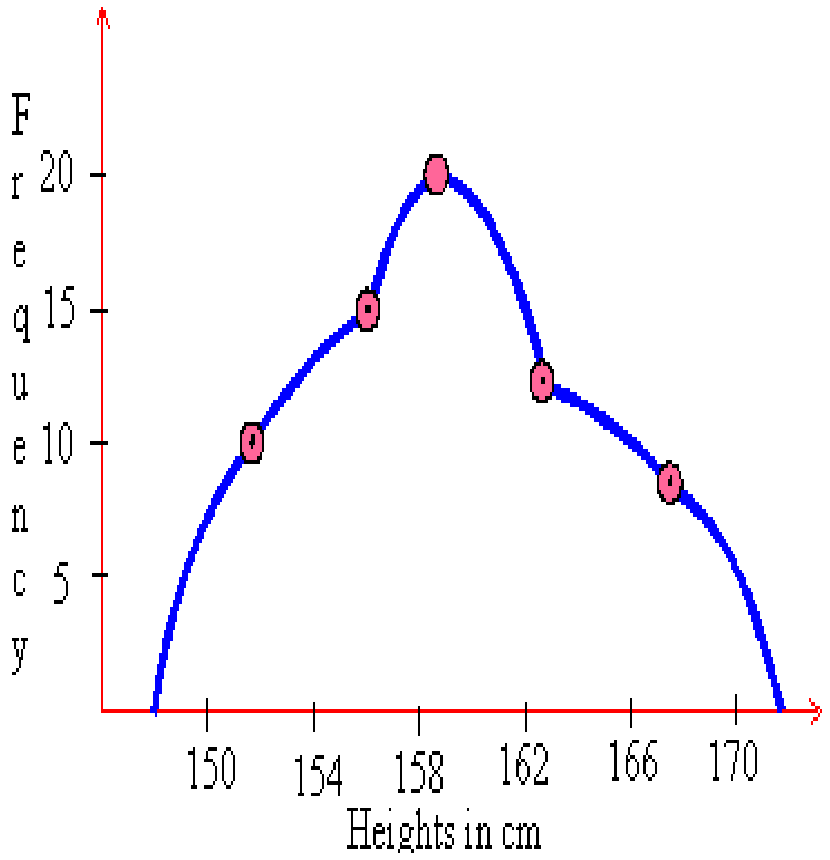
The polygon is closed at the base by extending it on both its sides ( ends ) to the midpoints of two hypothetical classes, at the extremes of the distribution, with zero frequencies.




On comparing the Histogram and a frequency polygon, you will notice that, in frequency polygons the points replace the bars ( rectangles ). Also, when several distributions are to be compared on the same graph paper, frequency polygons are better than Histograms.

### 3) Frequency Distribution (Curve):-

Frequency distribution curves are like frequency polygons. In frequency distribution, instead of using straight line segments, a smooth curve is used to connect the points. The frequency curve for the above data is shown as:




### 4) Ogives or Cumulative Frequency Curves:-

When frequencies are added, they are called cumulative frequencies. The curve obtained by plotting cumulating frequencies is called a cumulative frequency curve or an ogive ( pronounced ogive ).

To construct an Ogive:-

1) Add up the progressive totals of frequencies, class by class, to get the cumulative frequencies.

2) Plot classes on the horizontal ( x-axis ) and cumulative frequencies on the vertical ( y-axis).

3) Join the points by a smooth curve. Note that Ogives start at (i) zero on the vertical axis, and (ii) outside class limit of the last class. In most of the cases it looks like 'S'. Note that cumulative frequencies are plotted against the 'limits' of the classes to which they refer.

(A) Less than Ogive:- To plot a less than ogive, the data is arranged in ascending order of magnitude and the frequencies are cumulated starting from the top. It starts from zero on the y-axis and the lower limit of the lowest class interval on the x-axis.

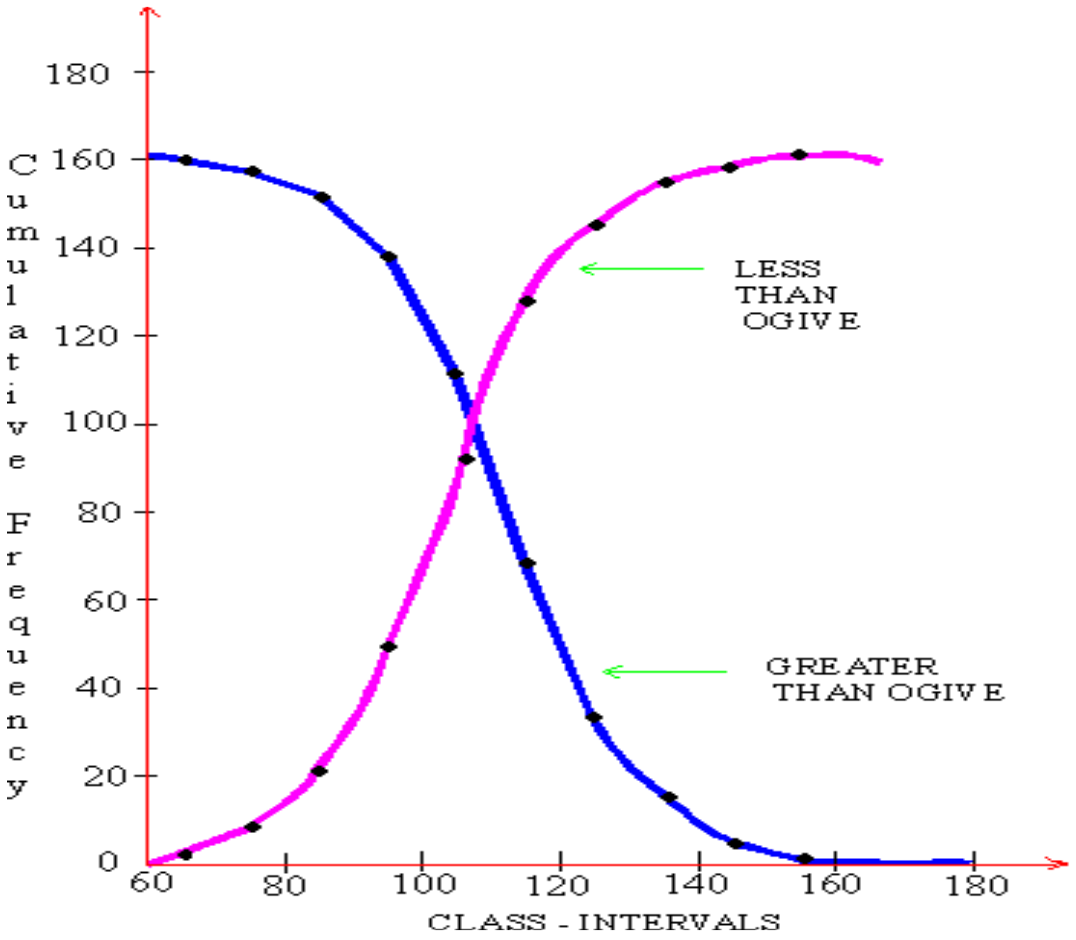
(B) Greater than Ogive:- To plot this ogive, the data are arranged in the ascending order of magnitude and frequencies are cumulated from the bottom. This curve ends at zero on the the y-axis and the upper limit of the highest class interval on the x-axis.

**Illustrations:-** On a graph paper, draw the two ogives for the data given below of the I.Q. of 160 students.

| Class -intervals: | 60 - 70 | 70 - 80 | 80 - 90 | 90 - 100 | 100 - 110 |
|---|---|---|---|---|---|
| No. of students: | 2 | 7 | 12 | 28 | 42 |

| 110 - 120 | 120 - 130 | 130 - 140 | 140 - 150 | 150 - 160 |
|---|---|---|---|---|
| 36 | 18 | 10 | 4 | 1 |

| Solution classes | f | c.f. less than | c.f. greater tha |
|---|---|---|---|
| 60 - 70 | 2 | 2 | 159+1 = 160 |
| 70 - 80 | 7 | 2+7 = 9 | 151+7 = 158 |
| 80 - 90 | 12 | 9+12 = 21 | 139+12 = 151 |
| 90 -100 | 28 | 21+28 = 49 | 111+28 = 139 |
| 100 -110 | 42 | 49+4 = 91 | 69+42 = 111 |
| 110 -120 | 36 | 91+36 = 127 | 33+36 = 69 |
| 120 -130 | 18 | 127+18 = 145 | 15+18 = 33 |
| 130 -140 | 10 | 145+10 = 155 | 5+10 = 15 |
| 140 -150 | 4 | 155+4 = 159 | 1+4 = 5 |
| 150 -160 | 1 | 159+1 = 160 | 1 |
| | Σf=160 | | |



**Uses:-**

Certain values like median, quartiles, deciles, quartile deviation, coefficient of skewness etc. can be located using ogives. it can be used to find the percentage of items having values less than or greater than certain value. Ogives are helpful in the comparison of the two distributions.

# MEASURES OF CENTRAL TENDENCY

## Introduction

In the previous chapter, we have studied how to collect raw data, its classification and tabulation in a useful form, which contributes in solving many problems of statistical concern. Yet, this is not sufficient, for in practical purposes, there is need for further condensation, particularly when we want to compare two or more different distributions. We may reduce the entire distribution to one number which represents the distribution.

A single value which can be considered as typical or representative of a set of observations and around which the observations can be considered as Centered is called an 'Average' (or average value) or a Center of location. Since such typical value tends to lie centrally within a set of observations when arranged according to magnitudes, averages are called measures of central tendency.

In fact the distribution have a typical value (average) about which, the observations are more or less symmetrically distributed. This is of great importance, both theoretically and practically. Dr. A.L. Bowley correctly stated, "Statistics may rightly be called the science of averages."

The fundamental measures of tendencies are:

(1) Arithmetic mean

(2) Median

(3) Mode

(4) Geometric mean

(5) Harmonic mean

However, the most common measures of central tendencies or Locations are Arithmetic mean, median and mode. We therefore, consider the Arithmetic mean

**ARITHMETIC MEAN: (A.M)**

This is the most commonly used average which you have also studied and used in lower grades. Here are two definitions given by two great masters of statistics.

**Horace Sacrist :** Arithmetic mean is the amount secured by dividing the sum of values of the items in a series by their number.

**W.I. King :** The arithmetic average may be defined as the sum of aggregate of a series of items divided by their number.

Thus, the students should add all observations (values of all items) together and divide this sum by the number of observations (or items).

**Ungrouped Data**

Suppose, we have 'n' observations (or measures) $x_1$ , $x_2$ , $x_3$, ....., $x_n$ then the Arithmetic mean is obviously $\frac{x_1 + x_2 + x_3 + \ldots\ldots + x_n}{n}$

We shall use the symbol x (pronounced as x bar) to denote the Arithmetic mean. Since we have to write the sum of observations very frequently, we use the usual symbol ' □ ' (pronounced as sigma) to denote the sum. The symbol $x_i$ will be used to denote, in general the 'i' th observation. Then the sum, $x_1 + x_2 + x_3 + \ldots\ldots + x_n$ will be represented by $\sum_{i=1}^{n} x_i$ or $\sum x_i$ simply

Therefore the Arithmetic mean of the set $x_1 + x_2 + x_3 + .....+ x_n$ is given by,

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + x_3 + .... + x_n}{n}$$

This method is known as the "Direct Method".

**Example** A variable takes the values as given below. Calculate the arithmetic mean of 110, 117, 129, 195, 95, 100, 100, 175, 250 and 750.

**Solution:** Arithmetic mean = $\frac{\sum x_i}{n}$

$\sum x_i = 110 + 117 + 129 + 195 + 95 + 100 + 100 + 175 + 250 + 750 = 2021$

and n = 10

**Indirect Method (Assumed Mean Method)**

$\bar{u} = \frac{\sum u_i}{n}$ where $u = x_i - A$

A = Assumed Mean = $A + \bar{u}$

**Calculations:**

Let A = 175 then

- $u_i$ = -65, -58, -46, +20, -80, -75,-75, +0, + 75, +575

  = 670 - 399

  = 271/10 = 27.1

- $\bar{x} = A + \bar{u}$

= 175 + 27.1

= 202.1

**Example** M.N. Elhance's earnings for the past week were:

| | |
|---|---|
| Monday | $ 450 |
| Tuesday | $ 375 |
| Wednesday | $ 500 |
| Thursday | $ 350 |
| Friday | $ 270 |

Find his average earning per day.

**Solution:**

$$\sum x_i = 450 + 375 + 500 + 350 + 270 = \$1945$$

n = 5

Arithmetic mean = $\frac{\sum x_i}{n} = \frac{1945}{5} = \$389$

Therefore, Elhance's average earning per day is $389.

**Short-cut Method :**

Sometimes the values of x are very big and in that case, to simplify the calculation the short-cut method is used. For this, first you assume a mean (called as the assumed mean). Let it be A. Now find the deviations of all the values of x from A. We now get a new variable $u_i = x_i - A$

Now find

$\bar{u} = \frac{\sum u_i}{n}$ then $\bar{x} = A + \frac{\sum u_i}{n}$ or $\bar{x} = A + \bar{u}$

**Example** The expenditure of ten families in dollars are given below :

Family : A B C D E F G H I J

Expenditure : 300 700 100 750 500 80 120 250 100 370
(in dollars).

Calculate the Arithmetic mean.

**Solution:** Let the assumed mean be $ 500. (as. = assume)

| **Families** | **Expenditure ($)** | **Deviation from as. mean** |
|---|---|---|
| $(x_i)$ | | $(u_i = x_i - A)$ |
| A | 300 | -200 |
| B | 700 | 200 |
| C | 100 | - 400 |
| D | 750 | 250 |
| E | 500 | 0 |
| F | 80 | -420 |
| G | 120 | -380 |
| H | 250 | -250 |
| I | 100 | -400 |
| J | 370 | -130 |
| $n = 10$ | | $\Sigma u_i = - 2180 + 450 = -1730$ |

**Calculations** :

$$\bar{u} = \frac{\Sigma u_i}{n} = \frac{-1730}{10} = -173$$

$$\bar{x} = A + \bar{u}$$

$$\therefore \bar{x} = 500 + (-173) = 327$$

**Discrete Series:** There is a difference in the methods for finding the arithmetic means of the individual series and a discrete series. In the discrete series, every term (i.e. value of x) is multiplied by its corresponding frequency ($f_ix_i$) and then their total (sum) is found $\left(\sum f_i x_i\right)$. The arithmetic mean is then obtained by dividing the total frequency $\left(\sum f\right)$ by the above sum so obtained $\left(\sum f_i x_i\right)$.

