Introduction to Statistics

Origin and Growth of Statistics

The origin of statistics can be traced back to the primitive man, who put notches on trees to keep an account of his belongings. During 5000 BC, kings used to carry out census of populations and resources of the state. Kings of olden days made their crucial decisions on wars, based on statistics of infantry, and elephantry units of their own and that of their enemies. Later it enhanced its scope in their kingdoms' tax management and administrative domains. Thus, the word 'Statistics' has its root either to Latin word 'Status' or Italian word 'Statista' or German word 'Statistik' each of which means a 'political state'. The word 'Statistics' was primarily associated with the presentation of facts and figures pertaining to demographic, social and political situations prevailing in a state/government. Its evolution over time formed the basis for most of the science and art disciplines. Statistics is used in the developmental phases of both theoretical and applied areas, encompassing the field of Industry, Agriculture, Medicine, Sports, and Business analytics.

Statistics is concerned with scientific method for collecting, organizing, summarizing, presenting, analyzing, and interpreting of data. The word statistics is normally referred either as numerical facts or methods.

Statistics is used in two different forms-singular and plural. In plural form it refers to the numerical figures obtained by measurement or counting in a systematic manner with a definite purpose such as number of accidents in a busy road of a city in a day, number of people died due to a chronic disease during a month in a state and so on. In its singular form, it refers to statistical theories and methods of collecting, presenting, analyzing, and interpreting numerical figures.

Though the importance of statistics was strongly felt, its tremendous growth was in the twentieth century. During this period, lot of new theories, applications in various disciplines were introduced. With the contribution of renowned statisticians several theories and methods were introduced, naming a few are Probability Theory, Sampling Theory, Statistical Inference, Design of Experiments, Correlation and Regression Methods, Time Series and Forecasting Techniques.

In early 1900s, statistics and statisticians were not given much importance but over the years due to advancement of technology it had its wider scope and gained attention in all fields of science and management. We also tend to think statistician as a small profession but a steady growth in the last century is impressive. It is pertinent to note that the continued growth of statistics is closely associated with information technology. As a result, several new inter- disciplines have emerged. They are Data Mining, Data Warehousing, Geographic Information System, Artificial Intelligence etc. Now-a-days, statistics can be applied in hardcore technological spheres such as Bioinformatics, Signal processing, Telecommunications, Engineering, Medicine, Crimes, Ecology, etc.

Today's business managers need to learn how analytics can help them make better decisions that can generate better business outcomes. They need to understand the statistical concepts that can help analyze and simplify the flood of data around them. They should be able to leverage analytical techniques like decision trees, regression analysis, clustering and association to improve business processes.

Definitions of Statistics

Statistics has been defined by various statisticians.

'Statistics is the science of counting' ***-A. L .Bowley***

'Statistics is the science which deals with the collection, presentation, analysis and interpretation of numerical data' **- *Croxton and Cowden***

Wallist and **Roberts** defines statistics as *"Statistics is a body of methods for making decisions in the face of uncertainty"*

Ya-Lun-Chou slightly modifies Wallist and Roberts definition and come with the following definition: *"Statistics is a method of decision making in the face of uncertainty on the basis of numerical data and calculated risk."*

It may be seen that most of the above definitions of statistics are restricted to numerical measurements of facts and figures of a state. But modern thinkers like Sacrist defines statistics as

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

Among them, the definition by Croxton and Cowden is considered as the most preferable one due to its comprehensiveness. It is clear from this definition that statistics brings out the following characteristics.

Characteristics of Statistics:

- **Aggregate of facts collected in systematic manner for a specific purpose**.

  Statistics deals with the aggregate of facts and figures. A single number cannot be called as statistics. For example, the weight of a person with 65kg is not statistics but the weights of a class of 60 persons is statistics, since they can be studied together, and meaningful comparisons are made in relation to the other. This reminds us of Joseph Stalin's well known quote, "One death is a tragedy; a million is a statistic." Further the purpose for which the data is collected is to be made clear, otherwise the whole exercise will be futile. The data so collected must be in a systematic way and should not be haphazard.

- **Affected by large number of causes to marked extent.**

  Statistical data so collected should be affected by various factors at the same time. This will help the statistician to identify the factors that influence the statistics. For example, the sales of commodities in the market are affected by causes such as supply, demand, and import quality etc. Similarly, as mentioned earlier if a million deaths occur the policy makers will be immediately in action to find out the causes for these deaths to see that such events will not occur.

- **Numerically expressed.**

The statistical facts and figures are collected numerically for meaningful inference. For instance, the service provided by a telephone company may be classified as poor, average, good, very good and excellent. They are qualitative in nature and cannot be called statistics. They should be expressed numerically such as 0 to denote poor,1 for average, 2 for good, 4 to denote very good and 5 for excellent. Then this can be regarded as statistics and is suitable for analysis. The other types of quality characteristics such as honesty, beauty, intelligence, defective etc. which cannot be measured numerically cannot be called statistics. They should be suitably expressed in the form of numbers so that they are called statistics.

