Unit- 1 Introduction to biostatistics

Biostatistics (a contraction of biology and statistics; sometimes referred to as **biometry** or **biometrics**) is the application of statistics to a wide range of topics in biology. The science of biostatistics encompasses the design of biological experiments, especially in medicine and agriculture; the collection, summarization, and analysis of data from those experiments; and the interpretation of, and inference from, the results.

INTRODUCTION

- Statistics plays a vitally important role in the research.
- Health information is very often explained in statistical terms
- Many decisions in the Health Sciences are created through statistical studies
- It enables you:

  - to read and evaluate reports and other literature
  - to take independent research investigations
  - to describe the data in meaningful terms

Applications of biostatistics

- Public health, including epidemiology, health services research, nutrition, and environmental health
- Design and analysis of clinical trials in medicine
- Population genetics, and statistical genetics in order to link variation in genotype with a variation in phenotype. This has been used in agriculture to improve crops and farm animals (animal breeding). In biomedical research, this work can assist in finding candidates for gene alleles that can cause or influence predisposition to disease in human genetics
- Analysis of genomics data, for example from microarray or proteomics experiments. Often concerning diseases or disease stages.
- Ecology, ecological forecasting
- Biological sequence analysis
- Systems biology for gene network inference or pathways analysis.

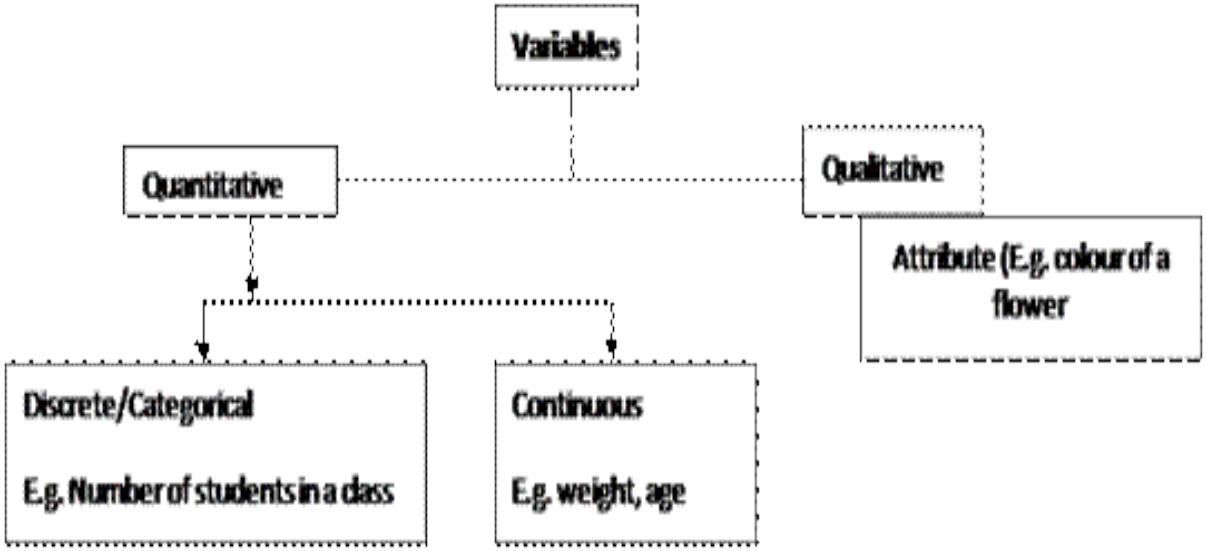
Statistical methods are beginning to be integrated into medical informatics, public health informatics, bioinformatics and computational biology.

DEFINITIONS

1. **Statistics:** is the study of how to collect, organizes, analyze, and interpret data.
2. **Data:** the values recorded in an experiment or observation.
3. **Population:** refers to any collection of individual items or units that are the subject of investigation.
4. **Sample:** A small representative sample of a population is called sample.
5. **Observation:** each unit in the sample provides a record, as a measurement which is called observation.
6. **Sampling:** getting sample from a population
7. **Variable:** the value of an item or individual is called variable
8. **Raw Data**: Data collected in original form.
9. **Frequency**: The number of times a certain value or class of values occurs.
10. **Tabulation:** can be defined as the logical and systematic arrangement of statistical data in rows and columns.
11. **Frequency Distribution:** The organization of raw data in table form with classes and frequencies.
12. **Class Limits:** Separate one class in a grouped frequency distribution from another. The limits could actually appear in the data and have gaps between the upper limit of one class and the lower limit of the next.
13. **Class Boundaries**: Separate one class in a grouped frequency distribution from another.
14. **Cumulative Frequency**: The number of values less than the upper class boundary for the current class. This is a running total of the frequencies.
15. **Histogram**: A graph which displays the data by using vertical bars of various heights to represent frequencies.
16. **Frequency Polygon**: it is a line graph. The frequency is placed along the vertical axis and the class midpoints are placed along the horizontal axis. These points are connected with lines.
17. **Pie Chart**: Graphical depiction of data as slices of a pie. The frequency determines the size of the slice. The number of degrees in any slice is the relative frequency times 360 degrees.
18. **Central tendency -** a typical or representative value for a dataset.

VARIABLES

- The value of an item or individual is called variable.
- Variables are of two types:
    - **Quantitative**: a variable with a numeric value. E.g. age, weight.
    - **Qualitative**: a variable with a category or group value. E.g. Gender (M/F), Religion (H/M/C), Qualification (degree/PG)
- Quantitative variable are two types:
    - Discrete /categorical variables
    - Continuous variables



- Variables can be
    - **Independent**
        - Are not influenced by other variables.
        - Are not influenced by the event, but could influence the event.
    - **Dependent**
        - The variable which is influenced by the others is often referred as dependent variable.

E.g. In an experimental study on relaxation intervention for reducing HTN, blood pressure is the dependent variable and relaxation training, age and gender are independent variable.

SAMPLING

- Sampling is the process of getting a representative fraction of a population.
- Analysis of the sample gives an idea of the population.
- Methods of sampling:
    - ❖ Random Sampling or Probability sampling
        - Simple random sampling
        - Stratified random Sampling
        - Cluster sampling
    - ❖ Non-random sampling
        - Convenient Sampling
        - Purposive Sampling
        - Quota Sampling

- In **Simple Random sampling,** each individual of the population has an equal chance of being included in the sample. Two methods are used in simple random sampling:
- Random Numbers method
- Lottery method
- In **stratified random sampling**, the population is divide in to groups or strata on the basis of certain characteristics.
- In **cluster sampling**, the whole population is divided in to a number of relatively small cluster groups. Then some of the clusters are randomly selected.
- **Convenience sampling** is a type of non-probability sampling which involves the sample being drawn from that part of the population which is selected because it is readily available and convenient.
- **Purposive sampling** is a type of non-probability sampling in which researcher selects participants based on fulfillment of some criteria. E.g. schizophrenia treatment naive.

SCALES OF MEASUREMENT

- Four measurement scales are used: nominal, ordinal, interval and ratio.
- Each level has its own rules and restrictions.

Nominal Scale of measurement

- Nominal variables include categories of people, events, and other phenomena are named.
- Example: gender, age-class, religion, type of disease, blood groups A, B, AB, and O.
- They are exhaustive in nature, and are mutually exclusive.
- These categories are discrete and non-continuous.
    - Statistical operations permissible are: counting of frequency, Percentage, Proportion, mode, and coefficient of contingency.

Ordinal Scale of measurement

- It is second in terms of its refinement as a means of classifying information.
- It incorporates the functions of nominal scale.
- The ordinal scale is used to arrange (or rank) individuals into a sequence ranging from the highest to lowest.
- Ordinal implies rank-ordered from highest to lowest.
    - Grade A+, A, B+, B, C+, C
    - 1st , 2nd , 3rd etc

Interval scale of Measurement

- Interval scale refers to the third level of measurement in relation to complexity of statistical techniques used to analyze data.
- It is quantitative in nature
- The individual units are equidistant from one point to the other.
- The interval data does not have an absolute zero.
    - E.g. temperature is measured in Celsius or Fahrenheit.

Ratio Scale of Measurement

- Equal distances between the increments
- This scale has an absolute zero.

- Ratio variables exhibit the characteristics of ordinal and interval measurement
- E.g. variable like time, length and weight are ratio scales and also be measured using nominal or ordinal scale.

[The mathematical properties of interval and ratio scales are very similar, so the statistical procedures are common for both the scales.

In statistics and survey methodology, **sampling** is concerned with the selection of a subset of individuals from within a population to estimate characteristics of the whole population.

Researchers rarely survey the entire population because the cost of a census is too high. The three main advantages of sampling are that the cost is lower, data collection is faster, and since the data set is smaller it is possible to ensure homogeneity and to improve the accuracy and quality of the data.

Each observation measures one or more properties (such as weight, location, color) of observable bodies distinguished as independent objects or individuals. In survey sampling, weights can be applied to the data to adjust for the sample design, particularly stratified sampling (blocking). Results from probability theory and statistical theory are employed to guide practice. In business and medical research, sampling is widely used for gathering information about a population.

Process of sampling

The sampling process comprises several stages:

- Defining the population of concern
- Specifying a sampling frame, a set of items or events possible to measure
- Specifying a sampling method for selecting items or events from the frame
- Determining the sample size
- Implementing the sampling plan
- Sampling and data collecting

Population definition

Successful statistical practice is based on focused problem definition. In sampling, this includes defining the population from which our sample is drawn. A population can be defined as including all people or items with the characteristic one wishes to understand. Because there is very rarely enough time or money to gather information from everyone or everything in a population, the goal becomes finding a representative sample (or subset) of that population.

Sometimes that which defines a population is obvious. For example, a manufacturer needs to decide whether a batch of material from production is of high enough quality to be released to the customer, or should be sentenced for scrap or rework due to poor quality. In this case, the batch is the population.

Note also that the population from which the sample is drawn may not be the same as the population about which we actually want information. Often there is large but not complete overlap between these two groups due to frame issues etc. (see below). Sometimes they may be entirely separate - for instance, we might study rats in order to get a better understanding of human health, or we might study records from people born in 2008 in order to make predictions about people born in 2009.