Therefore, if the observations $x_1 + x_2 + x_3 + .....+ x_n$ are repeated $f_1 + f_2 + f_3 + .....+ f_n$ times, then we have:

Arithmetic mean $\left(\bar{x}\right) = \frac{\sum f_i x_i}{\sum f_i}$ i.e. $\frac{f_1 x_1 + f_2 x_2 + f_3 x_3 + .....+ f_n x_n}{f_1 + f_2 + f_3 + .....+ f_n}$

The formulae for Arithmetic mean by direct method and by the short-cut methods are as follows:

**Short-cut method**

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} \qquad \bar{x} = A + \bar{u} \text{ where } \bar{u} = \frac{\sum f_i u_i}{\sum f_i}$$

and $u = x_i - A$

Therefore, $\bar{x} = A + \frac{\sum f_i u_i}{\sum f_i}$

**Example** Find the mean of the following 50 observations.

19, 19, 20, 20, 20, 19, 20, 18, 21, 19,

20, 20, 19, 19, 20, 19, 21, 19, 19, 21,

18, 20, 18, 18, 17, 20, 20, 22, 20, 20,

20, 20, 20, 21, 20, 17, 23, 18, 17, 21,

20, 21, 20, 20, 20, 18, 21, 19, 20, 19

**Solution:** We may tabulate the given observations as follows.

| Observations ($x_i$) | Frequency ($f_i$) | $fx_i$ |
|---|---|---|
| 17 | 3 | $17 \times 3 = 51$ |
| 18 | 6 | $18 \times 6 = 108$ |
| 19 | 11 | $19 \times 11 = 209$ |
| 20 | 20 | $20 \times 20 = 400$ |
| 21 | 8 | $21 \times 8 = 168$ |
| 22 | 1 | $22 \times 1 = 22$ |
| 23 | 1 | $23 \times 1 = 23$ |
| Total | $\sum f_i = 50$ | $\sum f_i x_i = 981$ |

The arithmetic mean is $\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{981}{50} = 19.62$

**Example** Eight coins were tossed together and the number of times they fell on the side of heads was observed. The activity was performed 256 times and the frequency obtained for different values of x, (the number of times it fell on heads) is shown in the following table. Calculate then mean by:

i) Direct method ii) Short-cut method

x : 0 1 2 3 4 5 6 7 8

f : 1 9 26 59 72 52 29 7 1

**Solution:**

| No. of coins | Frequency | Direct method | Short-cut method Dev. from A (4) i.e. $u_i = x_i - A$ | |
|---|---|---|---|---|
| $x_i$ | $f_i$ | $f_i x_i$ | $u_i$ | $\Sigma f_i u_i$ |
| 0 | 1 | 0 | -4 | -4 |
| 1 | 9 | 9 | -3 | -27 |
| 2 | 26 | 52 | -2 | -52 |
| 3 | 59 | 117 | -1 | -59 |
| 4 | 72 | 288 | 0 | 0 |
| 5 | 52 | 260 | 1 | +52 |
| 6 | 29 | 174 | 2 | +58 |
| 7 | 7 | 49 | 3 | +21 |
| 8 | 1 | 8 | 4 | +4 |
| | $\Sigma f = 256$ | $\Sigma fx = 1017$ | | +135 - 142 $\Sigma fu = -7$ |

**Mean for Grouped data**

**Continuous series:** The procedure of finding the arithmetic mean in this series, is the same as we have used in the discrete series. The only difference is that in this series, we are given class-intervals, whose mid-values (class-marks) are to be calculated first.

Formula, Arithmetic mean $(\bar{x}) = \frac{\Sigma f_i x_i}{\Sigma f_i}$

where x = mid-value

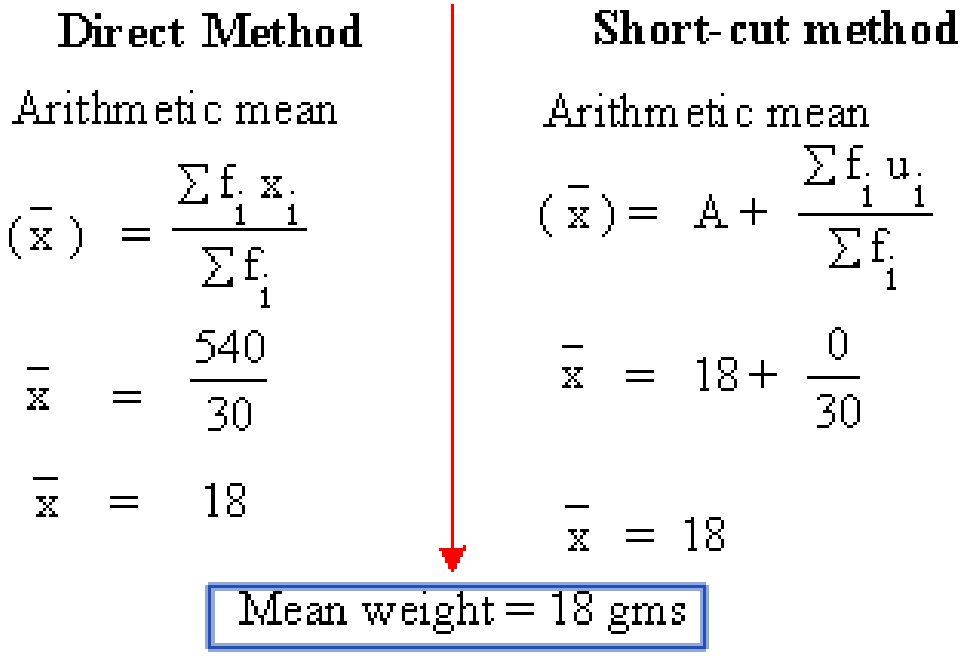
**Example** The weights (in gms) of 30 articles are given below :

14, 16, 16, 14, 22, 13, 15, 24, 23, 14, 20, 17, 21, 18, 18, 19, 20, 17, 16, 15, 11, 22, 21, 20, 17, 18, 19, 22, 23.

Form a grouped frequency table, by dividing the variate range into intervals of equal width, one class being 11-13 and then compute the arithmetic mean.

**Solution:**

| Weights | mid-values $x_i$ | frequency $f_i$ | Direct method $f_i x_i$ | Short-cut method A = 18 | |
|---|---|---|---|---|---|
| | | | | $u_i = x_i - A$ | $f_i u_i$ |
| 11 - 13 | 12 | 3 | 36 | -6 | -18 |
| 13 - 15 | 14 | 4 | 56 | -4 | -16 |
| 15 - 17 | 16 | 5 | 80 | -2 | -20 |
| 17 - 19 | 18 | 6 | 108 | 0 | 0 |
| 19 - 21 | 20 | 5 | 100 | +2 | +10 |
| 21 - 23 | 22 | 4 | 88 | +4 | +16 |
| 23 - 25 | 24 | 3 | 72 | +6 | +18 |
| Total | | $\Sigma f_i = 30$ | $\Sigma f_i x_i = 540$ | | $\Sigma f_i u_i = 0$ |




**Example** Find the arithmetic mean for the following :

Marks below : 10 20 30 40 50 60 70 80

No. of students : 15 35 60 84 96 127 198 250

**Solution:**

First, we have to convert the cumulative frequencies into frequencies of the respective classes.

| Marks | Mid-values $x_i$ | Frequencies | c.f. f. | U = X - A, A = 45 | $f_iu_i$ |
|---|---|---|---|---|---|
| 0 - 10 | 5 | 15 | 15 | - 40 | - 600 |
| 10 -20 | 15 | 35 | 20 | - 30 | - 600 |
| 20 - 30 | 25 | 60 | 25 | - 20 | - 500 |
| 30 - 40 | 35 | 84 | 24 | - 10 | - 240 |
| 40 - 50 | 45 □ A | 96 | 12 | 0 | 0 |
| 50 - 60 | 55 | 127 | 31 | +10 | +310 |
| 60 - 70 | 65 | 198 | 71 | +20 | +1420 |
| 70 - 80 | 75 | 250 | 52 | +30 | +1560 |
| Total | | $\sum f = 250$ | | | $\sum f_u = 1350$ |

$$\text{Arithmetic Mean } (\bar{x}) = A + \bar{u}$$

$$= A + \frac{\sum f_i u_i}{\sum f_i}$$

$$= 45 + \frac{1350}{250}$$

$$= 45 + 5.4$$

$$= 50.4 \text{ marks}$$

**Step-Deviation Method**

Here all class intervals are of the same width say 'c'. This method is employed in place of the Short-cut method. We measure all the class-marks (mid values) from some convenient value, say 'A', which generally should be taken as the class-mark of a class of maximum frequency or of a class which is the middle one. All the class marks happen to be multiples of c, since all class intervals are equal. We consider class frequencies as if they are centered at the corresponding class-marks.

**Theorem** If $x_1, x_2, x_3, \ldots\ldots, x_n$ are n values of the class marks with frequencies $f_1, f_2, f_3, \ldots\ldots f_n$ respectively and if each xi is expressed in terms of the new variable ui by the relation $x_i = A + cui$ then, with the usual notation, we have $\bar{x} = A + c\bar{u}$

where $\bar{u} = \frac{\sum f u_i}{\sum f_i}$ and $u_i = \frac{x_i - A}{c}$

This method is also known as the "Coding method."

**Example** Calculate the arithmetic mean from the following data :

Age (years) below : 25 30 35 40 45 50 55 60

No. of employees : 8 23 51 81 103 113 117 120

**Solution :**

| Age (years) | Mid - values $x_i$ | Frequencies | | $u_i = \frac{x_i - A}{c}$ | $fu_i$ |
|---|---|---|---|---|---|
| | | c.f | $f_i$ | | |
| 20 - 25 | 22.5 | 8 | 8 | - 4 | - 32 |
| 25 - 30 | 27.5 | 23 | 15 | - 3 | - 45 |
| 30 - 35 | 32.5 | 51 | 28 | - 2 | - 59 |
| 35 - 40 | 37.5 | 81 | 30 | - 1 | - 30 |
| 40 - 45 | 42.5 $\Rightarrow$ A | 103 | 22 | 0 | 0 |
| 45 - 50 | 47.5 | 113 | 10 | +1 | + 10 |
| 50 - 55 | 52.5 | 117 | 4 | +2 | + 8 |
| 55 - 60 | 57.5 | 120 | 3 | +3 | + 9 |
| Total | | | $\sum f = 120$ | | $\sum fu_i = -136$ |

$$\text{Arithmetic mean } (\bar{x}) = A + c\bar{u}$$

$$\text{where } \quad \bar{u} = \frac{\sum f u_i}{\sum f_i} = \frac{-136}{120} \quad \text{and } c = 5$$

$$\therefore \quad \bar{x} = 42.5 + \left(\frac{-136}{120}\right) \times 5$$

$$\therefore \quad \bar{x} = 42.5 - 5.67$$

$$\therefore \quad \bar{x} = 36.83 \text{ years}$$

**Properties of Arithmetic Mean**

1. The sum of the deviations, of all the values of x, from their arithmetic mean, is zero.

   Justification : $\sum f_i (x_i - \bar{x}) = \sum f_i x_i - \bar{x} \sum f_i = 0$

   Since $\bar{x}$ is a constant, $\bar{x} = \frac{\sum f_i x_i}{\sum f_i} \therefore \sum f_i x_i = \bar{x} \sum f_i$

2. The product of the arithmetic mean and the number of items gives the total of all items.

   Justification : $\bar{x} = \frac{\sum f_i x_i}{\sum f_i} \Rightarrow \sum f_i x_i = \bar{x} \sum f_i$

   or $\bar{x} = \frac{\sum x_i}{N} \Rightarrow \bar{x} \cdot N = \sum x_i$

3. If $\bar{x}_1$ and $\bar{x}_2$ are the arithmetic mean of two samples of sizes $n_1$ and $n_2$ respectively then, the arithmetic mean $\bar{x}$ of the distribution combining the two can be calculated as

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

This formula can be extended for still more groups or samples.

$$\bar{x}_1 = \frac{\sum x_{1i}}{n_1} \Rightarrow \sum x_{1i} = n_1 \bar{x}_1$$

Justification : $\bar{x}_1 = \frac{\sum x_{1i}}{n_1} \Rightarrow \sum x_{1i} = n_1 \bar{x}_1$ = total of the observations of the first sample

Similarly $\sum x_{2i} = n_2 \bar{x}_2$ = total of the observations of the first sample

The combined mean of the two samples

$$= \frac{\text{combined total}}{n_1 + n_2}$$

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

**Example** The average marks of three batches of students having 70, 50 and 30 students respectively are 50, 55 and 45. Find the average marks of all the 150 students, taken together.

**Solution :**

Let x be the average marks of all 150 students taken together.

| | Batch - I | Batch - II | Batch - III |
|---|---|---|---|
| | $\bar{x}_1$ | $\bar{x}_2$ | $\bar{x}_3$ |
| A. marks : | = 50 | = 55 | = 45 |
| No. of students | $n_1 = 70$ | $n_2 = 50$ | $n_3 = 30$ |

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + n_3 \bar{x}_3}{n_1 + n_2 + n_3} = \frac{70 \times 50 + 50 \times 55 + 30 \times 45}{70 + 50 + 30}$$

$$= \frac{7600}{150}$$

$\bar{x} = 50.67$ marks

**Merits**

1. It is rigidly defined. Its value is always definite.
2. It is easy to calculate and easy to understand. Hence it is very popular.
3. It is based on all the observations; so that it becomes a good representative.
4. It can be easily used for comparison.
5. It is capable of further algebraic treatment such as finding the sum of the values of the observations, if the mean and the total number of the observations are given; finding the combined arithmetic mean when different groups are given etc.
6. It is not affected much by sampling fluctuations.

**Demerits**

1. It is affected by outliers or extreme values. For example, the average (A.) mean of 10, 15, 25 and 500 is $\bar{x} = \frac{10+15+25+500}{4} = 137.5$

   Now observe first three values whose A. mean is $\frac{10+15+25}{3} = 16.67$ (approx.)

   Due to the outlier 500 the A. mean of the four numbers is raised to 137.5. In such a case A. mean is not a good representative of the given data.

2. It is a value which may not be present in the given data.

3. Many a times it gives absurd results like 4.4 children per family.

4. It is not possible to take out the averages of ratios and percentages.

5. We cannot calculate it when open-end class intervals are present in the data.

**MEDIAN: (M)**

It is the value of the size of the central item of the arranged data (data arranged in the ascending or the descending order). Thus, it is the value of the middle item and divides the series in to equal parts.

In Connor's words - "The median is that value of the variable which divides the group into two equal parts, one part comprising all values greater and the other all values lesser than the median." For example, the daily wages of 7 workers are 5, 7, 9, 11, 12, 14 and 15 dollars. This series contains 7 terms. The fourth term i.e. $11 is the median.

**Median in Individual Series (ungrouped Data)**

1. Set the individual series either in the ascending (increasing) or in the descending (decreasing) order, of the size of its items or observations.

2. If the total number of observations be 'n' then

   A. If 'n' is odd,

   The median = size of $\left(\frac{n+1}{2}\right)^{th}$ observation

   B. If 'n' is even, the median

   $$= \frac{1}{2}\left[\begin{array}{l}\text{size of } \left(\frac{n}{2}\right)^{th} \text{ observations} \\ + \text{size of } \left(\frac{n+2}{2}\right)^{th} \text{ observations}\end{array}\right]$$

**Example** The following figures represent the number of books issued at the counter of a Statistics library on 11 different days. 96, 180, 98, 75, 270, 80, 102, 100, 94, 75 and 200. Calculate the median.

**Solution:**

Arrange the data in the ascending order as 75, 75, 80, 94, 96, 98, 100, 102,180, 200, 270.