Sampling frame

In the most straightforward case, such as the sentencing of a batch of material from production (acceptance sampling by lots), it is possible to identify and measure every single item in the population and to include any one of them in our sample. However, in the more general case this is not possible. There is no way to identify all rats in the set of all rats

Probability and nonprobability sampling

A **probability sampling** scheme is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined. Probability sampling

includes: Simple Random Sampling, Systematic Sampling, Stratified Sampling, Probability Proportional to Size Sampling, and Cluster or Multistage Sampling.

Nonprobability sampling is any sampling method where some elements of the population have *no* chance of selection (these are sometimes referred to as 'out of coverage'/'undercovered'), or where the probability of selection can't be accurately determined. It involves the selection of elements based on assumptions regarding the population of interest, which forms the criteria for selection. Hence, because the selection of elements is nonrandom, nonprobability sampling does not allow the estimation of sampling errors. These conditions give rise to exclusion bias, placing limits on how much information a sample can provide about the population. Information about the relationship between sample and population is limited, making it difficult to extrapolate from the sample to the population.

Nonprobability Sampling includes: Accidental Sampling, Quota Sampling and Purposive Sampling.

Sampling methods / Techniques

Within any of the types of frame identified above, a variety of sampling methods can be employed, individually or in combination. Factors commonly influencing the choice between these designs include:

1. Nature and quality of the frame
2. Availability of auxiliary information about units on the frame
3. Accuracy requirements, and the need to measure accuracy
4. Whether detailed analysis of the sample is expected
5. Cost/operational concerns

1. **Simple random sampling**
2. **Systematic sampling**
3. **Stratified sampling**
4. **Probability proportional to size sampling**
5. **Cluster sampling**
6. **Quota sampling**
7. **Convenience sampling or Accidental Sampling**
8. **Line-intercept sampling**
9. **Panel sampling**

1. Simple random sampling

In a simple random sample ('SRS') of a given size, all such subsets of the frame are given an equal probability. Each element of the frame thus has an equal probability of selection: the frame is not subdivided or partitioned. Furthermore, any given *pair* of elements has the same chance of selection as any other such pair (and similarly for triples, and so on). This minimises bias and simplifies analysis of results. In particular, the variance between individual results within the sample is a good indicator of variance in the overall population, which makes it relatively easy to estimate the accuracy of results.

However, SRS can be vulnerable to sampling error because the randomness of the selection may result in a sample that doesn't reflect the makeup of the population. For instance, a simple random sample of ten people from a given country will *on average* produce five men and five women, but any given trial is likely to overrepresent one sex and underrepresent the other. Systematic and stratified techniques, discussed below, attempt to

overcome this problem by using information about the population to choose a more representative sample.

2. Systematic sampling

Systematic sampling relies on arranging the target population according to some ordering scheme and then selecting elements at regular intervals through that ordered list. Systematic sampling involves a random start and then proceeds with the selection of every *k*th element from then onwards. In this case, *k*=(population size/sample size). It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first to the *k*th element in the list. A simple example would be to select every 10th name from the telephone directory (an 'every 10th' sample, also referred to as 'sampling with a skip of 10').

As long as the starting point is randomized, systematic sampling is a type of probability sampling. It is easy to implement and the stratification induced can make it efficient, *if* the variable by which the list is ordered is correlated with the variable of interest. 'Every 10th' sampling is especially useful for efficient sampling from databases.

3. Stratified sampling

Where the population embraces a number of distinct categories, the frame can be organized by these categories into separate "strata." Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected.[1] There are several potential benefits to stratified sampling.

First, dividing the population into distinct, independent strata can enable researchers to draw inferences about specific subgroups that may be lost in a more generalized random sample.

Second, utilizing a stratified sampling method can lead to more efficient statistical estimates (provided that strata are selected based upon relevance to the criterion in question, instead of availability of the samples). Even if a stratified sampling approach does not lead to increased statistical efficiency, such a tactic will not result in less efficiency than would simple random sampling, provided that each stratum is proportional to the group's size in the population.

Third, it is sometimes the case that data are more readily available for individual, pre-existing strata within a population than for the overall population; in such cases, using a stratified sampling approach may be more convenient than aggregating data across groups (though this may potentially be at odds with the previously noted importance of utilizing criterion-relevant strata).

Finally, since each stratum is treated as an independent population, different sampling approaches can be applied to different strata, potentially enabling researchers to use the approach best suited (or most cost-effective) for each identified subgroup within the population.

4. Probability proportional to size sampling

In some cases the sample designer has access to an "auxiliary variable" or "size measure", believed to be correlated to the variable of interest, for each element in the population. These data can be used to improve accuracy in sample design. One option is to use the auxiliary variable as a basis for stratification, as discussed above.

Another option is probability-proportional-to-size ('PPS') sampling, in which the selection probability for each element is set to be proportional to its size measure, up to a maximum of 1. In a simple PPS design, these selection probabilities can then be used as the basis for Poisson sampling. However, this has the drawback of variable sample size, and different portions of the population may still be over- or under-represented due to chance variation in selections. To address this problem, PPS may be combined with a systematic approach.

5. Cluster sampling

Sometimes it is more cost-effective to select respondents in groups ('clusters'). Sampling is often clustered by geography, or by time periods. (Nearly all samples are in some sense 'clustered' in time - although this is rarely taken into account in the analysis.) For instance, if surveying households within a city, we might choose to select 100 city blocks and then interview every household within the selected blocks.

Clustering can reduce travel and administrative costs. In the example above, an interviewer can make a single trip to visit several households in one block, rather than having to drive to a different block for each household.

It also means that one does not need a sampling frame listing all elements in the target population. Instead, clusters can be chosen from a cluster-level frame, with an element-level frame created only for the selected clusters. In the example above, the sample only requires a block-level city map for initial selections, and then a household-level map of the 100 selected blocks, rather than a household-level map of the whole city.

6. Quota sampling

In **quota sampling**, the population is first segmented into mutually exclusive sub-groups, just as in stratified sampling. Then judgment is used to select the subjects or units from each segment based on a specified proportion. For example, an interviewer may be told to sample 200 females and 300 males between the age of 45 and 60.

7. Convenience sampling or Accidental Sampling

Convenience sampling (sometimes known as **grab** or **opportunity sampling**) is a type of nonprobability sampling which involves the sample being drawn from that part of the population which is close to hand. That is, a population is selected because it is readily available and convenient. It may be through meeting the person or including a person in the sample when one meets them or chosen by finding them through technological means such as the internet or through phone. The researcher using such a sample cannot scientifically make generalizations about the total population from this sample because it would not be representative enough. For example, if the interviewer were to conduct such a survey at a shopping center early in the morning on a given day, the people that he/she could interview would be limited to those given there at that given time, which would not represent the views of other members of society in such an area, if the survey were to be conducted at different times of day and several times per week. This type of sampling is most useful for pilot testing.

8. Line-intercept sampling

Line-intercept sampling is a method of sampling elements in a region whereby an element is sampled if a chosen line segment, called a "transect", intersects the element.

9. Panel sampling

Panel sampling is the method of first selecting a group of participants through a random sampling method and then asking that group for the same information again several times over a period of time. Therefore, each participant is given the same survey or interview at two or more time points; each period of data collection is called a "wave".

Replacement of selected units

Sampling schemes may be *without replacement* ('WOR' - no element can be selected more than once in the same sample) or *with replacement* ('WR' - an element may appear multiple times in the one sample). For example, if we catch fish, measure them, and immediately return them to the water before continuing with the sample, this is a WR design, because we might end up catching and measuring the same fish more than once. However, if we do not return the fish to the water (e.g. if we eat the fish), this becomes a WOR design.

Sample size

Formulas, tables, and power function charts are well known approaches to determine sample size.

❖ Sampling and data collection

Good data collection involves:

- Following the defined sampling process
- Keeping the data in time order
- Noting comments and other contextual events
- Recording non-responses

Sampling errors and biases

Sampling errors and biases are induced by the sample design. They include:

1. **Selection bias**: When the true selection probabilities differ from those assumed in calculating the results.
2. **Random sampling error**: Random variation in the results due to the elements in the sample being selected at random.

Non-sampling error

Non-sampling errors are caused by other problems in data collection and processing. They include:

1. **Overcoverage**: Inclusion of data from outside of the population.
2. **Undercoverage**: Sampling frame does not include elements in the population.
3. **Measurement error**: E.g. when respondents misunderstand a question, or find it difficult to answer.
4. **Processing error**: Mistakes in data coding.
5. **Non-response**: Failure to obtain complete data from all selected individuals.

Sampling bias

In statistics, **sampling bias** is when a sample is collected in such a way that some members of the intended population are less likely to be included than others. It results in a **biased sample**, a non-random sample[1] of a population (or non-human factors) in which all individuals, or instances, were not equally likely to have been selected.[2] If this is not accounted for, results can be erroneously attributed to the phenomenon under study rather than to the method of sampling.

Medical sources sometimes refer to sampling bias as **ascertainment bias**.[3][4] Ascertainment bias has basically the same definition,[5][6] but is still sometimes classified as a separate type of bias. selection bias and sampling bias are often used synonymously

Types of sampling bias

- Selection from a **specific area**. For example, a survey of high school students to measure teenage use of illegal drugs will be a biased sample because it does not include home-schooled students or dropouts. A sample is also biased if certain members are underrepresented or overrepresented relative to others in the population. For example, a "man on the street" interview which selects people who walk by a certain location is going to have an overrepresentation of healthy individuals who are more likely to be out of the home than individuals with a chronic illness. This may be an extreme form of biased sampling, because certain members of the population are totally excluded from the sample (that is, they have zero probability of being selected).
- **Self-selection** bias, which is possible whenever the group of people being studied has any form of control over whether to participate. Participants' decision to participate may be correlated with traits that affect the study, making the participants a non-representative sample. For example, people who have strong opinions or substantial knowledge may be more willing to spend time answering a survey than those who do not. Another example is online and phone-in polls, which are biased samples because the respondents are self-selected. Those individuals who are highly motivated to respond, typically individuals who have strong opinions, are overrepresented, and individuals that are indifferent or apathetic are less likely to respond. This often leads to a polarization of responses with extreme perspectives being given a disproportionate weight in the summary. As a result, these types of polls are regarded as unscientific.
- **Pre-screening** of trial participants, or **advertising** for volunteers within particular groups. For example a study to "prove" that smoking does not affect fitness might recruit at the local fitness center, but advertise for smokers during the advanced aerobics class, and for non-smokers during the weight loss sessions.
- **Exclusion** bias results from exclusion of particular groups from the sample, e.g. exclusion of subjects who have recently migrated into the study area (this may occur when newcomers are not available in a register used to identify the source population). Excluding subjects who move out of the study area during follow-up is rather equivalent of dropout or nonresponse, a selection bias in that it rather affects the internal validity of the study.

- **Healthy user bias**, when the study population is likely healthier than the general population, e.g. workers (i.e. someone in ill-health is unlikely to have a job as manual laborer).
- **Overmatching**, matching for an apparent confounder that actually is a result of the exposure. The control group becomes more similar to the cases in regard to exposure than the general population.

Problems caused by sampling bias

A biased sample causes problems because any statistic computed from that sample has the potential to be consistently erroneous. The bias can lead to an over- or underrepresentation of the corresponding parameter in the population

The word bias in common usage has a strong negative word connotation, and implies a deliberate intent to mislead or other scientific fraud

Statistical corrections for a biased sample

If entire segments of the population are excluded from a sample, then there are no adjustments that can produce estimates that are representative of the entire population. But if some groups are underrepresented and the degree of underrepresentation can be quantified, then sample weights can correct the bias.

For example, a hypothetical population might include 10 million men and 10 million women. Suppose that a biased sample of 100 patients included 20 men and 80 women. A researcher could correct for this imbalance by attaching a weight of 2.5 for each male and 0.625 for each female. This would adjust any estimates to achieve the same expected value as a sample that included exactly 50 men and 50 women, unless men and women differed in their likelihood of taking part in the survey.

Data

Statistics is the field of science that deals with organization, interpretation and analyzing of a data. The term statistical data refers to the

data collected form different sources through methods experiments, surveys and analysis. This data is then interpreted by statistical methods and formulae for their analysis.

There are mainly four types of statistical data:

1. Primary statistical data
2. Secondary statistical data
3. Qualitative statistical data
4. Quantitative statistical data.

Some classifications divide the data into two broad types i.e. primary and secondary and qualitative and quantitative. But in this classification each of the type is divided individually.

1. Primary Statistical Data

The primary data is the data that is collected directly and is not taken from a source. This data includes the data collected through direct interviews, surveys and experiments. Basically this data comprises of results obtained through surveys on which the statistical operations have not been applied.

The primary data is of following types:

- **Data Collected Through Personal Investigation:**

  This data is collected by the investigator or the researcher or the researching team personally. The researcher or the team of researchers collects this data through the surveys that they conduct themselves. This type of data is more accurate as it is collected directly by the researcher.

- **Data Collected Through Workers:**

  The researcher or the researching team may hire another person or a group of people for conducting a survey to collect a data.

2. The Secondary Statistical Data

The secondary statistical data is the data that is obtained by the researcher from another source. This data may be collected form various sources.

Many important and current data which was initially a primary data can be published in books, research papers, journals or can be kept in record books to be used as the secondary data for another researcher.

The primary data on which the statistical operations have been applied is also defined as the secondary data.

The secondary data can be official i.e. obtained form a ministry or a department for example the ministry of health or the department of health and sciences or the secondary data can be semi official i.e. obtained form books and journals

3. Qualitative Statistical Data

The qualitative statistical data is the data which is expressed in words rather than in numbers. In other words, the qualitative data is the data in which the measurement of a category is expressed in words. For example in a qualitative data measurement of height will be explained a tall, short or medium.

There are two types of qualitative data:

- Nominal data which is the data which includes measurements of categories such as gender, religion and sports.
- The ordinal data which has variable measurements of variable categories such as size, color and behavior.

The qualitative data only tells us about something but does not tell the extent of something.

For example if we make a survey of climate of different cities of a country we may record the climate of different cities as cold, warm or moderate, this data would simple summarize the kind of climate of the city but this would not tell us the average maximum or minimum temperature of the city which would explain the extent of the weather in the city.

4. Quantitative Statistical Data

The quantitative statistical data is the data in which the measurements are numerically expressed. For example the temperature of a city in this data would be given in accurate measurement like 25 degrees C.

The quantitative data represents measurements taken with a scale includes the variables such as temperature, weight and size that can be measured in a precise scale are expressed in their actual measurements in quantitative data

Other numeric expressions in the qualitative data can be for example: the number of people in a town, the social security numbers of the citizens or the number of deaths occurring due to a disease.

The quantitative data is more accurate than the qualitative data as it tells us the extent of something.

There are two common scales used the quantitative measures:

- The ratio scale which is used to measure variables such as age and money. The observations that can be counted are measured by this scale.
- The interval scale which is used to measure variables such as temperatures, height and weight

➢ Types of Primary Data Collection

Primary data are always collected from the source. It is collected either by the investigator himself or through his agents. There are different methods of collecting primary data. Each method has its relative merits and demerits.

1. Direct Personal observation:

2. Indirect Oral Interviews :

3. Mailed Questionnaire method:

4. Schedule Method:

5. From Local Agents:

1. Direct Personal observation:

Here the investigator directly contacts the informants, solicits their cooperation and enumerates the data. The information are collected by direct personal interviews. It is neither difficult for the enumerator nor the informants. Because both are present at the spot of data collection. This method provides most accurate information as the investigator collects them personally. But as the investigator alone is involved in the process, his personal bias may influence the accuracy of the data. So it is necessary that the investigator should be honest, unbiased and experienced. In such cases the data collected may be fairly accurate. However, the method is quite costly and time-consuming. So the method should be used when the scope of enquiry is small.

2. Indirect Oral Interviews:

This is an indirect method of collecting primary data. Here information are not collected directly from the source but by interviewing persons closely related with the problem. This method is applied to apprehend culprits in case of theft, murder etc. The informations relating to one's personal life or which the informant hesitates to reveal are better collected by this method. Here the investigator prepares 'a small list of questions relating to the enquiry. The answers (information) are collected by interviewing persons well connected with the incident. The investigator should cross-examine the informants to get correct information.

This method is time saving and involves relatively less cost. The accuracy of the information largely depends upon the integrity of the investigator. It is desirable that the investigator should be experienced and capable enough to inspire and create confidence in the informant to collect accurate data.

3. Mailed Questionnaire method:

This is a very commonly used method of collecting primary data. Here information are collected through a set of questionnaire. A questionnaire is a document prepared by the investigator containing a set of questions. These questions relate to the problem of enquiry directly or indirectly. Here first the

questionnaires are mailed to the informants with a formal request to answer the question and send them back.

For better response the investigator should bear the postal charges. The questionnaire should carry a polite note explaining the aims and objective of the enquiry, definition of various terms and concepts used there. Besides this the investigator should ensure the secrecy of the information as well as the name of the informants, if required.

Success of this method greatly depends upon the way in which the questionnaire is drafted. So the investigator must be very careful while framing the questions

4. Schedule Method:

Here the questionnaires are sent through the enumerators to collect informations. Enumerators are persons appointed by the investigator for the purpose. They directly meet the informants with the questionnaire. They explain the scope and objective of the enquiry to the informants and solicit their cooperation. The enumerators ask the questions to the informants and record their answers in the questionnaire and compile them. The success of this method depends on the sincerity and efficiency of the enumerators. So the enumerator should be sweet-tempered, good-natured, trained and well-behaved.

Schedule method is widely used in extensive studies. It gives fairly correct result as the enumerators directly collect the information. The accuracy of the information depends upon the honesty of the enumerators. They should be unbiased. This method is relatively more costly and time-consuming than the mailed questionnaire method.

5. From Local Agents:

Sometimes primary data are collected from local agents or correspondents. These agents are appointed by the sponsoring authorities. They are well conversant with the local conditions like language, communication, food habits, traditions etc. Being on the spot and well acquainted with the nature of the enquiry they are capable of furnishing reliable information.

The accuracy of the data collected by this method depends on the honesty and sincerity of the agents. Because they actually collect the information from the spot. Information from a wide area at less cost and time can be collected by this method. The method is generally used by government agencies, newspapers, periodicals etc. to collect data.

- **Types of Secondary Data Collection**

Secondary data are second hand informations. They are not collected from the source as the primary data. In other words, secondary data are those which have already been collected. So they may be relatively less accurate than the primary data. Secondary data are generally used when the time of enquiry is short and the accuracy of the enquiry can be compromised to some extent.

Secondary data are already collected informations. They might have been collected for some specific purposes. So they must be used with caution. It is generally very different to verify such information to find out inconsistencies, errors, omissions etc. Therefore scrutiny of secondary data is essential. Because the data might be inaccurate, unsuitable or inadequate. Thus it is very risky to use statistics collected by other people unless they have been thoroughly edited and found reliable, adequate and suitable for the purpose.

Secondary data can be collected from a number of sources which can broadly be classified into two categories.

i) Published sources

ii) Unpublished sources

i) Published Sources:

Mostly secondary data are collected from published sources. Some important sources of published data are the following.

1. Published reports of Central and State Governments and local bodies.

2. Statistical abstracts, census reports and other reports published by different ministries of the Government.

3. Official publications of the foreign Governments.

4. Reports and Publications of trade associations, chambers of commerce, financial institutions etc.

5. Journals, Magazines and periodicals.

6. Periodic Publications of Government organizations like Central Statistical Organization (C. S. O.), National Sample Survey Organization (NSSO).

7. Reports submitted by Economists, Research Scholars, Bureaus etc.

8. Published works of research institutions and Universities etc.

ii) Unpublished Sources:

Statistical data can also be collected from various unpublished sources. Some of the important unpublished sources from which secondary data can be collected are:

1. The research works carried out by scholars, teachers and professionals.
2. The records maintained by private firms and business enterprises. They may not like to publish the information considering them as business secret.
3. Records and statistics maintained by various departments and offices of the Central and State Governments, Corporations, Undertakings etc.