Now the total number of items 'n'= 11 (odd)

Therefore, the median = size of $\left(\frac{n+1}{2}\right)^{th}$ item

$$= \text{size of } \left(\frac{11+1}{2}\right)^{th} \text{ item}$$

$$= \text{size of } 5^{th} \text{ item}$$

$$= 98 \text{ books per day}$$

**Example** The population (in thousands) of 36 metropolitan cities are as follows :2468, 591, 437, 20, 213, 143, 1490, 407, 284, 176, 263, 19, 181, 777, 387, 302, 213, 204, 153, 733, 391, 176 178, 122, 532, 360, 65, 260, 193, 92, 672, 258, 239, 160, 147, 151. Calculate the median.

**Solution:**

Arranging the terms in the ascending order as :

20, 65, 92, 131, 142, 143, 147, 151, 153, 160, 169, 176, 178, 181, 193, 204, (213, 39), 258, 263, 260, 384, 302, 360, 387, 391, 407, 437, 522, 591, 672, 733, 777, 1490, 2488.

Since total number of items n = 36 (Even).

the median

$$= \frac{1}{2}\left[\text{size of } \left(\frac{n}{2}\right)^{th} \text{ item} + \text{size of} \left(\frac{n+2}{2}\right)^{th} \text{ item}\right]$$

$$= \frac{1}{2}\left[\text{size of } 18^{th} \text{ item} + \text{size of } 19^{th} \text{ item}\right]$$

$$= \frac{1}{2}[213+239]$$

$$= \frac{1}{2}[252]$$

$$= 226 \text{ thousands}$$

**Median in Discrete Series**

Steps:

1. Arrange the data in ascending or descending order of magnitude.

2. Find the cumulative frequencies.

3. Apply the formula:

   A. If 'n' = $\Sigma f_i$ (odd) then,

   $$\text{Median} = \text{size of } \left(\frac{n+1}{2}\right)^{th} \text{ item}$$

   B. If 'n' = $\Sigma f_i$ (even) then,

   $$\text{Median} = \frac{1}{2}\left[\begin{array}{l}\text{size of } (n/2)^{th} \text{ item} \\ + \text{ size of} \left(\frac{n+2}{2}\right)^{th} \text{ item}\end{array}\right]$$

**Example** Locate the median in the following distribution.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Size | : | 8 | 10 | 12 | 14 | 16 | 18 | 20 |
| Frequency: | | 7 | 7 | 12 | 28 | 10 | 9 | 6 |

**Solution:**

| Size $(x_i)$ | Frequency $f_i$ | Cumulative frequency (c.f) |
|---|---|---|
| 8 | 3 | 3 |
| 10 | 7 | $3+7=10$ |
| 12 | 12 | $12+10=22$ |
| 14 | 28 | $28+22=50$ |
| 16 | 10 | $50+10=60$ |
| 18 | 9 | $60+9=69$ |
| 20 | 6 | $69+6=75$ |
| | $n=\Sigma f_i=75\,(\text{odd})$ | |

Therefore, the median = $\text{size of } \left(\frac{n+1}{2}\right)^{th} \text{ item}$

$$= \text{size of } \left(\frac{75+1}{2}\right)^{th} \text{ item}$$

$$= \text{size of } 38^{th} \text{ item}$$

In the order of the cumulative frequency, the 38th term is present in the 50th cumulative frequency, whose size is 14.

Therefore, the median = 14

**Median in Continuous Series (grouped Data)**

Steps:

1. Determine the particular class in which the value of the median lies.

   Use $n/2$ as the rank of the median and not $\left(\frac{n+1}{2}\right)$

2. After ascertaining the class in which median lies, the following formula is used for determining the exact value of the median.

   Median = $\ell_1 + \left[\frac{N/2 - c.f}{f}\right](\ell_2 - \ell_1)$

   Where, $\ell_1$ = lower limit of the median class, the class in which the middle item of the distribution lies.

   $\ell_2$ = upper limit of the median class

   c.f = cumulative frequency of the class preceding the median class

   f = sample frequency of the median class

   It should be noted that while interpolating the median value of frequency distribution it is assumed that the variable is continuous and that there is an orderly and even distribution of items within each class.

**Example** Calculate the median for the following and verify it graphically.

| Age (years) | : | 20-25 | 25-30 | 30-35 | 35-40 | 40-45 |
|---|---|---|---|---|---|---|
| No. of person | : | 70 | 80 | 180 | 150 | 20 |

**Solution:**

| Age (years)<br>C.I. | No. of persons<br>$f_i$ | Cumulative frequency<br>c.f. |
|---|---|---|
| 20 - 25 | 70 | 70 |
| 25 - 30 | 80 | 70 + 80 = 150 |
| 30 - 35 | 180 | 150 + 180 = 330 |
| 35 - 40 | 150 | 330 + 150 = 480 |
| 40 - 45 | 20 | 480 + 20 = 500 |
| | $n = \Sigma f_i = 500$ | |

$$\text{Now median} = \text{size of } (n/2)^{th} \text{ item}$$

$$= \text{size of } \left(\frac{500}{2}\right)^{th}$$

$$= \text{size of } 250^{th} \text{ item which lies in } (30 - 35) \text{ class interval}$$

$$\text{Median} = \ell_1 + \left[\frac{N/2 - c.f}{f}\right](\ell_2 - \ell_1)$$

Here $\ell_1 = 30$, $\ell_2 = 35$, $n/2 = 250$, c.f. = 150 and f = 180

Therefore, Median

$$= 30 + \left[\frac{250 - 150}{180}\right](35 - 30)$$

$$= 30 + \frac{100}{180} \times 5$$

$$= 32.78 \text{ years}$$

Sometimes the series is given in the descending order of magnitude. In this situation convert the series in the ascending order of magnitude and then using the regular formula, the median can be calculated or the series can be put in the descending order of the magnitude and an alternative formula be used to calculate the median.

**Example** Marks: 40 -50 30- 40 20-30 10-20 0 -10

No. of students: 10 12 40 30 8

**Solution:**

| Marks | No. of students | c.f. |
|---|---|---|
| 0 -10 | 8 | 8 |
| 10 - 20 | 30 | 8 + 30 = 38 |
| 20 - 30 | 40 | 38 + 40 = 78 |
| 30 - 40 | 12 | 78 + 12 = 90 |
| 40 - 50 | 10 | 90 + 10 = 100 |
| | $n = \sum f = 100$ | |

$$\text{Median} = \text{size of } (n/2)^{\text{th}} \text{ item}$$

$$= \text{size of } \left(\frac{100}{2}\right)^{\text{th}} \text{ item}$$

$$= \text{size of } 50^{\text{th}} \text{ item which comes into class interval } (20-30)$$

By interpolation

$$\text{Median} = \ell_1 + \left[\frac{n/2 - c.f}{f}\right](\ell_2 - \ell_1)$$

$$= 20 + \left[\frac{50-38}{40}\right](30-20)$$

$$= 20 + \frac{12}{40} \times 10$$

$$= 20 + 3$$

$$\therefore \quad \text{Median} = 23 \text{ marks}$$

Note that, while calculating the median of a series, it must be put in the 'exclusive class-interval' form. If the original series is in inclusive type, first convert it into the exclusive type and then find its median.

**Example** The following distribution represents the number of minutes spent by a group of teenagers in watching movies. What is the median ?

Minutes/Weeks:0-99 100-199 200-299 300-399 400 - 499 500 - 599 600 & more

No. of teenagers : 27 32 65 78 58 32 8

**Solution:**

| Minutes/ weeks | Real class intervals | Frequency | Cumulative frequency |
|---|---|---|---|
| 0 -99 | 0.5 - 99.5 | 27 | 27 |
| 100 - 199 | 99.5 - 199.5 | 32 | 27 + 32 = 59 |
| 200 - 299 | 199.5 - 299.5 | 65 | 59 + 65 = 124 |
| 300 - 399 | 299.5 - 399.5 | 78 | 124 + 78 = 202 |
| 400 - 499 | 399.5 - 499.5 | 58 | 202 + 58 = 260 |
| 500 - 599 | 499.5 - 599.5 | 32 | 260 + 32 = 292 |
| 600 &more | 599.5 & more | 82 | 92 + 8 = 300 |
| | | $n = \sum f_i = 300$ | |

$$\text{Median} = \text{size of } (n/2)^{th} \text{ item}$$
$$= \text{size of } \left(\frac{300}{2}\right)^{th} \text{ item}$$
$$= \text{size of } 150^{th} \text{ item which}$$
$$\text{lies in } (299.5 - 399.5) \text{ class interval}$$

By using interpolation

$$\text{Median} = \ell_1 - \left[\frac{n/2 - \text{c.f}}{f}\right](\ell_2 - \ell_1)$$
$$= 299.5 + \left[\frac{150 - 124}{78}\right](399.5 - 299.5)$$
$$= 299.5 + \left[\frac{26}{78}\right] \times 100$$
$$= 299.5 + 33.33 \text{ (approximately)}$$
$$\text{Median} = 392.88 \text{ minutes/ week}$$

**Merits of Median**

1. It is rigidly defined.

2. It is easy to calculate and understand.

3. It is not affected by extreme values like the arithmetic mean. For example, 5 persons have their incomes $2000, $2500, $2600, $3000, $5000. The median would be $2600 while the arithmetic mean would be $3020.

4. It can be found by mere inspection.

5. It is fully representative and can be computed easily.

6. It can be used for qualitative studies.

7. Even if the extreme values are unknown, median can be calculated if one knows the number of items.

8. It can be obtained graphically.

**Demerits of Median**

1. It may not be representative if the distribution is irregular and abnormal.

2. It is not capable of further algebraic treatment.

3. It is not based on all observations.

4. It is affected by sample fluctuations.

5. The arrangement of the data in the order of magnitude is absolutely necessary.

**MODE: (Z)**

It is the size of that item which possesses the maximum frequency. According to Professor Kenney and Keeping, the value of the variable which occurs most frequently in a distribution is called the mode.

It is the most common value. It is the point of maximum density.

**Ungrouped Data**

**Individual series:** The mode of this series can be obtained by mere inspection. The number which occurs most often is the mode.

**Example** Locate mode in the data 7, 12, 8, 5, 9, 6, 10, 9, 4, 9, 9

**Solution:** On inspection, it is observed that the number 9 has maximum frequency. Therefore 9 is the mode.

Note that if in any series, two or more numbers have the maximum frequency, and then the mode will be difficult to calculate. Such series are called as Bi-modal, Tri-modal or Multi-modal series.

**Grouped Data**

Steps:

1. Determine the modal class which as the maximum frequency.
2. By interpolation the value of the mode can be calculated as -

$$\text{Mode} = \ell_1 - \left[\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right](\ell_2 - \ell_1)$$

Where

$\ell_1$ = lower limit of the modal class

$\ell_2$ = upper limit of the modal class

$f_1$ = frequency of the modal class

$f_0$ = frequency of the class preceding to the modal class

$f_2$ = frequency of the class succeding the modal class

**Example** Calculate the modal wages.

Daily wages in $: 20 -25 25-30 30-35 35-40 40-45 45-50

No. of workers: 1 3 8 12 7 5

Verify it graphically.

**Solution:**

Here the maximum frequency is 12, corresponding to the class interval (35 - 40) which is the modal class.

Therefore $\ell_1 = 35,\ \ell_2 = 40,\ f_0 = 8,\ f_1 = 12,\ f_2 = 7$

By interpolation

Mode = $$\ell_1 - \left[\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right](\ell_2 - \ell_1)$$

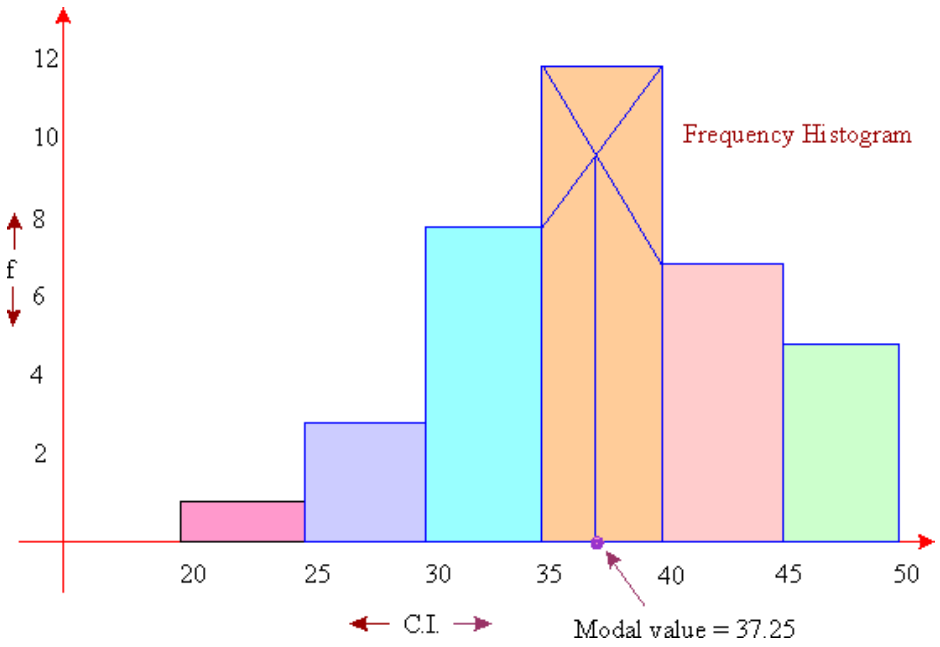
$$= 40 + \left[\frac{12-8}{24-8-7}\right](40-35)$$

$$= 40 + \left[\frac{4}{9}\right] \times 5$$

$$= 40 + 2.22$$

$$= 37.22$$

Modal wages is $37.22




**MERITS OF MODE**

1. It is simple to calculate.

2. In individual or discrete distribution it can be located by mere inspection.

3. It is easy to understand. Everyone is used to the idea of average size of a garment, an average American etc.

4. It is not isolated like the median as it is the most common item.

5. Like the Average mean, it is not a value which cannot be found in the series.

6. It is not necessary to know all the items. What we need the point of maximum density frequency.

7. It is not affected by sampling fluctuations.

**DEMERITS**

1. It is ill defined.

2. It is not based on all observations.

3. It is not capable of further algebraic treatment.

4. It is not a good representative of the data.

5. Sometimes there are more than one values of mode.

**Geometric Mean (G.M):**

Geometric mean is defined as the N th root of the product of N items. If there are two items, we take the square root; if three, the cube root and so on. The geometric mean is never larger than the arithmetic mean; on occasion it may turn out to be the same as the arithmetic mean, but usually it is smaller. If there are zeros or negative values in the series, the geometric mean cannot be used. Thus, geometric mean is obtained by multiplying together all the values of the series and then calculating the root of their product corresponding to the number of items in the group. To solve a question to find out the geometric mean, help is taken from logarithms so as to save the time and labour. Therefore, geometric mean is the antilog of the arithmetic average of the logarithms of different items.

$$\text{Geometric Mean} = \text{Antilog of} \quad \frac{\text{Log } X_1 + \log X_2 + \log X_3 \ldots . \text{Log } X_n}{N}$$

Or

$$\text{Geometric Mean} = \text{antilog of} \quad \frac{\log X}{\mathbf{N}}$$

**Calculation of Geometric Mean – Individual Series:**

**Steps:**

1. Find out the logarithm of each value or the size of the item from the log table- log X
2. Add all the values of log X - $\sum$ log X
3. The sum of log ($\sum$ log X) is divided by the number of items. $\frac{\sum \log X}{N}$
4. Find out the antilog of the quotient (from step 3). This is the geometric mean of the data.