- **Methods of data presentation,**

The purpose of data presentation is to summarize data and present them in a form which is more precise, but still
gives an accurate view of the raw data. There are several ways in which data can be summarized in diagrams, and we shall classify the most important of these as:

1. Tables of numerical data
2. Graphs
3. Graphical representation to show relationships between variables
   (i) Bar graphs
   (ii) Histograms and Frequency polygons
   (iii) Ogive curves
   (iv) Pictographs
   (v) Pie charts

Pie charts, Bar charts and Pictograms show relative frequencies and Histograms show relative frequencies of continuous data

1. Tables or Tabulation:

Tabulation is the process of condensation. It is the systematic and orderly presentation of classified data in a definite form so as to elucidate the characteristics of the data. In statistical tables the numerical information is presented in such a form that the information so presented turns to be readily understandable. Tables are designed to summaries facts revealed by enquiry and to present them in such a way that all the important factors contained in the data under review are displayed. Tables tend to simplify the presentation and facilitate comparison between related facts. Tabular presentation takes the form of arranging statistical date in columns and rows. The idea of a table will be clear if we look to the different parts of a table.

❖ Construction of a table:

The preparation of a good table is an art. The purpose of tabulation must always be kept in mind before the preparation of a statistical table. **A good statistical table must contain** at least the following components.

i. Table number: A table should always be numbered for easy identification and reference in future
ii. Title of the table: A table mu8st have a suitable title. Title is the description of the contents of the table. So the title should be clear, brief and self-explanatory.

iii. Caption: Caption refers to the column headings. It explains what the column represents. A caption should be brief, concise and self-explanatory. Captions are usually written in the middle of the columns in small letters to economies space.
iv. Stubs: These refer to the headings of horizontal rows. They are at the extreme left.
v. Body: The body of the table contains the numerical information. This is the most vital part of the table. Data presented in the body arranged according to descriptions are classifications of the captions and stubs.
vi. Head note : It is a brief explanatory statement applying to all or a major part of the material in the table, and in placed below the title and enclosed in brackets.
vii. Footnote: Anything in a table that the reader may find difficult to understand form the title, captions and stubs should be explained in footnotes. If footnotes are needed, they are placed directly below the body of the table. In most cases footnotes are used to mention the source of data especially in case of secondary data.

Types of Tables:

Tables may broadly be classified into two categories:

i. Simple and complex tables; and
ii. General purpose and special purpose (or summary) tables

(i) Simple table or one-way table. In this type of table only one characteristic is shown. This is the simplest of tables. The following is the illustration of such a table:

Number of Employees in an organization According to age Group

Age (in years)	No. of Employees
Below 25	50
25-35	67
35-45	43
45-55	15
55 and above	5
	Total 180

(ii) Two-way table. Such a table shows two characteristics and is formed when either the stub or the caption is divided into two coordinate parts. The example given on page 56 illustrates the nature of such a table:

Number of Employees in an organization according to age and sex

Age (in years)	Employees		Total
	Males	Females	
Below 25	32	18	50
25-35	40	27	67
35-45	25	18	43
45-55	10	5	15
55 & above	5		5
Total	112	68	180

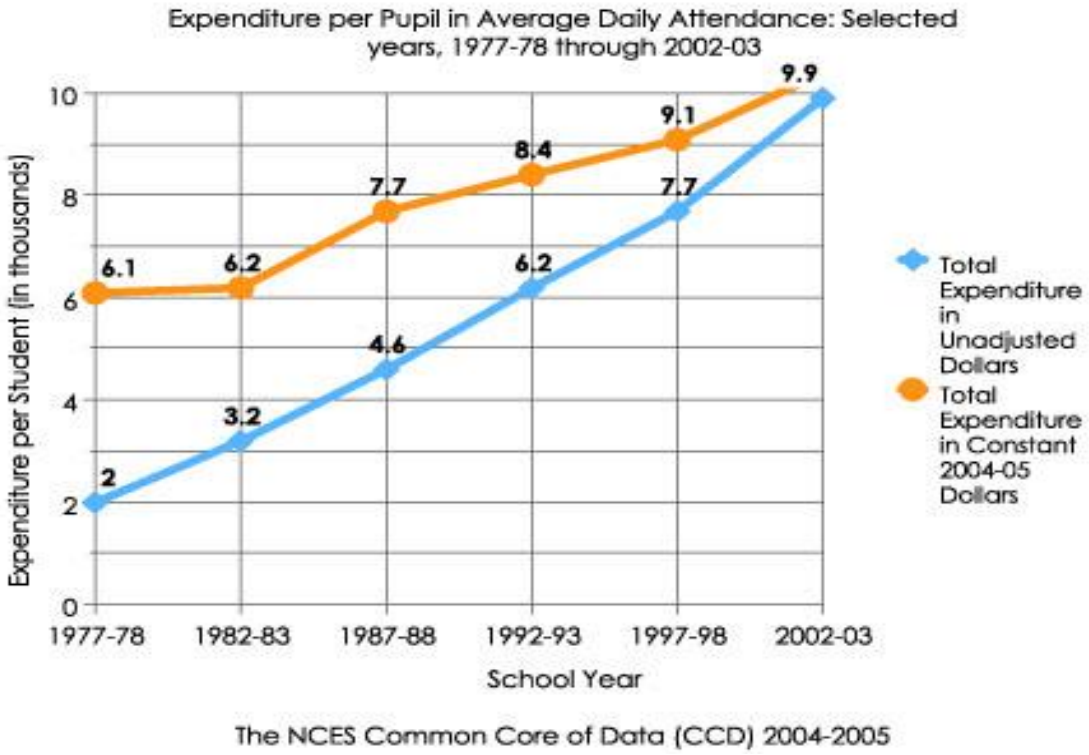
2. GRAPHS

Besides formal tables, statistical **data** can also be presented in the form of various types of graphs. Graphs are a useful way of conveying information very quickly and briefly. With the same ease and efficiency, they help in comparing data over time and space. They are visual aids and have **a** powefi impact on the people.

It is ofien said, "a ***picture is worth a thousand words".*** They attract a reader's attention to what they are supposed to convey about the data. Further, they may help us to estimate some values at a glance, and serve as a pictorial check on
the accuracy of our solutions.

However, **graphical** presentation of data, although used in different ways mentioned above, is only one method of describing data. T ' s cannot and is not a substitute for other forms of presentation **as** well **as** firher statistical analysis. In the following, we discuss some of the graphical methods of presentation.

Line graphs can be used to show how something changes over time. Line graphs are good for plotting data that has peaks (ups) and valleys (downs), or that was collected in a short time period. The following pages describe the different parts of a line graph.



The Title

The title offers a short explanation of what is in your graph. This helps the reader identify what they are about to look at. It can be creative or simple as long as it tells what is in the graph. The title of this graph tells the reader that the graph contains information about the changes in money spent on students of elementary and secondary schools from 1961 to 2002.

The Legend

The legend tells what each line represents. Just like on a map, the legend helps the reader understand what they are looking at. This legend tells us that the green line represents the actual dollar amount spent on each child and the purple line represents the amount spent when adjusted for inflation.

The Source

The source explains where you found the information that is in your graph. It is important to give credit to those who collected your data! In this graph, the source tells us that we found our information from NCES.

Y-Axis

In line graphs, the y-axis runs vertically (up and down). Typically, the y-axis has numbers for the amount of stuff being measured. The y-axis usually starts counting at 0 and can be divided into as many equal parts as you want to. In this line graph, the y-axis is measuring the amount of money spent on individual students for public education.

The Data

The most important part of your graph is the information, or data, it contains. Line graphs can present more than one group of data at a time. In this graph, two sets of data are presented.

X-Axis

In line graphs, like the one above, the x-axis runs horizontally (flat). Typically, the x-axis has numbers representing different time periods or names of things being compared. In this line graph, the x-axis measured different school years.

3. Graphical Representation of Data

INTRODUCTION

Whenever verbal problems involving a certain situation is presented visually before the learners, it makes easier for the learner to understand the problem and attempt its solution. Similarly, when the data are presented pictorially (or graphically) before the learners, it makes the presentation eye-catching and more intelligible. The learners can easily see the salient features of the data and interpret them.

There are many forms of representing data graphically. They are

(i) Bar graph
(ii) Histograms and Frequency polygons
(iii) Ogive curves
(iv) Pictographs
(v) Pie charts

i) Bar Graph

A bar graph is a graphical representation of frequency distributions of ungrouped data. It is a pictorial representation of the numerical data by a number of bars (rectangles) of uniform width erected vertically (or horizontally) with equal spacing between them.

Construction of Bar Graphs

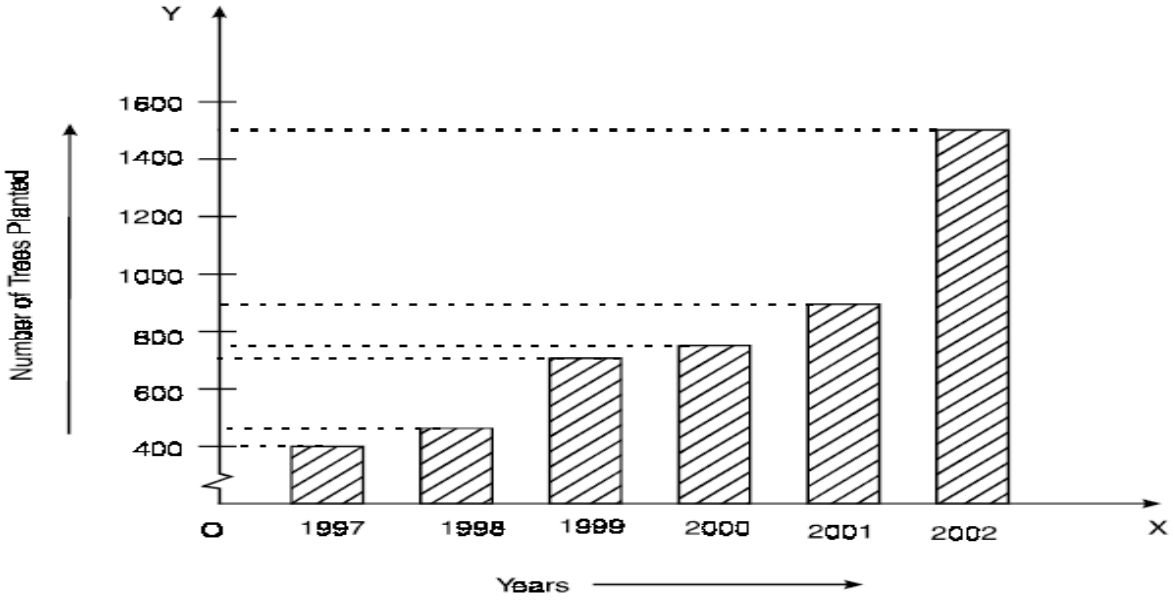
For the construction of bar graphs, we go through the following steps:
Step 1 : We take a graph paper and draw two lines perpendicular to each other and call them horizontal and vertical axes.
Step 2 : Along the horizontal axis, we take the values of the variables and along the vertical axis, we take the frequencies.
Step 3 : Along the horizontal axis, we choose the uniform (equal) width of bars and the uniform gap between the bars, according to the space available.
Step 4 : Choose a suitable scale to determine the heights of the bars. The scale is chosen according to the space available.
Step 5 : Calculate the heights of the bars, according to the scale chosen and draw the bars.
Step 6 : Mark the axes with proper labelling

Let us take some examples to illustrate :

The number of trees planted by an agency in different years is given below

Years	1997	1998	1999	2000	2001	2002	Total
Number of trees planted	400	450	700	750	900	1500	4700

Solution : The bar graph is given below in



Step 1 : We draw two perpendicular lines OX and OY.
Step 2 : On OX, we represent years, from 1997–2002 and on OY we represent the number of trees planted.
Step 3 : On OY, we start with 400 and marks points at equal intervals of 200.
Step 4 : The height of the bars are calculated according to the number of trees.

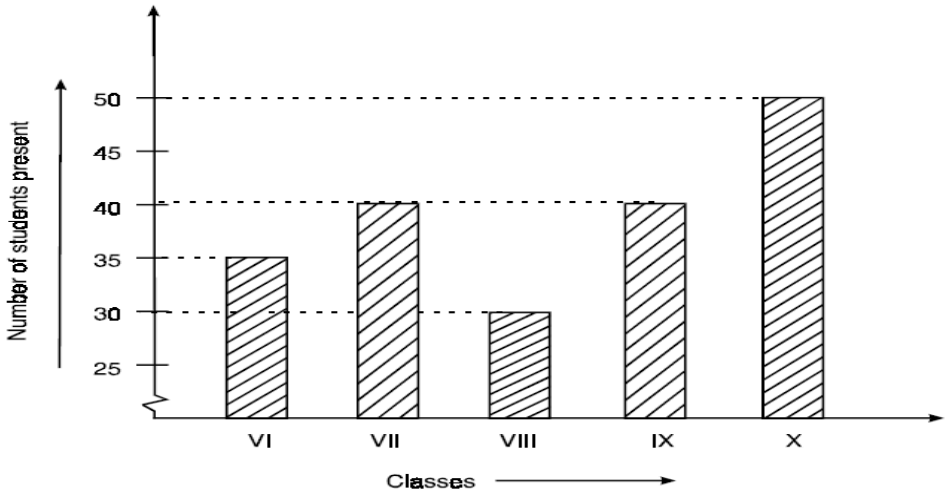
A kink (~) has been shown on the vertical axis showing that the marking on the vertical axis starts from zero but has been shown to start from 400 as the data needs.

Example - 01

The data below shows the number of students present in different classes on a particular day :

Classes	VI	VII	VIII	IX	X
Number of students present	35	40	30	40	50

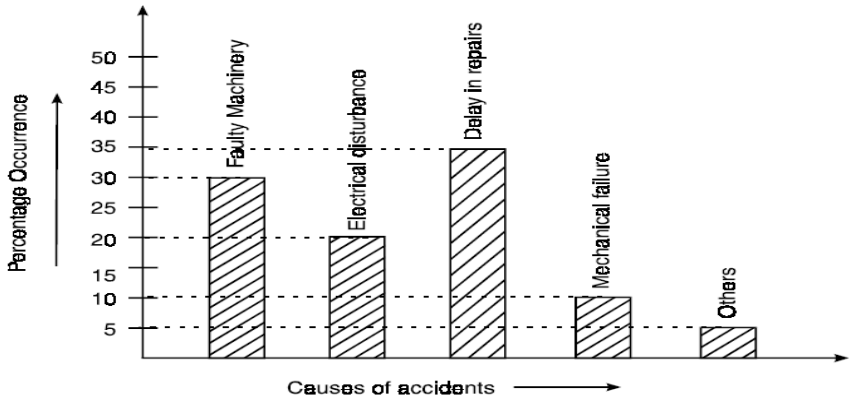
Represent the above data by a bar graph.



Example – 02

Causes	Percentage of Occurrence
Faulty Machinery	30%
Electrical Disturbance	20%
Delay in repairs	35%
Mechanical Failure	10%
Others	5%

Draw a bar graph to represent the above data.



Interpretation of Bar graphs

After drawing a bar graph, we can draw some conclusions, which is called interpreting bar graphs.
Let us take some examples and do the same.

Example – 1

i. In which year the maximum number of trees were planted ?
ii. What trend the number of trees planted show ?
iii. In which years the number of trees planted differ by 50 only ?

Example – 2

i. Which cause is responsible for maximum accidents in factories ? Which is for minimum ?
ii. Can you think of one of the "other" causes ?
iii. How many percent of accidents could have been avoided by timely action?

Bar Diagram

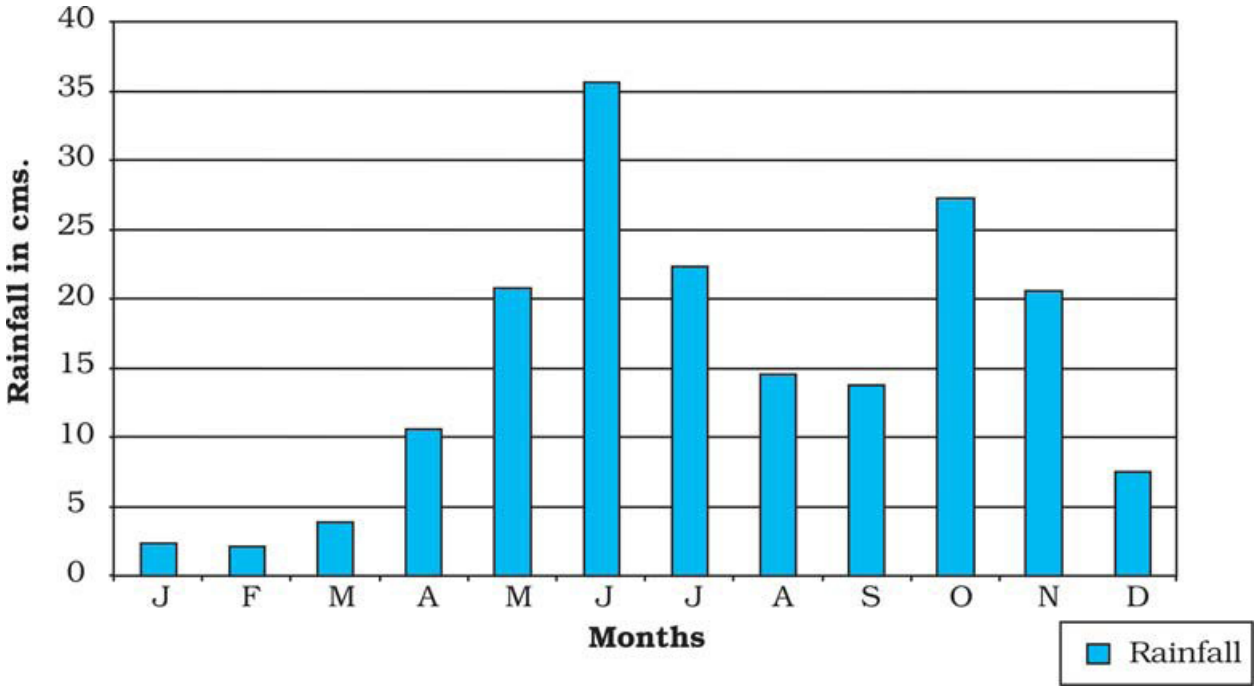
It is also called a columnar diagram. The bar diagrams are drawn through columns of equal width. Following rules were observed while constructing a bar diagram:
(a) The width of all the bars or columns is similar.
(b) All the bars should are placed on equal intervals/distance.
(c) Bars are shaded with colours or patterns to make them distinct and attractive.

Three types of bar diagrams are used to represent different data sets:

- The simple bar diagram
- Compound bar diagram
- Polybar diagram.

- **Simple Bar Diagram**

A simple bar diagram is constructed for an immediate comparison. It is advisable to arrange the given data set in an ascending or descending order and plot the data variables accordingly. However, time series data are represented according to the sequencing of the time period.



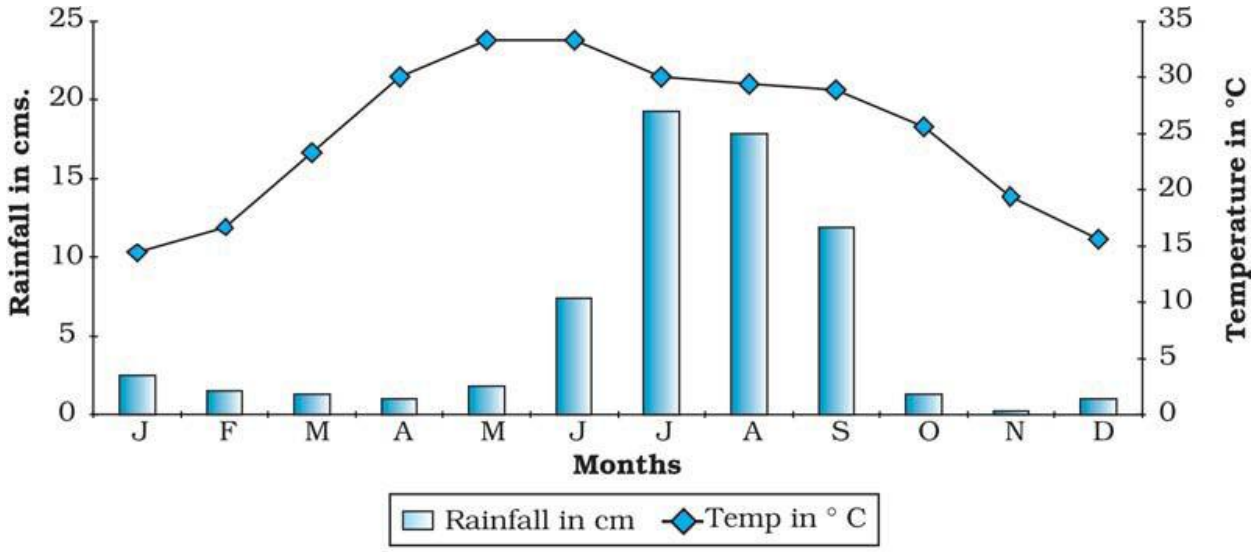
Construction Steps:

Draw X and Y-axes on a graph paper. Take an interval and mark it on Y-axis to plot data.