**Example:**

Calculate geometric mean from the following:

50 72 54 82 93

**Solution:**

| X | Log of X |
|---|---|
| 50 | 1.6990 |
| 72 | 1.8573 |
| 54 | 1.7324 |
| 82 | 1.9138 |
| 93 | 1.9685 |
| | $\sum$log X= 9.1710 |

$$\text{Geometric Mean} = \text{antilog of } \frac{\sum \log X}{N}$$

$$= \text{Antilog of } \frac{9.1710}{5}$$

$$= \text{Antilog } 1.8342 = 68.26$$

**Calculation of Geometric Mean – Discrete Series:**

**Steps:**

1. Find out the logarithm of each value- log X
2. Multiply the log of each size by its frequency- f log X
3. Add all the products thus we get ∑f log X
4. Divide the total of products by the total frequency (N) $\frac{\sum f \log X}{N}$
5. The antilog of the step 4 is the result.

$$G.M. = \text{Antilog} \frac{\sum f \log X}{N}$$

**Example:**

The following table gives the weight of 31 persons in sample survey. Calculate geometric mean.

| Weight(lbs) | 130 | 135 | 140 | 145 | 146 | 148 | 149 | 157 |
|---|---|---|---|---|---|---|---|---|
| No.of persons | 3 | 4 | 6 | 613 | 5 | 2 | 1 | 1 |

**Solution:**

| Size of item X | Frequency | Log X | f log X |
|---|---|---|---|
| 130 | 3 | 2.1139 | 6.3417 |
| 135 | 4 | 2.1303 | 8.5212 |
| 140 | 6 | 2.1461 | 12.5766 |
| 145 | 6 | 2.1614 | 12.9684 |
| 146 | 3 | 2.1644 | 6.4932 |
| 148 | 5 | 2.1703 | 10.8515 |
| 149 | 2 | 2.1732 | 4.3464 |
| 150 | 1 | 2.1761 | 2.1761 |
| 157 | 1 | 2.1959 | 2.1959 |
| | N = 31 | | ∑f log X = 66.7710 |

$$\text{G.M.} = \text{Antilog} \frac{\sum f \log X}{N}$$

$$\frac{66.7710}{31} = \text{Antilog of } 2.1539$$

G.M. weight = 142.5 lbs

**Calculation of Geometric Mean – Continuous Series:**

**Steps:**

1. Find out the mid value of each class-m
2. Find the logarithm of the mid value log m
3. Multiply the log m by their respective frequency f log m
4. Add up all the products- $\sum f \log m$
5. Divide f log m by N - $\frac{\sum f \log m}{N}$
6. Find out the antilog of the result of step 5 and this will give the answer

The formula is:

$$\text{G.M.} = \text{Antilog} \frac{\sum f \log m}{N}$$

**Example:**

Calculate the geometric mean:

| Yield of wheat(mounds) | 7.5-10.5 | 10.5-13.5 | 13.5-16.5 | 16.5-19.5 | 19.5-22.5 | 22.5-25.5 | 25.5-28.5 |
|---|---|---|---|---|---|---|---|
| No. of farms | 5 | 9 | 19 | 23 | 7 | 4 | 1 |

**Solution:**

| Mid value (m) | Log m | f | f log m |
|---|---|---|---|
| 9 | 0.9542 | 5 | 4.7710 |
| 12 | 1.0792 | 9 | 9.7128 |
| 15 | 1.1761 | 19 | 22.3459 |
| 18 | 1.2553 | 23 | 28.8719 |
| 21 | 1.3222 | 7 | 9.2554 |
| 24 | 1.3802 | 4 | 5.5208 |
| 27 | 1.4314 | 1 | 1.4314 |
| | | N = 68 | ∑f log m = 81.9092 |

$$G.M. = \text{Antilog}\ \frac{\sum f \log m}{N}$$

$$= \frac{81.9092}{68} = 1.2045$$

**Merits of Geometric Mean:**

1. It is based on all observations
2. It is rigidly defined
3. It is capable of further algebraic treatment
4. It is less affected by the extreme values
5. It is useful in studying economic and social data

**Demerits of Geometric Mean:**

1. It is difficult to understand
2. Non-mathematical persons cannot do calculations
3. It has restricted application

**Uses of Geometric Mean:**

1. Geometric Mean is highly useful in averaging ratios, percentages and rate of increase between two periods

2. It is important in the construction of index numbers
3. In economic and social sciences, where we want to give more weight to smaller items and smaller weight to large items, geometric mean is appropriate

**Harmonic Mean (H.M):**

Harmonic Mean is a measure of central tendency in solving special type of problems. Harmonic mean is the reciprocal of values of various items in the variable. The reciprocal of a number is that value which is obtained by dividing one by the value. For example the reciprocal of 7 is 1/7; of 9 is 1/9. The reciprocal can be obtained from logarithm tables.

**Calculation of Harmonic Mean- Individual Series:**

Steps:

1. Find out the reciprocal of each size i.e., 1/x
2. Add all the reciprocals of all values ($\sum$ 1/x)
3. Apply the formula;

$$H.M. = \frac{N}{1/x_1 + 1/x_2 + 1/x_3 \ldots 1/x_n}$$

Or

$$H.M. = \frac{N}{\sum 1/x}$$

$x_1$ $x_2$ $x_3$ …… $x_n$ refer to the various values of the observations.

Example:

The monthly income of 10 families in rupees in a certain village is given below: Calculate Harmonic Mean

| Family | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Income | 85 | 70 | 10 | 75 | 500 | 8 | 42 | 250 | 40 | 36 |

**Solution:**

| Family | Income | Reciprocals (1/x) |
|---|---|---|
| 1 | 85 | 0.01176 |
| 2 | 70 | 0.01426 |
| 3 | 10 | 0.10000 |
| 4 | 75 | 0.01333 |
| 5 | 500 | 0.00200 |
| 6 | 8 | 0.12500 |
| 7 | 42 | 0.02318 |
| 8 | 250 | 0.00400 |
| 9 | 40 | 0.02500 |
| 10 | 36 | 0.02778 |
| | | $\sum 1/x = 0.34631$ |

$$H.M. = \frac{N}{1/x_1 + 1/x_2 + 1/x_3 \ldots 1/x_n}$$

Or

$$H.M. = \frac{N}{\sum 1/x}$$

$$= \frac{10}{0.34631} = \text{Rs. } 28.87$$

**Calculation of Harmonic Mean. - Discrete Series:**

**Steps:**

1. Find out the reciprocal of each size of items (1/x)
2. Multiply the reciprocal of each size by its frequency (f 1/x)
3. Add up all the products- $\sum$ f (1/x)

4. Apply the formula

$$H.M. = \frac{N}{\sum f(1/x)}$$

**Example:**

Calculate H.M. from the following:

| Size of item | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|
| Frequency | 4 | 6 | 9 | 5 | 2 | 8 |

**Solution:**

| Size of item | Frequency | Reciprocals (1/x) | Product of reciprocal (f 1/x) |
|---|---|---|---|
| 6 | 4 | 0.1667 | 0.6668 |
| 7 | 6 | 0.1429 | 0.8574 |
| 8 | 9 | 0.1250 | 1.1250 |
| 9 | 5 | 0.1111 | 0.5555 |
| 10 | 2 | 0.1000 | 0.2000 |
| 11 | 8 | 0.0909 | 0.7272 |
| | | | $\sum$ 1/x =4.1319 |

$$H.M. = \frac{N}{\sum f(1/x)} = \frac{34}{4.1319} = 8.23.$$

**Calculation of Harmonic Mean – Continuous Series:**

**Steps:**

1. Find out the reciprocal of each class m
2. Find out the reciprocal of each midvalue- 1/m
3. Multiply the reciprocal of each midvalue by its frequency f1/m
4. Add up all the products- $\sum$ f (1/m)
5. Apply the formula:

$$H.M. = \frac{N}{f_1 1/m_1 + f_2 1/m_2 + f_n 1/m_n}$$

or

$$H.M. = \frac{N}{\sum f\,(1/m)}$$

**Example:**

Calculate H.M. of the following data:

| Marks | 30-40 | 40-50 | 50-60 | 60-70 | 70-80 | 80-90 | 90-100 |
|---|---|---|---|---|---|---|---|
| Frequency | 15 | 13 | 8 | 6 | 15 | 7 | 6 |

**Solution:**

| Marks | Midvalue (m) | Frequency | Reciprocal (1/m) | f × reciprocal (f 1/m ) |
|---|---|---|---|---|
| 30-40 | 35 | 15 | 0.02857 | 0.42855 |
| 40-50 | 45 | 13 | 0.02222 | 0.28886 |
| 50-60 | 55 | 8 | 0.01818 | 0.14544 |
| 60-70 | 65 | 6 | 0.01534 | 0.09204 |
| 70-80 | 75 | 15 | 0.01333 | 0.19995 |
| 80-90 | 85 | 7 | 0.01176 | 0.08232 |
| 90-100 | 95 | 6 | 0.01053 | 0.06318 |
| | | $\sum f = 70$ | | $\sum f\ 1/m = 1.30034$ |

**Merits of Harmonic Mean:**

1. It is rigidly defined
2. It is based on all the observations of the series
3. It is suitable in case of series having wide dispersion
4. It is suitable for further mathematical treatment
5. It gives less weight to large items and more weight to small items

**Relationship between Mean, Geometric mean and Harmonic Mean:**

If all the items in a variable are the same the arithmetic mean, the geometric mean and harmonic mean are equal. If all the items in a distribution have the same value then,

$$\bar{X} > G.M > H.M$$

But if the size vary, as will generally be the case, mean will be greater than the geometric mean and geometric mean will be greater than the harmonic mean. This is because of the property of the geometric mean to give larger weight to smaller item and of the harmonic mean to give the largest weight to the smallest items. Hence,

$$\bar{X} > G.M > H.M.$$

**Demerits of Harmonic Mean:**

1. It is difficult to calculate and is not understandable
2. All the values must be available for computation
3. It is not popular
4. It is usually a value which does not exist in series

# MEASURES OF DISPERSION

## Introduction

The measures of central tendencies (i.e. means) indicate the general magnitude of the data and locate only the center of a distribution of measures. They do not establish the degree of variability or the spread out or scatter of the individual items and their deviation from (or the difference with) the means.

i) According to Nciswanger, "Two distributions of statistical data may be symmetrical and have common means, medians and modes and identical frequencies in the modal class. Yet with these points in common they may differ widely in the scatter or in their values about the measures of central tendencies."

ii) Simpson and Kafka said, "An average alone does not tell the full story. It is hardly fully representative of a mass, unless we know the manner in which the individual item. Scatter around it .... a further description of a series is necessary, if we are to gauge how representative the average is."

From this discussion we now focus our attention on the scatter or variability which is known as **dispersion**. Let us take the following three sets.

| Students | Group X | Group Y | Group Z |
|---|---|---|---|
| 1 | 50 | 45 | 30 |
| 2 | 50 | 50 | 45 |
| 3 | 50 | 55 | 75 |
| $\therefore$ mean $\bar{x} \Rightarrow$ | 50 | 50 | 50 |

Thus, the three groups have same mean i.e. 50. In fact the median of group X and Y are also equal. Now if one would say that the students from the three groups are of equal capabilities, it is totally a wrong conclusion then. Close examination reveals that in group X students have equal marks as the mean,

students from group Y are very close to the mean but in the third group Z, the marks are widely scattered. It is thus clear that the measures of the central tendency is alone not sufficient to describe the data.

**Definition of dispersion:** The arithmetic mean of the deviations of the values of the individual items from the measure of a particular central tendency used. Thus the 'dispersion' is also known as the "**average of the second degree.**" Prof. Griffin and Dr. Bowley said the same about the dispersion.

In measuring dispersion, it is imperative to know the amount of variation (absolute measure) and the degree of variation (relative measure). In the former case we consider the range, mean deviation, standard deviation etc. In the latter case we consider the coefficient of range, the coefficient mean deviation, the coefficient of variation etc.

**Methods of Computing Dispersion**

(I) Method of limits:

(1) The range (2) Inter-quartile range

(II) Method of Averages:

(1) Quartile deviation (2) Mean deviation

(3) Standard Deviation and (4) Lorenz curve

Note that, we are going to study some of these and not all.

**RANGE**

In any statistical series, the difference between the largest and the smallest values is called as the range.

Thus Range (R) = L - S $\begin{cases} L = \text{Largest value of the series.} \\ S = \text{Smallest value of the series.} \end{cases}$

**Coefficient of Range:** The relative measure of the range. It is used in the comparative study of the dispersion co-efficient of Range = $\frac{L-S}{L+S}$

**Example** (Individual series ) Find the range and the co-efficient of the range of the following items :

110, 117, 129, 197, 190, 100, 100, 178, 255, 790.

**Solution:** R = L - S = 790 - 100 = 690

Co-efficient of Range = $\frac{L-S}{L+S} = \frac{790-100}{790+100} = \frac{690}{890} = 0.78$

**Example** (Continuous series) find the range and its co-efficient from the following data.

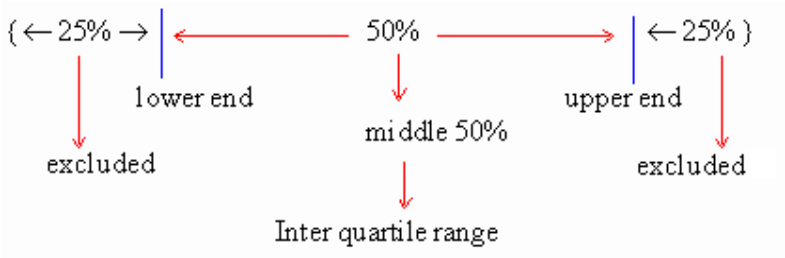
| Size: | 10 - 20 | 20 - 30 | 30 - 40 | 40 - 50 | 50 - 10 |
|---|---|---|---|---|---|
| Frequency : | 2 | 3 | 5 | 4 | 2 |

**Solution:** R = L - S = 100 - 10 = 90

Co-efficient of range = $\frac{L-S}{L+S} = \frac{100-10}{100+10} = \frac{90}{110} = 0.82$

**QUARTILES AND INTERQUARTILE RANGE**

If we concentrate on two extreme values ( as in the case of range ), we don't get any idea about the scatter of the data within the range ( i.e. the two extreme values ). If we discard these two values the limited range thus available might be more informative. For this reason the concept of interquartile range is developed. It is the range which includes middle 50% of the distribution. Here 1/4 ( one quarter of the lower end and 1/4 ( one quarter ) of the upper end of the observations are excluded.