Divide X-axis into equal parts to draw bars. The actual values will be plotted according to the selected scale.

➢ Line and Bar Graph

The line and bar graphs as drawn separately may also be combined to depict the data related to some of the closely associated characteristics such as the climatic data of mean monthly temperatures and rainfall.

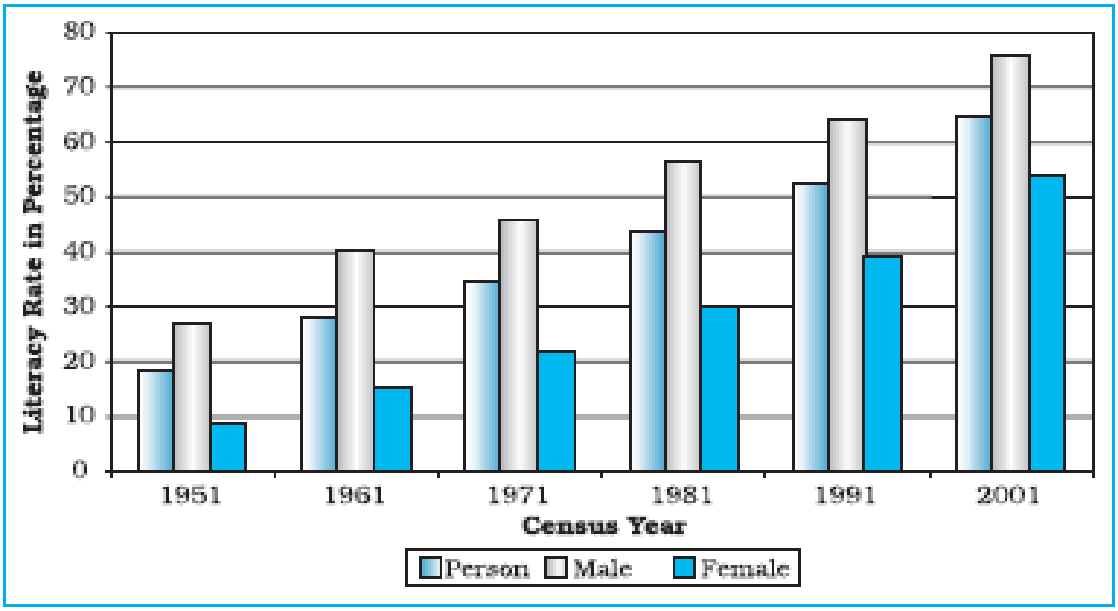


Construction:

a) Draw X and Y-axes of a suitable length and divide X-axis into parts to show months in a year.
b) Select a suitable scale with equal intervals on the Y-axis and label it at its right side.
c) Similarly, select a suitable scale with equal intervals on the Y-axis and label at its left side.
d) Plot data using line graph and columnar diagram.

➢ Multiple Bar Diagram

Multiple bar diagrams are constructed to represent two or more than two variables for the purpose of comparison. For example, a multiple bar diagram may be constructed to show proportion of males and females in the total, rural and urban population or the share of canal, tube well and well irrigation in the total irrigated area in different states.

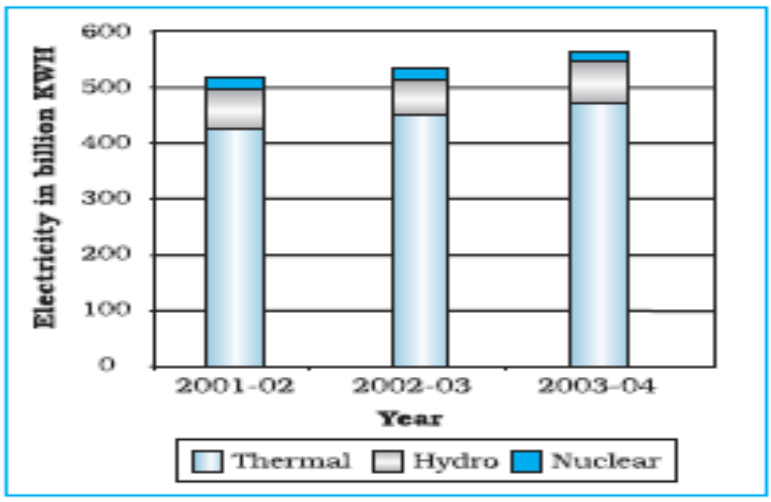


Construction

a) Mark time series data on X-axis and variable data on Y-axis as per the selected scale.
b) Plot the data in closed columns.

- **Compound Bar Diagram**

When different components are grouped in one set of variable or different variables of one component are put together, their representation is made by a compound bar diagram. In this method, different variables are shown in a single bar with different rectangles.



a) Arrange the data in ascending or descending order.
b) A single bar will depict the set of variables by dividing the total length of the bar as per percentage.

(ii) Histograms and Frequency polygons

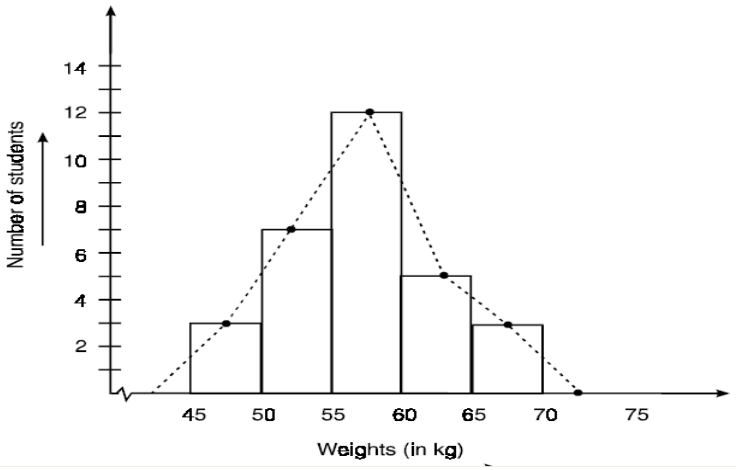
A histogram is a graphical representation of a continuous frequency distribution i.e. grouped frequency distributions. It is a graph, including vertical rectangles, with no space between the rectangles. The class-intervals are taken along the horizontal axis and the respective class frequencies on the vertical axis using suitable scales on each axis. For each class, a rectangle is drawn with base as width of the class and height as the class frequency. The area of the rectangles must be proportional to the frequencies of the respective classes.

A frequency polygon is the join of the mid-points of the tops of the adjoining rectangles. The mid-points of the first and the last classes are joined to the mid-points of the classes preceding and succeding respectively at zero frequency to complete the polygon.
Let us illustrate these with the help of examples.

Example – 1 The following is the frequency distribution of weights of 30 students of class IX of a school. Draw a histogram to represent the data.

Classes :	45–50	50–55	55–60	60–65	65–70	Total
Frequency :	3	7	12	5	3	30

Solution : For drawing a histogram we go through the steps similar to those of a bar graph.
They are given below :
Step 1 : On a paper, we draw two perpendicular lines and call them horizontal and vertical axes.
Step 2 : Along the horizontal axis, we take classes of equal width :
45–50, 50–55, ... As the axis starts from 45–50, we take one interval 40–45 before it and put a kink on axis before that.
Step 3 : Choose a suitable scale on the vertical axis to represent the frequency. It can start from 0 to 12, with a step of 2, *i.e.*, 0, 2, 4, 6, ..., 12, 14
Step 4 : Draw the rectangles as shown in Fig. shows the histogram required.



Note : *A frequency polygon has been shown in dotted lines, as explained in the steps shown above.*

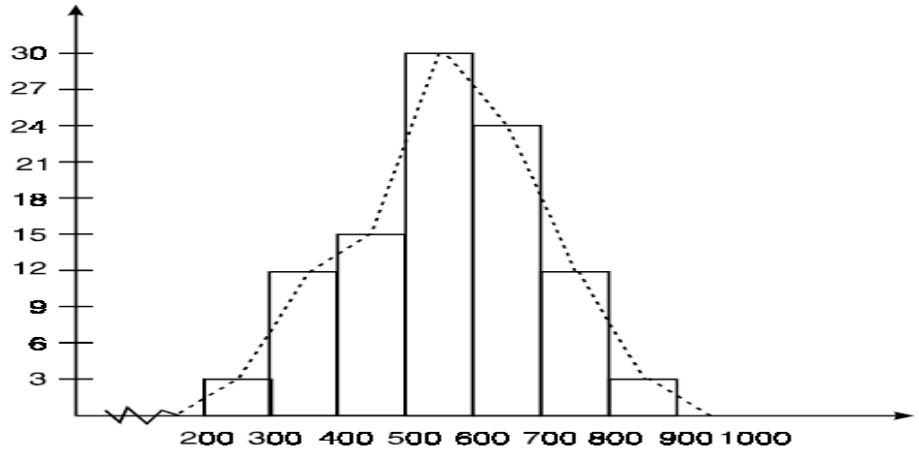
Example – 2

The daily earnings of 100 shopkeepers are given below :

Daily earnings (in Rs)	200-300	300-400	400-500	500-600	600-700	700-800	800-900
No. of shops	3	12	15	30	25	12	3

Draw a histogram and a frequency polygon to represent the above data.