Now the lower quartile ( $Q_1$ ) is the 25th percentile and the upper quartile ( $Q_3$ ) is the 75th percentile. It is interesting to note that the 50th percentile is the middle quartile ( $Q_2$ ) which is in fact what you have studied under the title ' Median ". Thus symbolically

Inter quartile range = $Q_3 - Q_1$

If we divide ( $Q_3 - Q_1$ ) by 2 we get what is known as Semi-Inter quartile range.

i.e. $\frac{Q_3 - Q_1}{2}$. It is known as Quartile deviation ( Q. D or SI QR ).

Therefore Q. D. = $\frac{Q_3 - Q_1}{2}$

$$\text{Coefficient of Q.D} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Quartile Deviation is an improved measure over the range, as it is not calculated from extreme items, but on quartiles. For a symmetrical distribution, we have

$$\text{Median} + \text{Q.D} = Q_3$$

$$\text{Median} = \text{Q.D} + Q_1 = \frac{Q_3 - Q_1}{2} + Q_1 + \frac{Q_3 + Q_1}{2}$$

$$Q_1 = \text{Median} - \text{Q.D}$$

Quartile deviation gives an idea of the distribution of the middle half of the items around the median.

**Example:**

Calculate Quartile Deviation and its coefficient

| Age in years | 20 | 30 | 40 | 50 | 60 | 70 | 80 |
|---|---|---|---|---|---|---|---|
| No. of members | 3 | 61 | 132 | 153 | 140 | 51 | 3 |

**Solution:**

| Age in years | No. of members | C.F |
|---|---|---|
| 20 | 3 | 3 |
| 30 | 61 | 64 |
| 40 | 132 | 196 |
| 50 | 153 | 349 |
| 60 | 140 | 489 |
| 70 | 51 | 540 |
| 80 | 3 | 543 |

$$Q_1 = \text{value of } \frac{N+1}{4} \text{ th item}$$

$$= \text{value of } \frac{543+1}{4} \text{ th item}$$

$$= 136^{th} \text{ item}$$

$$= 40 \text{ years}$$

$$Q_3 = \text{value of } 3\left(\frac{N+1}{4}\right) \text{ th item}$$

$$= \text{value of } 3\left[\frac{543+1}{4}\right] \text{ th item}$$

$$= \text{value of } 3 \times 136^{th} \text{ item}$$

$$= 408^{th} \text{ item which is 60 years}$$

$$\text{Therefore Q. D.} = \frac{Q_3 - Q_1}{2}$$

$$= \frac{60-40}{2} = \frac{20}{2} = 10$$

$$\text{Coefficient of Q.D} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{60-40}{60+40} = \frac{20}{100} = 0.2$$

Calculate the Quartile deviation and its coefficient

| Wages (Rs) | 30-32 | 32-34 | 34-36 | 36-38 | 38-40 | 40-42 | 42-44 |
|---|---|---|---|---|---|---|---|
| Workers | 12 | 18 | 16 | 14 | 12 | 8 | 6 |

$$Q_1 = \text{size of } \frac{N}{4} \text{ th item}$$

$$= \frac{86}{4} = 21.5$$

$Q_1$ lies in the group 32-34

$$Q_1 = L + \frac{N/4 - c.f}{F} \times i$$

$$= 32 + \frac{21.5-12}{18} \times 2$$

$$= 32 + 1.06$$

$$= 33.06$$

$$Q_3 = \text{size of } \frac{3N}{4} \text{ th item}$$

$$= \frac{3 \times 86}{4} = 64.5^{th} \text{ item}$$

$Q_3$ lies in the group 38-40

$$Q_3 = L + \frac{3N/4 - c.f}{F} \times i$$

$$= 38 + \frac{64.5-60}{12} \times 2$$

$$= 38 + 0.75 = 38.75$$

Therefore Q. D. = $\frac{Q_3 - Q_1}{2}$

$$Q.D = \frac{38.75-33.06}{2}$$

$$= \frac{5.69}{2} = 2.85$$

$$\text{Coefficient of Q.D} = \frac{Q_3-Q_1}{Q_3+Q_1}$$

$$= \frac{38.75-33.06}{38.75+33.06} = 0.08$$

**Merits of Quartile Deviation:**

1. It is simple to understand and easy to compute
2. It is not influenced by the extreme values
3. It can be found out with open end distribution
4. It is not affected by presence of extreme values

**Demerits of Quartile Deviation:**

1. It ignores the first 25% of the items and the last 25% of the items
2. It is a positional average; hence not amenable to further mathematical treatment
3. Its value is affected by sampling fluctuations
4. It gives only a rough measure
5. It is not the representative value of data

**MEAN DEVIATION**

Average deviations ( mean deviation ) is the average amount of variations (scatter) of the items in a distribution from either the mean or the median or the mode, ignoring the signs of these deviations by Clark and Senkade.

**Individual Series**

**Steps:**

(1) Find the mean or median or mode of the given series.

(2) Using and one of three, find the deviations (differences) of the items of the series from them.

i.e. xi - x, xi - Me and xi - Mo.

Me = Median and Mo = Mode.

(3) Find the absolute values of these deviations i.e. ignore there positive (+) and negative (-) signs.

i.e. | xi - x | , | xi - Me | and xi - Mo |.

(4) Find the sum of these absolute deviations.

i.e. $\Sigma$ | xi - x | + , $\Sigma$ | xi - Me | , and $\Sigma$ | xi - Mo | .

(5) Find the mean deviation using the following formula.

$$M.D = (\partial x) = \frac{\Sigma \text{ xi} - x}{}, \quad \partial(Me) = \frac{\Sigma \text{ xi} - Me}{}$$

$$\text{or } \partial(Mo) = \frac{\text{xi} \quad Mo}{}$$

Note that:

(i) Generally M. D. obtained from the median is the best for the practical purpose.

(ii) co-efficient of M. D. = $\frac{\partial \bar{x}}{\bar{x}} = \frac{\partial(Me)}{Me} = \frac{\partial(Mo)}{Mo}$

**Example** Calculate Mean deviation and its co-efficient for the following salaries:

$ 1030, $ 500, $ 680, $ 1100, $ 1080, $ 1740. $ 1050, $ 1000, $ 2000, $ 2250, $ 3500 and $ 1030.

**Solution :**

| Size ($x_i$) | Frequency (f) | C. f. | $\lvert x_i - Me \rvert$ | $\lvert x_i - Me \rvert$ |
|---|---|---|---|---|
| 4 | 2 | 2 | $\lvert 8-4 \rvert = 4$ | $2 \times 4 = 8$ |
| 6 | 4 | $2+4=6$ | $\lvert 8-6 \rvert = 2$ | $4 \times 2 = 8$ |
| 8 | 5 | $6+5=11$ | $\lvert 8-8 \rvert = 0$ | $5 \times 0 = 0$ |
| 10 | 3 | $11+3=14$ | $\lvert 8-10 \rvert = 2$ | $3 \times 2 = 6$ |
| 12 | 2 | $14+2=16$ | $\lvert 8-12 \rvert = 4$ | $2 \times 4 = 8$ |
| 14 | 1 | $16+1=17$ | $\lvert 8-14 \rvert = 4$ | $1 \times 6 = 6$ |
| 16 | 4 | $17+4=21$ | $\lvert 8-16 \rvert = 8$ | $4 \times 8 = 32$ |
| | $n = 21$ | | | $\lvert x_i - Me \rvert = 68$ |

**Calculations:**

i) Median (Me) = Size of $\left(\frac{n+1}{2}\right)^{th}$ item = size of $\left(\frac{21+1}{2}\right)^{th}$ item

= Size of $11_{th}$ item.

Therefore, Median (Me) = 8

ii) M. D. = $\frac{\Sigma \lvert xi - Me \rvert}{Me} = \frac{68}{21} = 3.24$

**Example** (Continuous series) Calculate the mean deviation and the coefficient of mean deviation from the following data using the mean.

Difference in ages between boys and girls of a class

| Diff. in years: | No. of students: |
|---|---|
| 0 - 5 | 449 |
| 5 - 10 | 705 |
| 10 - 15 | 507 |
| 15 - 20 | 281 |
| 20 - 25 | 109 |
| 25 - 30 | 52 |
| 30 - 35 | 16 |
| 35 - 40 | 4 |

**Solution :**

| Diff. in age | Mid-values ($x_i$) | frequency ($f_i$) | $f_i x_i$ | $\lvert x_i - \bar{x} \rvert$ | $f_i \lvert x_i - \bar{x} \rvert$ |
|---|---|---|---|---|---|
| 0 - 5 | 2.5 | 449 | 1122.5 | 8 | 3592 |
| 5 - 10 | 7.5 | 705 | 5287.5 | 3 | 2115 |
| 10 - 15 | 12.5 | 507 | 6337.5 | 2 | 1014 |
| 15 - 20 | 17.5 | 281 | 4917.5 | 7 | 1967 |
| 20 - 25 | 22.5 | 109 | 2452.5 | 12 | 1308 |
| 25 - 30 | 27.5 | 52 | 1430.0 | 17 | 884 |
| 30 - 35 | 32.5 | 16 | 520.0 | 22 | 352 |
| 35 - 40 | 37.5 | 4 | 150.0 | 27 | 108 |
| | | $n = 2123$ | $\Sigma f_i x_i = 22217.5$ | | $\Sigma f_i \lvert x_i - \bar{x} \rvert = 11440$ |

**Calculation:**

1) $\bar{X} = \frac{\Sigma f_i x_i}{n} = \frac{22217.5}{2123} = 10.5 \text{ (approx.)}$

2) M. D. $= \frac{\Sigma f_i \left| x_i - \bar{x} \right|}{n} = \frac{11440}{2123} = 5.4$

3) Co –efficient of M. D. $= \frac{M.D.}{\bar{x}} = \frac{5.4}{10.5} = 0.514$

**VARIANCE**

The term variance was used to describe the square of the standard deviation R.A. Fisher in 1913. The concept of variance is of great importance in advanced work where it is possible to split the total into several parts, each attributable to one of the factors causing variations in their original series. Variance is defined as follows:

Variance = $\frac{\Sigma (x_i - \bar{x})^2}{n}$

**STANDARD DEVIATION (S. D.)**

It is the square root of the arithmetic mean of the square deviations of various values from their arithmetic mean. it is denoted by S.D or σ.

Thus, S.D. ($\sigma$ x ) = $\sqrt{\frac{\Sigma \left( x_i - \bar{x} \right)^2}{n}}$ for the ungrouped data

$= \sqrt{\frac{\Sigma f_i \left( x_i - \bar{x} \right)^2}{n}}$ for the grouped data

Where n = Σ fi

**Merits:**

(1) It is rigidly defined and based on all observations.

(2) It is amenable to further algebraic treatment.

(3) It is not affected by sampling fluctuations.

(4) It is less erratic.

**Demerits:**

(1) It is difficult to understand and calculate.

(2) It gives greater weight to extreme values.

Note that variance V(x) = $\frac{\Sigma\left(x_i - \bar{x}\right)^2}{n}$ and

and s. d. ( $\sigma$ x ) = $\sqrt{\frac{\Sigma\left(x_i - \bar{x}\right)^2}{n}}$ and $\sqrt{\frac{\Sigma f_i\left(x_i - \bar{x}\right)^2}{n}}$

Then V ( x ) = $\sigma_x^2$

**CO-EFFICIENT OF VARIATION ( C. V. )**

To compare the variations (dispersion) of two different series, relative measures of standard deviation must be calculated. This is known as co-efficient of variation or the co-efficient of s. d. Its formula is

$$\text{C. V.} = \frac{\sigma x}{\bar{x}} \times 100$$

Thus it is defined as the ratio s. d. to its mean.

Remark: It is given as a percentage and is used to compare the consistency or variability of two more series. The higher the C. V. , the higher the variability and lower the C. V., the higher is the consistency of the data.

**Example** Calculate the standard deviation and its co-efficient from the following data.

| A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|
| 10 | 12 | 16 | 8 | 25 | 30 | 14 | 11 | 13 | 11 |

**Solution :**

| No<br>No. | $x_i$ | $(x_i - x)$ | $(x_i - x)^2$ |
|---|---|---|---|
| A | 10 | -5 | 25 |
| B | 12 | -3 | 9 |
| C | 16 | +1 | 1 |
| D | 8 | -7 | 49 |
| E | 25 | +10 | 100 |
| F | 30 | +15 | 225 |
| G | 14 | -1 | 1 |
| H | 11 | -5 | 16 |
| I | 13 | -2 | 4 |
| J | 11 | -4 | 16 |
| n= 10 | Σ xi = 150 | | Σ∈lxi - x l2= 446 |

**Calculations:**

i) $\bar{x} = \frac{\Sigma x_i}{n} = \frac{150}{10} = 15$

ii) $s.d.(\sigma x) = \sqrt{\frac{\Sigma\left(x_i - \bar{x}\right)^2}{n}} = \sqrt{\frac{446}{10}} = 6.7$

iii) $\text{co-efficient of s. d.} = \frac{\sigma x}{\bar{x}} = \frac{6.7}{15} = 0.45$

**Example** Prove that s. d. $(sx) = \sqrt{\frac{\Sigma xi^2}{n} - (\bar{x})^2}$

**Solution :** Now consider

$$\frac{\Sigma(xi - \bar{x})^2}{n} = \frac{\Sigma(xi^2 - 2xi\bar{x} + \bar{x}^2)}{n}$$

$$= \frac{\Sigma xi^2}{n} - 2\bar{x}\frac{\Sigma xi}{n} + \bar{x}^2\frac{\Sigma 1}{n}$$

$$= \frac{\Sigma x_i^2}{n} - 2\bar{x}(\bar{x}) + \bar{x}^2 \cdot \frac{n}{n}$$

$$\ldots(\because \frac{\Sigma x_i}{n} = \bar{x} \text{ and } \Sigma 1 = n)$$

$$= \frac{\Sigma x_i^2}{n} - 2(\bar{x})^2 + (\bar{x})^2$$

$$\text{Therefore, } \sigma x = \sqrt{\frac{\Sigma xi^2}{n} - (\bar{x})^2}$$

**This is known as a short-cut formula for standard deviation.**

The above problem can also be solved by this formula as:

**Solution :** We have $\Sigma x_i^2 = 2695$,

$n = 10$ and $\bar{x} = 15$ then

$$\text{s. d. } (sx) = \sqrt{\frac{2696}{10} - (15)^2}$$

$$= \sqrt{269.6 - 225} = \sqrt{44.6} = 6.67$$

**Example**: Calculate S.D of the marks of 100 students.

| Marks | No. of students ($f_i$) | Mid-values ($x_i$) | $f_i x_i$ | $f_i x_i^2$ |
|---|---|---|---|---|
| 0-2 | 10 | 1 | 10 | 10 |
| 2-4 | 20 | 3 | 60 | 180 |
| 4-6 | 35 | 5 | 175 | 875 |
| 6-8 | 30 | 7 | 210 | 1470 |
| 8-10 | 5 | 9 | 45 | 405 |
| | n = 100 | | $\Sigma f_i x_i = 500$ | $\Sigma f_i x_i^2 = 2940$ |

1) $$\bar{x} = \frac{\Sigma f_i x_i}{n} = \frac{5990}{1000} = 5.99$$

2) $$s.d.(\sigma x) = \sqrt{\frac{\Sigma f_i x_i^2}{n} - (\bar{x})^2} = \sqrt{\frac{38770}{1000} - (5.99)^2}$$

$$= \sqrt{38.77 - 35.88} = \sqrt{2.89} = 1.7 \ (\$ \text{ in } 100)$$

**Combined Standard deviation:** If two sets containing $n_1$ and $n_2$ items having means $x_1$ and $x_2$ and standard deviations $\sigma_1$ and $\sigma_2$ respectively are taken together then,

(1) Mean of the combined data is $$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

(2) S.D. of the combined set is $$\sigma = \sqrt{\frac{n_1\left(\sigma_1^2 + d_1^2\right) + n_2\left(\sigma_2^2 + d_2^2\right)}{n_1 + n_2}}$$

Where $d_1 = \bar{x}_1 - \bar{x}$ and $d_2 = \bar{x}_2 - \bar{x}$

**Example** The score of two teams A and B in 10 matches are as:

| A : | B: |
|---|---|
| 40 | 21 |
| 32 | 14 |
| 0 | 29 |
| 40 | 13 |
| 30 | 5 |
| 7 | 12 |
| 13 | 10 |
| 25 | 13 |
| 14 | 30 |
| 5 | 0 |

Find the variance for both the series. Which team is more consistent?

| A | | | B | | |
|---|---|---|---|---|---|
| $x_i$ | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ | $y_i$ | $(y_i - \bar{y})$ | $(y_i - \bar{y})^2$ |
| 40 | 19 | 361 | 21 | 4 | 16 |
| 32 | 11 | 121 | 14 | -3 | 9 |
| 0 | -21 | 441 | 14 | 12 | 144 |
| 40 | 19 | 361 | 30 | 13 | 169 |
| 30 | 9 | 81 | 5 | -12 | 144 |
| 7 | -8 | 196 | 12 | -5 | 25 |
| 13 | -8 | 64 | 10 | -7 | 49 |
| 25 | 4 | 16 | 13 | -4 | 16 |
| 14 | -7 | 49 | 30 | 13 | 169 |
| 9 | -12 | 144 | 6 | -11 | 121 |
| $\Sigma x_i = 210$ | | $\Sigma(x_i - \bar{x})^2 = 1834$ | $\Sigma x_i = 110$ | | $\Sigma(x_i - \bar{y})^2 = 862$ |

**Example** Calculate the coefficient of variation of a series on the basis of the following results:

$n = 50,\ \Sigma(x_i - 75) = -10,\ \Sigma(x_i - 7.5)^2 = 400$

**Solution:** 1) $\Sigma(x_i - 75) = \Sigma x_i - 375 = -10$ ...

(since $\Sigma 7.5 = 7.5n = 7.5 \times 50 = 375$)

$$\Sigma x_i = -10 + 375 = 365$$

Therefore, $\bar{x} = \frac{\Sigma x_i}{n} = \frac{365}{50} = 7.3$

2) $$\Sigma (x_i - 7.5)^2 = \Sigma (x_i - 7.3 - 0.2)^2$$

$$= \Sigma (x_i - 7.3 - 0.2)^2 - 0.4\,\Sigma (x_i - 7.3) + 0.04\,n$$

$$= \Sigma (x_i - \bar{x})^2 - 0.4\,\Sigma (x_i - \bar{x}) + 0.04 \times 50$$

But $\Sigma (x_i - \bar{x}) = 0$

$$\therefore \Sigma (x_i - 7.5)^2 = \Sigma (x_i - \bar{x})^2 - 0.4\,\Sigma (0) + 2 = 400$$

$$\therefore \Sigma (x_i - \bar{x})^2 = 398$$

Now $\sigma x =$

$$\sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n}} = \sqrt{\frac{398}{50}} = 2.8213$$

3) C. V. $=$

$$\frac{2.8213}{7.3} \times 100 = 38.65\%$$

**Solution**

1) $\bar{x} = \frac{\Sigma xi}{n} = \frac{210}{10}$ and $\bar{y} = \frac{\Sigma yi}{n} = \frac{170}{10} = 17$

$\therefore \sigma x = \sqrt{\frac{\Sigma (xi - \bar{x})^2}{n}} = \sqrt{\frac{1834}{10}} = \sqrt{183.4} = 13.54$

Also $\sigma y = \sqrt{\frac{\Sigma (yi - \bar{y})^2}{n}} = \sqrt{\frac{862}{10}} = \sqrt{86.2} = 9.28$

Therefore, $(C.V.)_A = \frac{13.54}{21} \times 100 = 64.47\%$

and $(C.V.)_B = \frac{9.28}{17} \times 100 = 54.69\%$

Since $(C.V.)_A < (C.V.)_B$

$\therefore$ The base ball team A is more consistent.

**LORENZ CURVE**:

In Economics, the **Lorenz curve** is a graphical representation of the distribution of income or of wealth. It is the form of a curve which is derived from the cumulative percentage of the given variables. This curve was given by Dr. Max O. Lorenz a popular Economic- Statistician. He studied distribution of Wealth and Income with its help.

It is graphic method to study dispersion. It helps in studying the variability in different components of distribution especially economic. The base of Lorenz Curve is that we take cumulative percentages along X and Y axis. Joining these points we get the Lorenz Curve. Lorenz Curve is of much importance in the comparison of two series graphically. It gives us a clear cut visual view of the series to be compared.

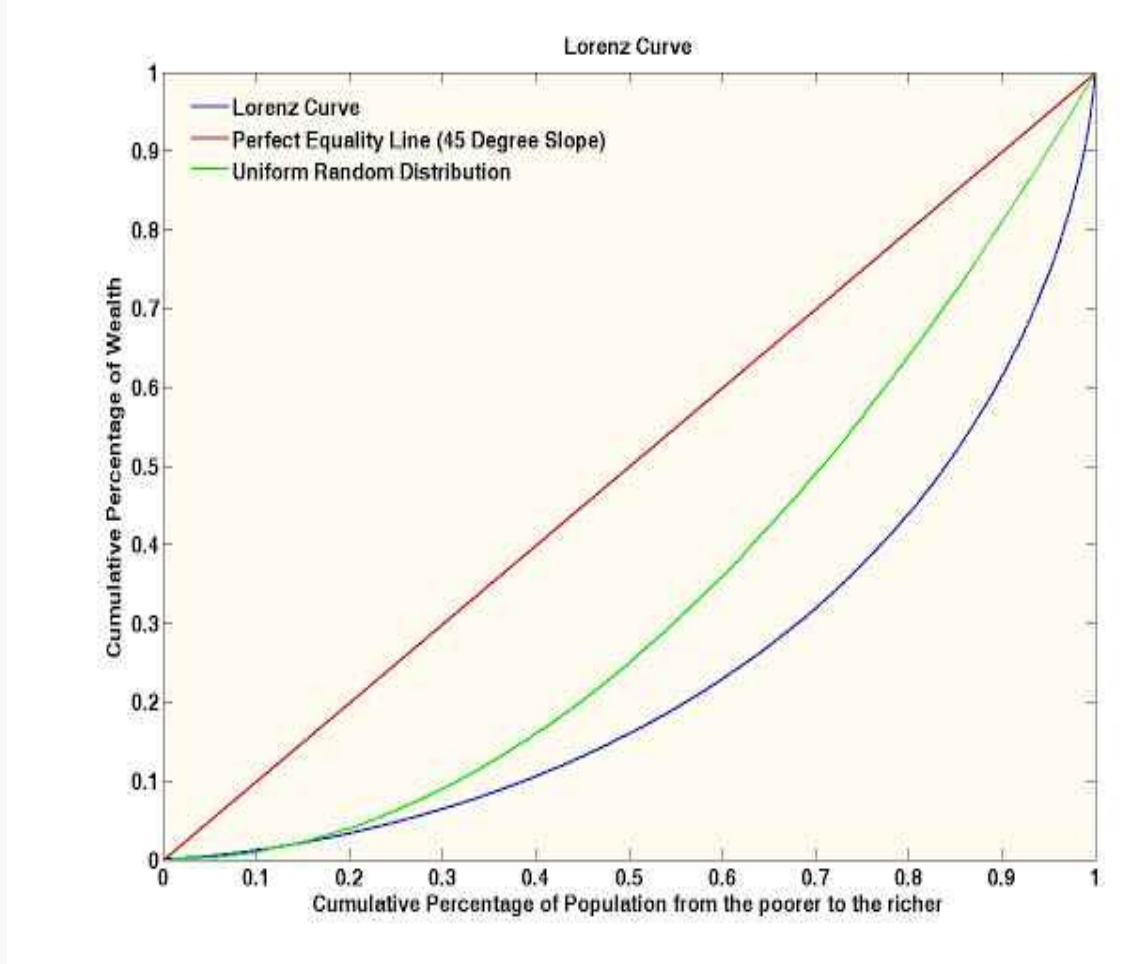
**Steps to plot 'Lorenz Curve':**

1. Cumulate both values and their corresponding frequencies.

2. Find the percentage of each of the cumulated figures taking the grand total of each corresponding column as 100.
3. Represent the percentage of the cumulated frequencies on X axis and those of the values on the Y axis.

4. Draw a diagonal line designated as the line of equal distribution.

5. Plot the percentages of cumulated values against the percentages of the cumulated frequencies of a given distribution and join the points so plotted through a free hand curve.



**Correlation and Regression:**

The word correlation is used in everyday life to denote some form of association. We might say that we have noticed a correlation between foggy days and attacks of wheeziness. However, in statistical terms we use correlation to denote association between two quantitative variables. We also assume that the association is linear, that one variable increases or decreases a fixed amount for a unit increase or decrease in the other. The other technique that is often used in these circumstances is regression, which involves estimating the best straight line to summarise the association.

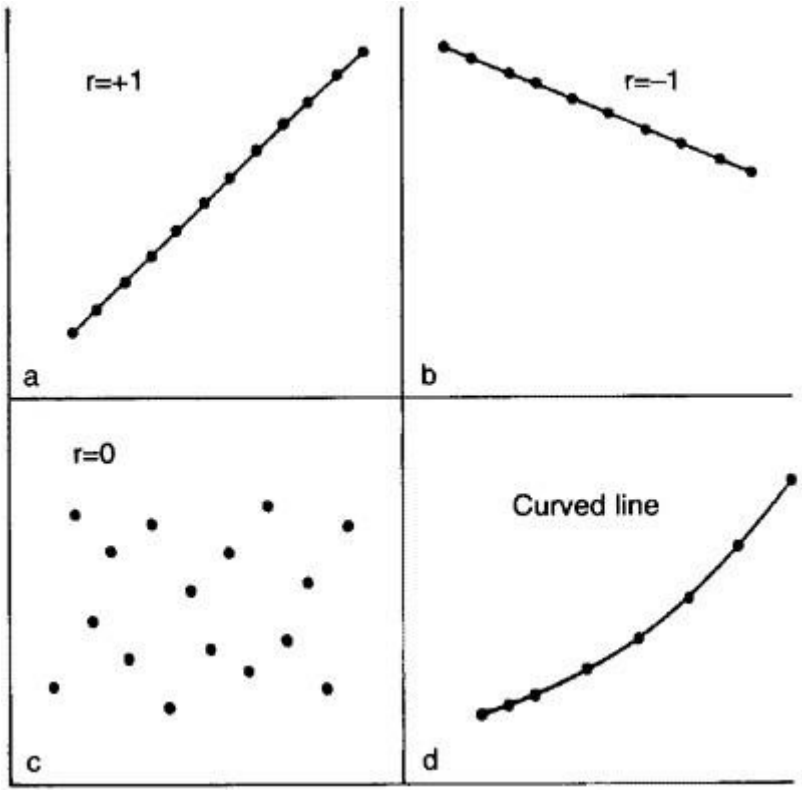
The value of $r$ is such that $-1 \leq r \leq +1$. The + and – signs are used for positive linear correlations and negative linear correlations, respectively.

- *Positive correlation:* If $x$ and $y$ have a strong positive linear correlation, $r$ is close to +1. An $r$ value of exactly +1 indicates a perfect positive fit. Positive values indicate a relationship between $x$ and $y$ variables such that as values for $x$ increases, values for $y$ also increase.
- *Negative correlation:* If $x$ and $y$ have a strong negative linear correlation, $r$ is close to -1. An $r$ value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between $x$ and $y$ such that as values for $x$ increase, values for $y$ decrease.
- *No correlation:* If there is no linear correlation or a weak linear correlation, $r$ is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables
- Note that $r$ is a dimensionless quantity; that is, it does not depend on the units employed.
- A *perfect* correlation of ± 1 occurs only when the data points all lie exactly on a straight line. If $r = +1$, the slope of this line is positive. If $r = -1$, the slope of this line is negative.
- A correlation greater than 0.8 is generally described as *strong*, whereas a correlation less than 0.5 is generally described as *weak*.

**Correlation coefficient**

The degree of association is measured by a correlation coefficient, denoted by r. It is sometimes called Pearson's correlation coefficient after its originator and is a measure of linear association. If a curved line is needed to express the relationship, other and more complicated measures of the correlation must be used.

The correlation coefficient is measured on a scale that varies from + 1 through 0 to - 1. Complete correlation between two variables is expressed by either + 1 or -1. When one variable increases as the other increases the correlation is positive; when one decreases as the other increases it is negative. Complete absence of correlation is represented by 0. Figure 11.1 gives some graphical representations of correlation.




# Scatter diagrams

When an investigator has collected two series of observations and wishes to see whether there is a relationship between them, he or she should first construct a scatter diagram. The vertical scale represents one set of measurements and the horizontal scale the other. If one set of observations

consists of experimental results and the other consists of a time scale or observed classification of some kind, it is usual to put the experimental results on the vertical axis. These represent what is called the "dependent variable". The "independent variable", such as time or height or some other observed classification is measured along the horizontal axis, or baseline.

The words "independent" and "dependent" could puzzle the beginner because it is sometimes not clear what is dependent on what. This confusion is a triumph of common sense over misleading terminology, because often each variable is dependent on some third variable, which may or may not be mentioned. It is reasonable, for instance, to think of the height of children as dependent on age rather than the converse but consider a positive correlation between mean tar yield and nicotine yield of certain brands of cigarette.' The nicotine liberated is unlikely to have its origin in the tar: both vary in parallel with some other factor or factors in the composition of the cigarettes. The yield of the one does not seem to be "dependent" on the other in the sense that, on average, the height of a child depends on his age. In such cases it often does not matter which scale is put on which axis of the scatter diagram. However, if the intention is to make inferences about one variable from the other, the observations from which the inferences are to be made are usually put on the baseline. As a further example, a plot of monthly deaths from heart disease against monthly sales of ice cream would show a negative association. However, it is hardly likely that eating ice cream protects from heart disease! It is simply that the mortality rate from heart disease is inversely related - and ice cream consumption positively related - to a third factor, namely environmental temperature.