The histogram and frequency polygon representing the above data are given below



Draw a frequency polygon for the following data without-drawing a histogram:

Pocket allowance (in rupees)	0–50	50–100	100–150	150–200	200–250	250–300
Number of students	16	25	13	26	15	5

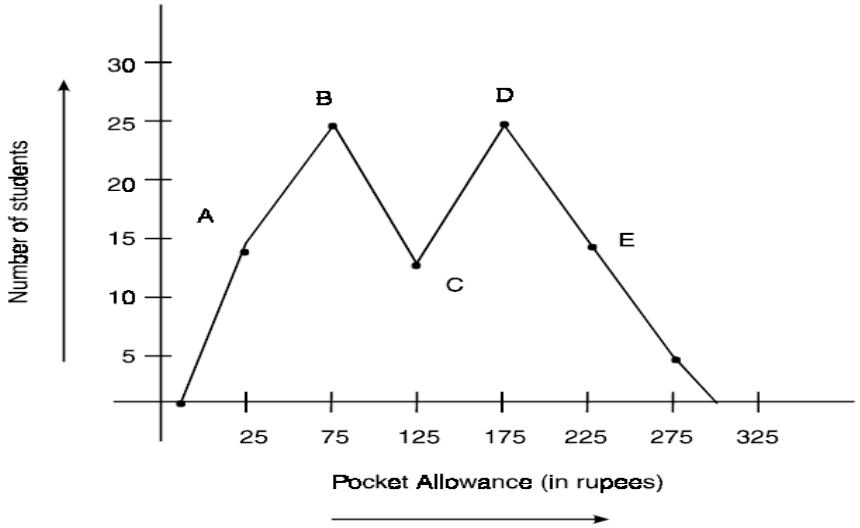
Solution : To draw a frequency polygon without-drawing a histogram we go through the following steps :

Step 1 : Draw two lines perpendicular to each other.

Step 2 : Find the class-marks of different classes. They are 25, 75, 125, 175, 225, 275

Step 3 : Plot the ordered pairs A (25, 16), B (75, 25), C(125, 13), D (175, 26), E(225, 15) and F(275, 5)

Step 4 : Join the points A, B, C, D, E and F and complete the polygon as explained before The frequency polygon is given below :



In addition to histograms and frequency polygons, we are sometimes faced with graphs of other types. When a patient is admitted in a hospital with fever the doctor/nurses prepare a temperature-time graph, which can be referred to any time for reference. Similarly, the velocity time graph and pressure-volume graph are of day-to-day use. We shall learn to draw these graphs and interpret them in the sections below :

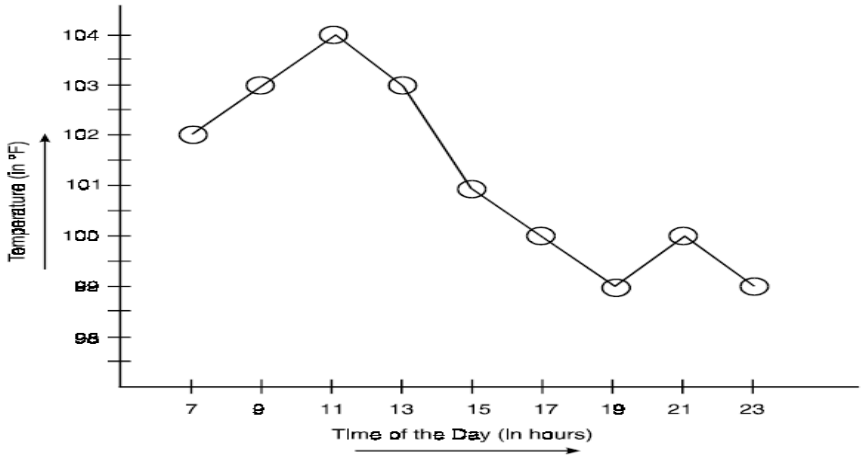
Temperature-Time Graph-Reading and Construction

The body temperature of a patient admitted in a hospital with typhoid fever at different times of a day are given below :

Time of the day	7 hrs	9 hrs	11 hrs	13 hrs	15 hrs	17 hrs	19 hrs	21 hrs	23 hrs
Temperature (in °F)	102	103	104	103	101	100	99	100	99

Draw a graph to represent the above data.

Solution : The graph of the above data is given in Fig. 28.10. The graph has been obtained by joining the points corresponding to pairs, like (7, 102), (9, 103), ..., (23, 99) in the rectangular system of coordinates, by line-segments.



Note : *While drawing the graph it has been assumed that during the time interval in between times, the same trend was present.*

Velocity Time Graph

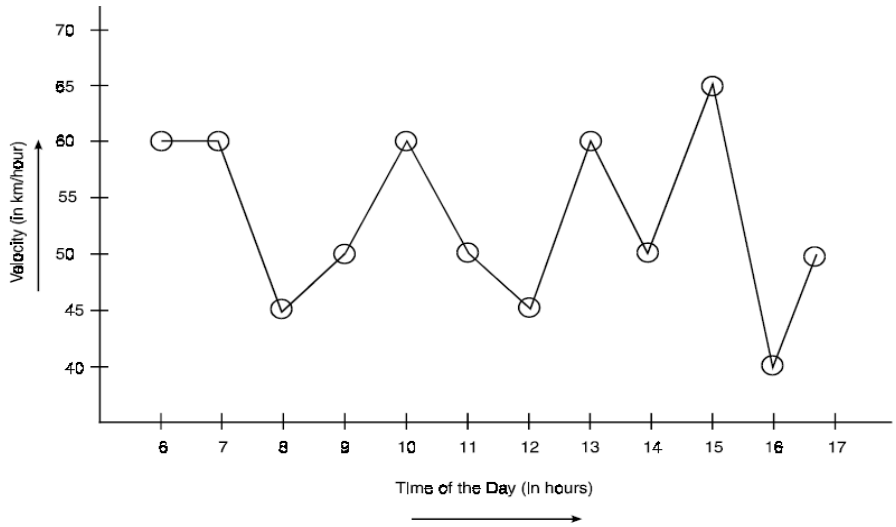
During a journey from one place to other, the speeds of vehicles keep on changing according to traffic congestions. This can be very well shown by a velocity-time graph. Let us illustrate it with the help of example:

During a journey from city A to city B by car the following data regarding
the time and velocity of the car was recorded :

Time of the day (in hours)	6	7	8	9	10	11	12	13	14	15	16	17
Velocity (in km/hour)	60	60	45	50	60	50	45	60	50	65	40	50

Represent the above data by a velocity time graph.

Solution : As before the graph can be obtained by plotting the ordered pairs (6, 60), (7, 60), ... (15, 65), ..., (17, 50) in the rectangular system of coordinates and then by joining them by line-segments.



Pressure-Volume Graph

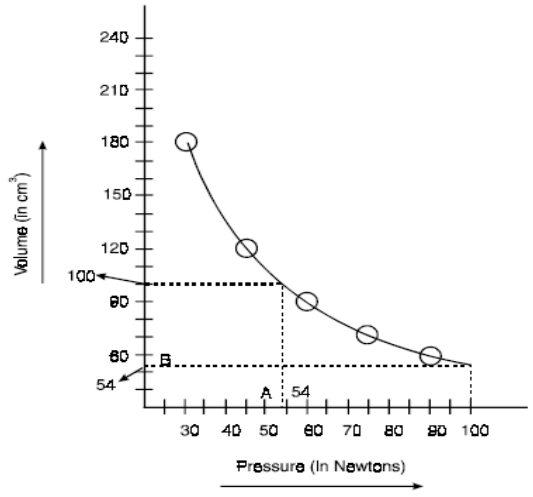
For a fixed quantity of a gas at a constant temperature, is there any relation between pressure and volume of the gas ? Let us see that from the following example :

The following data pertains to pressure and volume of a fixed quantity of gas:

Pressure (p) (in Newton)	60	90	45	30	75
Volume (v) (in cm^3)	90	60	120	180	72

Draw a graph to represent the above data.

Solution :



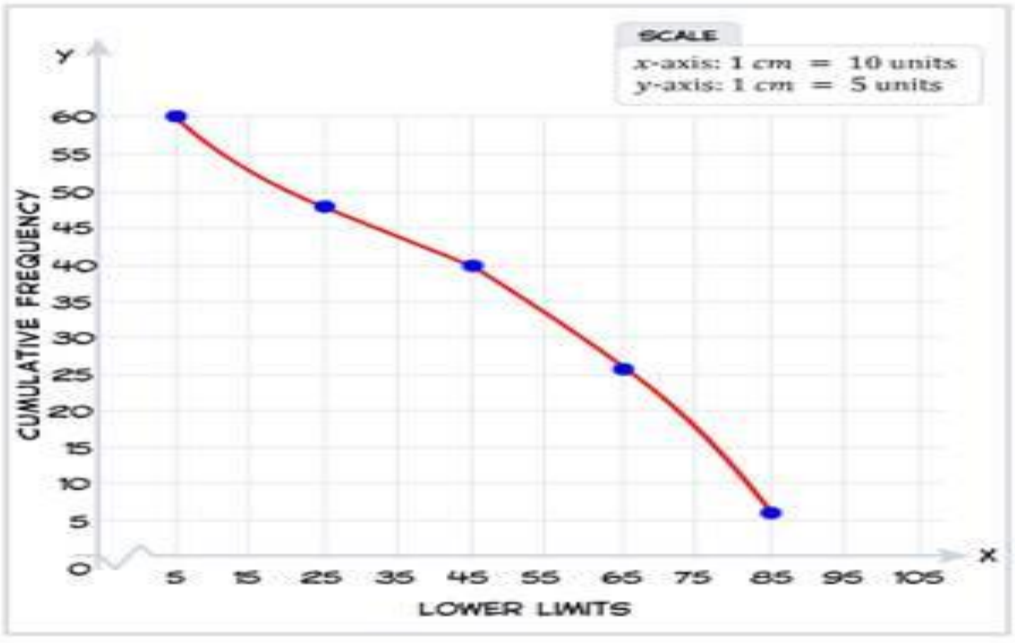
The graph is obtained by joining the plot of the ordered pairs (60, 90), (90, 60), ... (75, 72) by free hand curve.