## Calculation of the correlation coefficient

A paediatric registrar has measured the pulmonary anatomical dead space (in ml) and height (in cm) of 15 children. The data are given in table 11.1 and the scatter diagram shown in figure 11.2 Each dot represents one child, and it is placed at the point corresponding to the measurement of the height (horizontal axis) and the dead space (vertical axis). The registrar now inspects the pattern to see whether it seems likely that the area covered by the dots centres on a straight line or whether a curved line is needed. In this case the paediatrician decides that a straight line can adequately describe the general

trend of the dots. His next step will therefore be to calculate the correlation coefficient.

Table 11.1 Correlation between height and pulmonary anatomical dead space in 15 children

| Child number | Height (cm) | Dead space (ml), y |
|---|---|---|
| 1 | 110 | 44 |
| 2 | 116 | 31 |
| 3 | 124 | 43 |
| 4 | 129 | 45 |
| 5 | 131 | 56 |
| 6 | 138 | 79 |
| 7 | 142 | 57 |
| 8 | 150 | 56 |
| 9 | 153 | 58 |
| 10 | 155 | 92 |
| 11 | 156 | 78 |
| 12 | 159 | 64 |
| 13 | 164 | 88 |
| 14 | 168 | 112 |
| 15 | 174 | 101 |
| **Total** | **2169** | **1004** |
| **Mean** | **144.6** | **66.933** |

When making the scatter diagram to show the heights and pulmonary anatomical dead spaces in the 15 children, the pediatrician set out figures as in columns (1), (2), and (3) of table . It is helpful to arrange the observations in serial order of the independent variable when one of the two variables is clearly identifiable as independent. The corresponding figures for the dependent variable can then be examined in relation to the increasing series for the independent variable. In this way we get the same picture, but in numerical form, as appears in the scatter diagram.

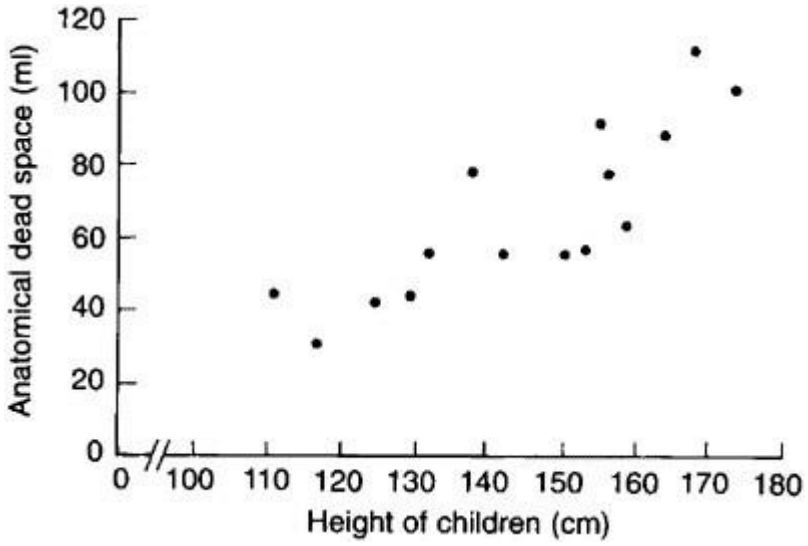



Figure Scatter diagram of relation in 15 children between height and pulmonary anatomical dead space.

The calculation of the correlation coefficient is as follows, with x representing the values of the independent variable (in this case height) and y representing the values of the dependent variable (in this case anatomical dead space). The formula to be used is:

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{[\Sigma(x - \bar{x})^2 (y - \bar{y})^2]}}$$

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{[\Sigma(x - \bar{x})^2 (y - \bar{y})^2]}}$$

which can be shown to be equal to:

$$r = \frac{\Sigma xy - \bar{n}\bar{x}\bar{y}}{(n - 1)SD\,(x)SD(y)}$$

## Calculator procedure

Find the mean and standard deviation of x, as described in $\bar{x}, SD(x)$

$\bar{x} = 144.6, SD(x) = 19.3769$

Find the mean and standard deviation of y: $\bar{y}, SD(y)$

$\bar{y} = 66.93, SD(y) = 23.6476$

Subtract 1 from n and multiply by SD(x) and SD(y), (n - 1)SD(x)SD(y)

$14 \times 19.3679 \times 23.6976 (6412.0609)$

This gives us the denominator of the formula. (Remember to exit from "Stat" mode.)

For the numerator multiply each value of x by the corresponding value of y, add these values together and store them.

110 x 44 = *Min*
116 x 31 = *M+*
etc.

This stores $\Sigma xy\ (150605)$ in memory. Subtract $n\bar{x}\bar{y}$

MR - 15 x 144.6 x 66.93 (5426.6)

Finally divide the numerator by the denominator.

r = 5426.6/6412.0609 = 0.846.

The correlation coefficient of 0.846 indicates a strong positive correlation between size of pulmonary anatomical dead space and height of child. But in interpreting correlation it is important to remember that correlation is not causation. There may or may not be a causative connection between the two correlated variables. Moreover, if there is a connection it may be indirect.

A part of the variation in one of the variables (as measured by its variance) can be thought of as being due to its relationship with the other variable and another part as due to undetermined (often "random") causes.

The part due to the dependence of one variable on the other is measured by Rho . For these data Rho= 0.716 so we can say that 72% of the variation between children in size of the anatomical dead space is accounted for by the height of the child. If we wish to label the strength of the association, for absolute values of r, 0-0.19 is regarded as very weak, 0.2-0.39 as weak, 0.40-0.59 as moderate, 0.6-0.79 as strong and 0.8-1 as very strong correlation, but these are rather arbitrary limits, and the context of the results should be considered.

## Spearman rank correlation

A plot of the data may reveal outlying points well away from the main body of the data, which could unduly influence the calculation of the correlation coefficient. Alternatively the variables may be quantitative discrete such as a mole count, or ordered categorical such as a pain score. A non-parametric procedure, due to Spearman, is to replace the observations by their ranks in the calculation of the correlation coefficient.

This results in a simple formula for Spearman's rank correlation, Rho.

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

where d is the difference in the ranks of the two variables for a given individual. Thus we can derive table 11.2 from the data in table 11.1 .

| Table 11.2 Derivation of Spearman rank correlation from data of table 11.1 | | | | |
|---|---|---|---|---|
| Child number | Rank height | Rank dead space | d | $d^2$ |
| 1 | 1 | 3 | 2 | 4 |
| 2 | 2 | 1 | -1 | 1 |
| 3 | 3 | 2 | -1 | 1 |
| 4 | 4 | 4 | 0 | 0 |
| 5 | 5 | 5.5 | 0.5 | 0.25 |
| 6 | 6 | 11 | 5 | 25 |
| 7 | 7 | 7 | 0 | 0 |
| 8 | 8 | 5.5 | -2.5 | 6.25 |
| 9 | 9 | 8 | -1 | 1 |
| 10 | 10 | 13 | 3 | 9 |
| 11 | 11 | 10 | -1 | 1 |
| 12 | 12 | 9 | -3 | 9 |
| 13 | 13 | 12 | -1 | 1 |
| 14 | 14 | 15 | 1 | 1 |
| 15 | 15 | 14 | -1 | 1 |
| Total | | | | 60.5 |

From this we get that

From this we get that

$$r_s = 1 - \frac{6 \times 60.5}{15 \times (225 - 1)} = (0.8920)$$

In this case the value is very close to that of the Pearson correlation coefficient. For n> 10, the Spearman rank correlation coefficient can be tested for significance using the t test given earlier.

## The regression equation

Correlation describes the strength of an association between two variables, and is completely symmetrical, the correlation between A and B is the same as the correlation between B and A. However, if the two variables are related it means that when one changes by a certain amount the other changes on an average by a certain amount. For instance, in the children described earlier greater height is associated, on average, with greater anatomical dead Space. If y represents the dependent variable and x the

independent variable, this relationship is described as the regression of y on x.

The relationship can be represented by a simple equation called the regression equation. In this context "regression" (the term is a historical anomaly) simply means that the average value of y is a "function" of x, that is, it changes with x.

The regression equation representing how much y changes with any given change of x can be used to construct a regression line on a scatter diagram, and in the simplest case this is assumed to be a straight line. The direction in which the line slopes depends on whether the correlation is positive or negative. When the two sets of observations increase or decrease together (positive) the line slopes upwards from left to right; when one set decreases as the other increases the line slopes downwards from left to right. As the line must be straight, it will probably pass through few, if any, of the dots. Given that the association is well described by a straight line we have to define two features of the line if we are to place it correctly on the diagram. The first of these is its distance above the baseline; the second is its slope. They are expressed in the following *Regression equation* :

With this equation we can find a series of values of $y_{fit}$ the variable, that correspond to each of a series of values of x, the independent variable. The parameters α and β have to be estimated from the data. The parameter signifies the distance above the baseline at which the regression line cuts the vertical (y) axis; that is, when y = 0. The parameter β (the *regression coefficient*) signifies the amount by which change in x must be multiplied to give the corresponding average change in y, or the amount y changes for a unit increase in x. In this way it represents the degree to which the line slopes upwards or downwards.

The regression equation is often more useful than the correlation coefficient. It enables us to predict y from x and gives us a better summary of the relationship between the two variables. If, for a particular value of x, x i, the regression equation predicts a value of y fit , the prediction error is $y_1 - y_{fit}$ . It can easily be shown that any straight line passing through the mean values x and y will give a total prediction error $\Sigma(y_1 - y_{fit})$ of zero because the positive and negative terms exactly cancel. To remove the negative signs we

square the differences and the regression equation chosen to minimise the sum of squares of the prediction errors, $S^2 = \Sigma(y_1 - y_{fit})^2$ We denote the sample estimates of Alpha and Beta by a and b. It can be shown that the one straight line that minimises $S_2$, the least squares estimate, is given by

$$b = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

and

$$a = \bar{y} - b\bar{x}$$

it can be shown that

$$b = \frac{\Sigma xy - n\bar{x}\bar{y}}{(n-1)SD(x)^2}$$

which is of use because we have calculated all the components of equation (11.2) in the calculation of the correlation coefficient.

The calculation of the correlation coefficient on the data in table 11.2 gave the following:

$\Sigma xy = 150605, SD(x) = 19.3679, \bar{y} = 66.93, \bar{x} = 144.6$

Applying these figures to the formulae for the regression coefficients, we have:

$$b = \frac{150605 - 15 \times 66.93 \times 144.6}{14 \times 19.3679^2} = \frac{5426.6}{5251.6} = 1.033 ml/cm$$

$$a = 66.39 - (1.033 \times 144.6) = -82.4$$

Therefore, in this case, the equation for the regression of y on x becomes

$$y = -82.4 + 1.033x$$

This means that, on average, for every increase in height of 1 cm the increase in anatomical dead space is 1.033 ml *over the range of measurements made*.

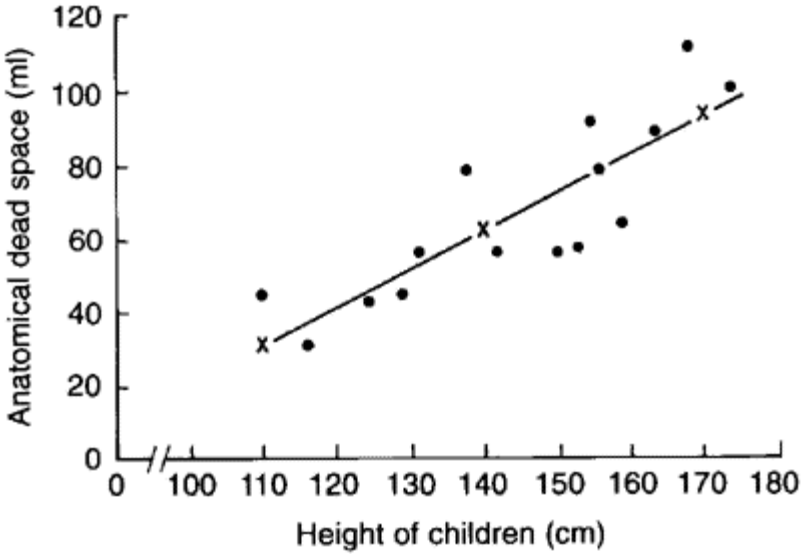
The line representing the equation is shown superimposed on the scatter diagram of the data in figure 11.2. The way to draw the line is to take three values of x, one on the left side of the scatter diagram, one in the middle and one on the right, and substitute these in the equation, as follows:

If x = 110, y = (1.033 x 110) - 82.4 = 31.2

If x = 140, y = (1.033 x 140) - 82.4 = 62.2

If x = 170, y = (1.033 x 170) - 82.4 = 93.2

Although two points are enough to define the line, three are better as a check. Having put them on a scatter diagram, we simply draw the line through them.




Regression line drawn on scatter diagram relating height and pulmonary anatomical dead space in 15 children

## Uses of Correlation and Regression

There are three main uses for correlation and regression.

- One is to test hypotheses about cause-and-effect relationships. In this case, the experimenter determines the values of the X-variable and sees whether variation in X causes variation in Y. For example, giving people different amounts of a drug and measuring their blood pressure.
- The second main use for correlation and regression is to see whether two variables are associated, without necessarily inferring a cause-and-effectrelationship. In this case, neither variable is determined by the experimenter; both are naturally variable. If an association is found, the inference is that variation in X may cause variation in Y, or variation in Y may cause variation in X, or variation in some other factor may affect both X and Y.
- The third common use of linear regression is estimating the value of one variable corresponding to a particular value of the other variable.

**Correlation** is a bivariate analysis that measures the strengths of association between two variables. In statistics, the value of the correlation coefficient varies between +1 and -1. When the value of the correlation coefficient lies around ± 1, then it is said to be a perfect degree of association between the two variables. As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker. Usually, in statistics, we measure three types of correlations: Pearson correlation, Kendall rank correlation and Spearman correlation.

***Pearson r correlation:*** Pearson r correlation is widely used in statistics to measure the degree of the relationship between linear related variables. For example, in the stock market, if we want to measure how two commodities are related to each other, Pearson *r* correlation is used to measure the degree of relationship between the two commodities. The following formula is used to calculate the Pearson r correlation:

$$r = \frac{N \sum xy - \sum (x)(y)}{\sqrt{N \sum x^2 - \sum (x^2)][N \sum y^2 - \sum (y^2)]}}$$

Where:
r = Pearson r correlation coefficient
N = number of value in each data set
∑xy = sum of the products of paired scores
∑x = sum of x scores
∑y = sum of y scores
∑x2= sum of squared x scores
∑y2= sum of squared y scores

## *Spearman rank correlation*:

Spearman rank correlation is a non-parametric test that is used to measure the degree of association between two variables. It was developed by Spearman, thus it is called the Spearman rank correlation. Spearman rank correlation test does not assume any assumptions about the distribution of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal.
The following formula is used to calculate the Spearman rank correlation:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

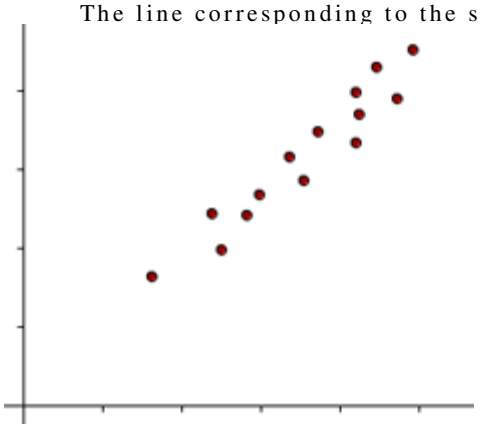
Where:
P= Spearman rank correlation
di= the difference between the ranks of corresponding values Xi and Yi
n= number of value in each data set

## Types of Correlation:

### Positive Correlation

Positive correlation occurs when an increase in one variable increases the value in another.