iii) Ogive curves

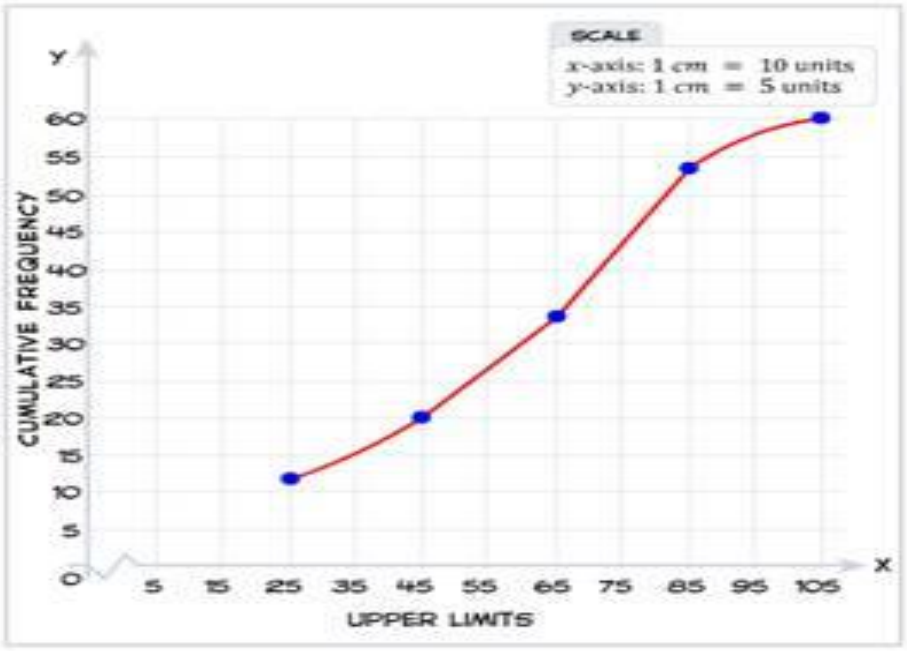
A **curve** that represents the **cumulative frequency distribution** of grouped data is called an **ogive** or **cumulative frequency curve**. The two **types of Ogives** are **more than type ogive** and **less than type ogive.**

An ogive representing a cumulative frequency distribution of '**more than' type** is called a **more than ogive**. An ogive representing a cumulative frequency distribution of **'less than' type** is called a **less than ogive**.

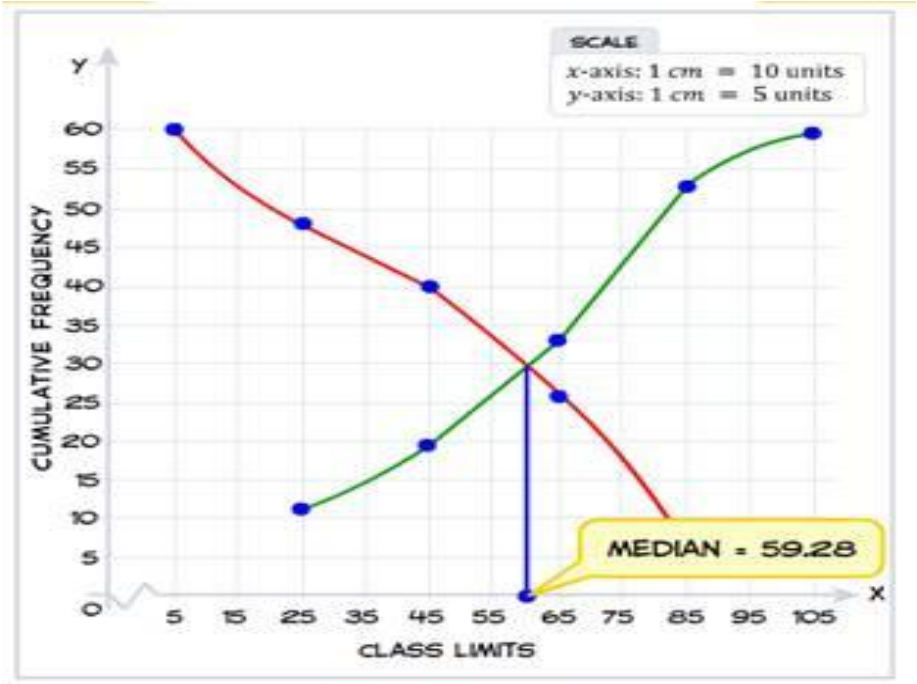
Ogives can be used to find the **median of a grouped data**. The median of grouped data can be obtained graphically by **plotting the Ogives** of the less than type and more than type and locate **the point of intersection of both the Ogives.** The x-coordinate of the point of intersection of two Ogives gives the median of the grouped data.



AN OGIVE REPRESENTING A CUMULATIVE FREQUENCY DISTRIBUTION OF THE 'MORE THAN' TYPE IS CALLED A 'MORE THAN' OGIVE.



AN OGIVE REPRESENTING A CUMULATIVE FREQUENCY DISTRIBUTION OF THE 'LESS THAN' TYPE IS CALLED A 'LESS THAN' OGIVE.

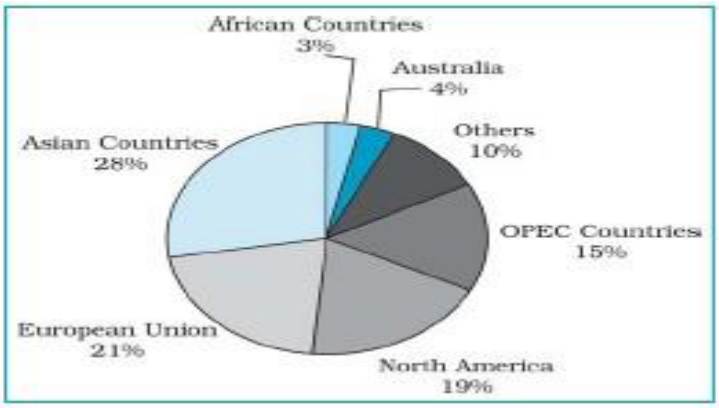


iv) Pie Diagram

Pie diagram is another graphical method of the representation of data. It is drawn to depict the total value of the given attribute using a circle. Dividing the circle into corresponding degrees of angle then represent the sub– sets of the data. Hence, it is also called as Divided Circle Diagram. The angle of each variable is calculated using the following formulae.

If data is given in percentage form, the angles are calculated using the given formulae.

$$\frac{\text{Value of given State/Region X 360}}{\text{Total Value of All States/Regions}}$$



If data is given in percentage form, the angles are calculated using the given formulae.

$$\frac{\text{Percentage of x X 360}}{100}$$

Calculation of Angles

a) Arrange the data on percentages in an ascending order.
b) Calculate the degrees of angles for showing the given values.
c) It could be done by multiplying percentage with a constant of 3.6 as derived by dividing the total number of degrees in a circle by 100, i. e. 360/100.
d) Plot the data by dividing the circle into the required number of divisions to show the share different regions/countries Construction.
e) Select a suitable radius for the circle to be drawn. A radius of 3, 4 or 5 cm may be chosen for the given data set.
f) Draw a line from the centre of the circle to the arc as a radius.

g) Measure the angles from the arc of the circle for each category of vehicles in an ascending order clock-wise, starting with smaller angle.
h) Complete the diagram by adding the title, sub – title, and the legend. The legend mark be chosen for each variable/category and highlighted by distinct shades/colours.

Precautions

a) The circle should neither be too big to fit in the space nor too small to be illegible.
b) Starting with bigger angle will lead to accumulation of error leading to the plot of the smaller angle difficult.

v) Pie Charts

Pie charts are useful to compare different parts of a whole amount. They are often used to present financial information. E.g. A company's expenditure can be shown to be the sum of its parts including different expense categories such as salaries, borrowing interest, taxation and general running costs (i.e. rent, electricity, heating etc).

A pie chart is a circular chart in which the circle is divided into sectors. Each sector visually represents an item in a data set to match the amount of the item as a percentage or fraction of the total data set.

Example

A family's weekly expenditure on its house mortgage, food and fuel is as follows:

Expense	$
Mortgage	300
Food	225
Fuel	75

Draw a pie chart to display the information.

Solution:

$$\begin{aligned}\text{The total weekly expenditure} &= \$300 + \$225 + \$75 \\ &= \$600\end{aligned}$$

We can find what percentage of the total expenditure each item equals.

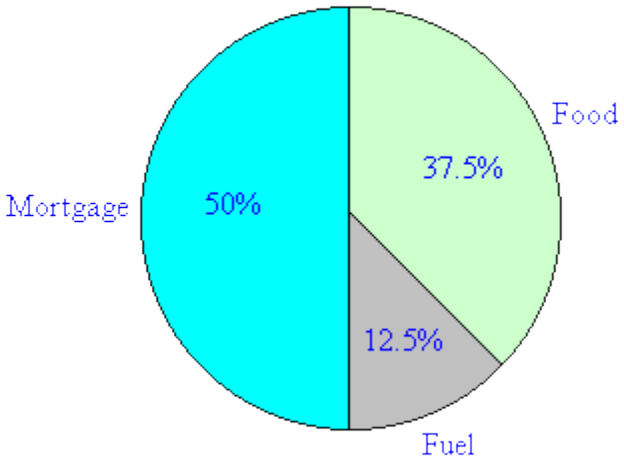
Percentage of weekly expenditure on:

$$\text{Mortgage} = \frac{300}{600} \times 100\% = 50\%$$

$$\text{Food} = \frac{225}{600} \times 100\% = 37.5\%$$

$$\text{Fuel} = \frac{75}{600} \times 100\% = 12.5\%$$

To draw a pie chart, divide the circle into 100 percentage parts. Then allocate the number of percentage parts required for each item.



Note:

- It is simple to read a pie chart. Just look at the required sector representing an item (or category) and read off the value. For example, the weekly expenditure of the family on food is 37.5% of the total expenditure measured.
- A pie chart is used to compare the different parts that make up a whole amount.