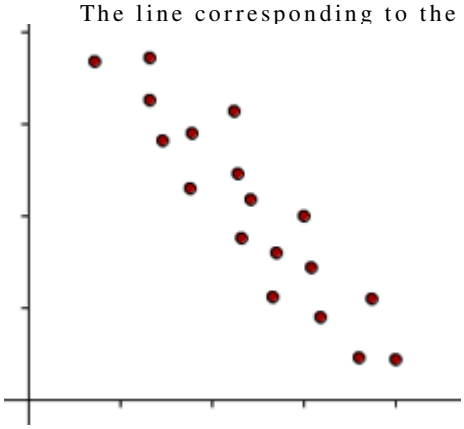
The line corresponding to the scatter plot is an increasing line.



### Negative Correlation

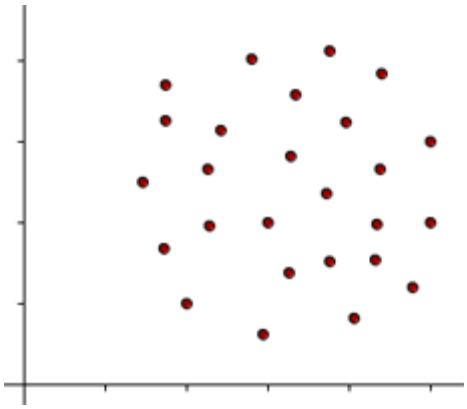
Negative correlation occurs when an increase in one variable decreases the value of another.

The line corresponding to the scatter plot is a decreasing line.



### *No Correlation*

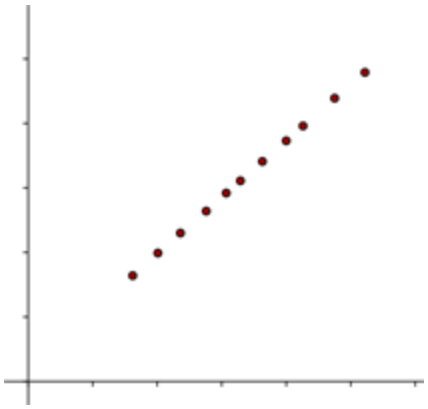
No correlation occurs when there is no linear dependency between the variables.



### *Perfect Correlation*

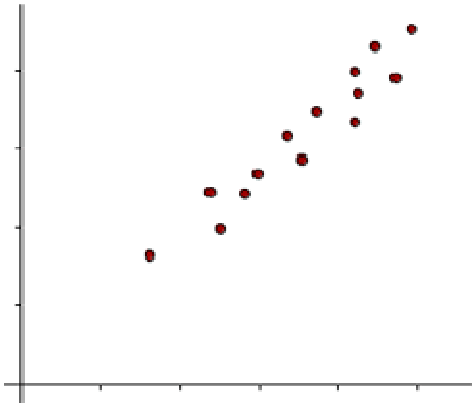
Perfect correlation occurs when there is a funcional dependency between the variables.

In this case all the points are in a straight line.



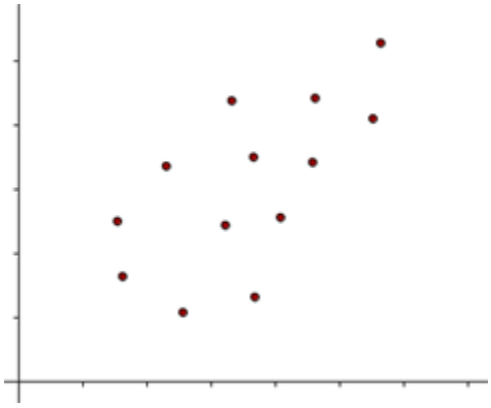
### *Strong Correlation*

A correlation is stronger the closer the points are located to one another on the line.



### *Weak Correlation*

A correlation is weaker the farther apart the points are located to one another on the line.



## Regression

Regression analysis is a statistical tool for the investigation of relationships betweenvariables.In **statistical** modeling, **regressionanalysis** is **statistical** process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors').

Regression coefficients represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant.

Regression analysis is a form of predictive modelling technique which investigates the relationship between a **dependent** (target) and **independent variable (s)**(predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

Regression analysis is an important tool for modelling and analyzing data. Here, we fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized.

There are multiple benefits of using regression analysis. They are as follows:

1. It indicates the **significant relationships** between dependent variable and independent variable.
2. It indicates the **strength of impact** of multiple independent variables on a dependent variable.

Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities. These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.

**Regression is a statistical measure that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independentvariables).**

Differences Between Correlation and Regression

The points given below, explains the difference between correlation and regression in detail:

1. A statistical measure which determines the co-relationship or association of two quantities is known as Correlation. Regression describes how an independent variable is numerically related to the dependent variable.
2. Correlation is used to represent the linear relationship between two variables. On the contrary, regression is used to fit the best line and estimate one variable on the basis of another variable.
3. In correlation, there is no difference between dependent and independent variables i.e. correlation between x and y is similar to y and x. Conversely, the regression of y on x is different from x on y.
4. Correlation indicates the strength of association between variables. As opposed to, regression reflects the impact of unit change in the independent variable on the dependent variable.

5. Correlation aims at finding a numerical value that expresses the relationship between variables. Unlike regression whose goal is to predict values of the random variable on the basis of the values of fixed variable.

   What is the line of regression?

   Linear **Regression**. Linear **regression** attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

   **A regression line is simply a single line that best fits the data (in terms of having the smallest overall distance from the line to the points). Statisticians call this technique for finding the best-fitting line a*simple linear regression analysis using the least squares method.***
   The formula for the *best-fitting line* (or *regression line*) is $y = mx + b$, where $m$ is the slope of the line and $b$ is the $y$-intercept.

   - The *slope* of a line is the change in $Y$ over the change in $X$. For example, a slope of

     $10/3$

     means as the $x$-value increases (moves right) by 3 units, the $y$-value moves up by 10 units on average.

   - The *y-intercept* is the value on the y-axis where the line crosses. For example, in the equation $y=2x - 6$, the line crosses the $y$-axis at the value $b= -6$. The coordinates of this point are $(0, -6)$; when a line crosses the $y$-axis, the $x$-value is always 0.

   To save a great deal of time calculating the best fitting line, first find the "big five," five summary statistics that you'll need in your calculations:

   1. **The mean of the *x* values**

(denoted by $\bar{x}$)

2. **The mean of the *y* values**

   (denoted by $\bar{y}$)

3. **The standard deviation of the *x* values (denoted $s_x$)**

4. **The standard deviation of the *y* values (denoted $s_y$)**

5. **The correlation between *X* and *Y* (denoted *r*)**

**Regression coefficient** -

when the regression line is linear (y = ax + b) the regressioncoefficient is the const ant (a) that represents the rate of change of one variable (y) as afunction of changes in the other (x); it is the slope of the regression line

# TIME SERIES AND INDEX NUMBERS

## TIME SERIES:

A time series is a collection of observations of well-defined data items obtained through repeated measurements over time. For example, measuring the value of retail sales each month of the year would comprise a time series. This is because sales revenue is well defined, and consistently measured at equally spaced intervals. Data collected irregularly or only once are not time series.

A time series is a sequence of numerical data points in successive order, usually occurring in uniform intervals. In plain English, a time series is simply a sequence of numbers collected at regular intervals over a period of time.

Time series analysis can be useful to see how a given asset, security or economic variable changes over time or how it changes compared to other variables over the same time period. For example, suppose you wanted to analyze a time series of daily closing stock prices for a given stock over a period of one year. You would obtain a list of all the closing prices for the stock over each day for the past year and list them in chronological order. This would be a one-year, daily closing price time series for the stock. The set of data depend on time is called time series.

If any data is collected on base of time and make a series that series is called time series.

**Importance of Time series:-**

There are many benefits of time series which can be written by us for business purposes

**♠ Helpful for study of past behaviour**

Time series are very helpful in study of past behaviour of business .On this

basis , we can invest our money in that type of business. It is duty of businessman to make time series of past sale or profit and see what is the trend of sale or profit in that type of business.

**♠ Helpful in forecasting**

Forecasting is science of estimation. Today is the day of competition so if you have to win from competition then you must learn this science , this science can be utilized if we make time series and on the basis we can read the history and then we can decide what happen in future . Suppose if we can make the time series of past strategy of our competitor then on this basis we can estimate future strategy of our competitor and on this base we can change our strategy for defeating our competitor.

**♠ Helpful in evaluating the achievements:-**

Time series is an equipment in your hand on this basis you can evaluate your business achievements if you did good , your performance shows your good face in the time series by up-word trend of your performance. If your business performance is very bad then you can make new policies to stable your business.

**♠ Helpful in comparison:-**

If we can calculate our two or more branches time series then we can compare the performance of our branches. On their performance we can give them rewards..

**Components of Time Series:**

1. Secular Trend
2. Seasonal Movements
3. Cyclical Movements
4. Irregular Fluctuations

**Secular Trend:**

The secular trend is the main component of a time series which results from long term effect of socio-economic and political factors. This trend may show the growth or decline in a time series over a long period. This is the type of tendency which continues to persist for a very long period. Prices, export and imports data, for example, reflect obviously increasing tendencies over time.

**Seasonal Trend:**

These are short term movements occurring in a data due to seasonal factors. The short term is generally considered as a period in which changes occur in a time series with variations in weather or festivities. For example, it is commonly observed that the consumption of ice-cream during summer us generally high and hence sales of an ice-cream dealer would be higher in some months of the year while relatively lower during winter months. Employment, output, export etc. are subjected to change due to variation in weather. Similarly sales of garments, umbrella, greeting cards and fire-work are subjected to large variation during festivals like Valentine's Day, Eid, Christmas, New Year etc. These types of variation in a time series are isolated only when the series is provided biannually, quarterly or monthly.

**Cyclic Movements:**

These are long term oscillation occurring in a time series. These oscillations are mostly observed in economics data and the periods of such oscillations are generally extended from five to twelve years or more. These oscillations are associated to the well known business cycles. These cyclic movements can be studied provided a long series of measurements, free from irregular fluctuations is available.

**Irregular Fluctuations:**

These are sudden changes occurring in a time series which are unlikely to be repeated, it is that component of a time series which cannot be explained by trend, seasonal or cyclic

movements .It is because of this fact these variations some-times called residual or random component. These variations though accidental in nature, can cause a continual change in the trend, seasonal and cyclical oscillations during the forthcoming period. Floods, fires, earthquakes, revolutions, epidemics and strikes etc,. are the root cause of such irregularities.

## INDEX NUMBERS

Index numbers measure the value of an item (or group of items) at a particular point in time, as a percentage of the value of an item (or group of items) at another point in time. They are commonly used in business and economics as indicators of changing business or economic activity. There are many kinds of index numbers, including price indices, quantity indexes, value indexes, and sociological indexes.

Some prominent definitions, given by statisticians, are given below: **According to the Spiegel** : "An index number is a statistical measure, designed to measure changes in a variable, or a group of related variables with respect to time, geographical location or other characteristics such as income, profession, etc." **According to Patterson** : " In its simplest form, an index number is the ratio of two index numbers expressed as a percent . An index is a statistical measure, a measure designed to show changes in one variable or a group of related variables over time, with respect to geographical location or other characteristics". **According to Tuttle** : "Index number is a single ratio (or a percentage) which measures the combined change of several variables between two different times, places or situations". We can thus say that index numbers are economic barometers to judge the inflation ( increase in prices) or deflationary (decrease in prices )

tendencies of the economy. They help the government in adjusting its policies in case of inflationary situations.

## CHARACTERISTICS OF INDEX NUMBERS:

Following are some of the important characteristics of index numbers: · Index numbers are expressed in terms of percentages to show the extent of relative change

- Index numbers measure relative changes.
- They measure the relative change in the value of a variable or a group of related variables over a period of time or between places. ·
- Index numbers measures changes which are not directly measurable.
- The cost of living, the price level or the business activity in a country are not directly measurable but it is possible to study relative changes in these activities by measuring the changes in the values of variables/factors which affect these activities.

## PROBLEMS IN THE CONSTRUCTION OF INDEX NUMBERS:

The decisions regarding the following problems/aspect have to be taken before starting the actual construction of any type of index numbers.

(i) Purpose of Index numbers under construction (ii) Selection of items (iii) Choice of an appropriate average (iv) Assignment of weights (importance) (v) Choice of base period. Let us discuss these one-by-one

### 1 Purpose of Index Numbers

An index number, which is designed keeping, specific objective in mind, is a very powerful tool. For example, an index whose purpose is to measure consumer price index, should not include wholesale rates of items and the index number meant for slum-colonies should not consider luxury items like A.C., Cars refrigerators, etc.

### 2. Selections of Items

After the objective of construction of index numbers is defined, only those items which are related to and are relevant with the purpose should be included.

**3. Choice of Average**

As index numbers are themselves specialised averages, it has to be decided first as to which average should be used for their construction. The arithmetic mean, being easy to use and calculate, is preferred over other averages (median, mode or geometric mean). In this lesson, we will be using only arithmetic mean for construction of index numbers.

**4. Assignment of weights**

Proper importance has to be given to the items used for construction of index numbers. It is universally agreed that wheat is the most important cereal as against other cereals, and hence should be given due importance.

**5 .Choice of Base year**

The index number for a particular future year is compared against a year in the near past, which is called base year. It may be kept in mind that the base year should be a normal year and economically stable year.

**TYPES OF INDEX NUMBERS:**

Index numbers are names after the activity they measure. Their types are as under:

**Price Index**:

Measure changes in price over a specified period of time. It is basically the ratio of the price of a certain number of commodities at the present year as against base year.

**Quantity Index:**

As the name suggest, these indices pertain to measuring changes in volumes of commodities like goods produced or goods consumed, etc.

**Value Index:**

These pertain to compare changes in the monetary value of imports, exports, production or consumption of commodities.

## CONSTRUCTION OF INDEX NUMBERS

Suppose one is interested in comparing the sum total of expenditure on a fixed number of commodities in the year 2003 as against the year 1998. Let us consider the following example.

The price in a selected year (such as 2003) is divided by the price in the base year. The base-period price is designated as p0, and a price other than the base period is often referred to as the given period or selected period and designated pt . To calculate the simple price index P using 100 as the base value for any given period use the formula:

$$\text{SIMPLE INDEX} \quad P = \frac{P_1}{P_0} \times 100$$

## . USES OF INDEX NUMBERS:

**i)** Index numbers are economic barometers. They measure the level of business and economic activities and are therefore helpful in gauging the economic status of the country.

**ii)** (ii) Index numbers measure the relative change in a variable or a group of related variable(s) under study.

**iii)** (iii) Consumer price indices are useful in measuring the purchasing power of money, thereby used in compensating the employees in the form of increase of allowances.