- **Enumerated or estimated with a reasonable degree of accuracy.**

The numerical data are collected by counting, measuring or by estimating. For example, to find out the number of patients admitted in a hospital, data is collected by actual counting or to find out the obesity of patients, data are collected by actual measurements on height and weight. In a large scale study like crop estimation, data are collected by estimation and using the powerful sampling techniques, because the actual counting may or may not be possible. Even if it is possible, the measurements involve more time and cost. The estimated figures may not be accurate and precise. However certain degree of accuracy must be maintained for a meaningful analysis.

- **To be placed in relation to the other.**

One of the main reasons for the collection of statistical data is for comparisons in order to make meaningful and valid comparisons, the data should be on the same characteristic as far as possible. For instance, we can compare the monthly savings of male employees to that of the female employees in a company. It is meaningless if we compare the heights of 20 year-old boys to the heights 20 year- old trees in a forest.

Types of Statistics

There are two kinds of Statistics, which are descriptive Statistics and inferential Statistics. In descriptive Statistics, the Data or Collection Data are described in a summarized way, whereas in inferential Statistics, we make use of it in order to explain the descriptive kind. Both are used on a large scale. Statistics is mainly divided into the following two categories.

- Descriptive Statistics
- Inferential Statistics

Descriptive Statistics

In the descriptive Statistics, the Data is described in a summarized way. The summarization is done from the sample of the population using different parameters like Mean or standard deviation. Descriptive Statistics are a way of using charts, graphs, and summary measures to organize, represent, and explain a set of Data.

- Data is typically arranged and displayed in tables or graphs summarizing details such as histograms, pie charts, bars, or scatter plots.
- Descriptive Statistics are just descriptive and thus do not require normalization beyond the Data collected.

Inferential Statistics

In the Inferential Statistics, we try to interpret the Meaning of descriptive Statistics. After the Data has been collected, analyzed, and summarized we use Inferential Statistics to describe the Meaning of the collected Data.

- Inferential Statistics use the probability principle to assess whether trends contained in the research sample can be generalized to the larger population from which the sample originally comes.
- Inferential Statistics are intended to test hypotheses and investigate relationships between variables and can be used to make population predictions.
- Inferential Statistics are used to draw conclusions and inferences, i.e., to make valid generalizations from samples.

Functions of Statistics

The functions of statistics can be elegantly expressed as 7 - C's as:

S.NO	Functions	What it does
1	Collection	The basic ingredient of statistics is data. It should be carefully and scientifically collected
2	Classification	The collected data is grouped based on similarities so that large and complex data are in understandable form.
3	Condensation	The data is summarized, precisely without losing information to do further statistical analysis.
4	Comparison	It helps to identify the best one and checking for the homogeneity of groups,
5	Correlation	It enables to find the relationship among the variables
6	Causation.	To evaluate the impact of independent variables on the dependent variables.
7	Chance	Statistics helps make correct decisions under uncertainty.

Scope and Applications

In ancient times the scope of statistics was limited. When people hear the word 'Statistics' they think immediately of either sports related numbers or a subject they have studied at college and passed with minimum marks. While statistics can be thought in these terms there is a wide scope for statistics. Today, there is no human activity which does not use statistics. There are two major divisions of statistical methods called descriptive statistics and inferential statistics and each of the divisions are important and satisfies different objectives. The descriptive statistics is used to consolidate a large amount of information. For example, measures of central tendency, like mean are descriptive statistics. Descriptive statistics just describes the data in a condensed form for solving some limited problems. They do not involve beyond the data at hand.

Inferential statistics, on the other hand, are used when we want to draw meaningful conclusions based on sample data drawn from a large population. For example, one might want to test whether

a recently developed drug is more efficient than the conventional drug. Hence, it is impossible to test the efficiency of the drug by administering to each patient affected by a particular disease, but we will test it only through a sample. A quality control engineer may be interested in the quality of the products manufactured by a company. He uses a powerful technique called acceptance sampling to protect the producer and consumer interests. An agricultural scientist wanted to test the efficacy of fertilizers should test by designed experiments. He may be interested in farm size, use of land and crop harvested etc. One advantage of working in statistics is that one can combine his interest with almost any field of science, technology, or social sciences such as Economics Commerce, Engineering Medicine, and Criminology and so on.

The profession of statistician is exciting, challenging and rewarding. Statistician is the most prevalent title but professionals like Risk analyst, Data analyst, Business analyst have been engaged in work related to statistics. We have mentioned earlier that statistic has applications to almost all fields. Here in this section, we highlight its applications to select branches.

- **Statistics and actuarial science**

  Actuarial science is the discipline that extensively applies statistical methods among other subjects involved in insurance and financial institutions. The professionals who qualify in actuarial science course are called actuaries. Actuaries, in the earlier days used deterministic models to assess the premiums in insurance sector. Nowadays, with modern computers and sophisticated statistical methods, science has developed vastly.

- **Statistics and Commerce**

  Statistical methods are widely used in business and trade solutions such as financial analysis, market research and manpower planning. Every business establishment irrespective of the type must adopt statistical techniques for its growth. They estimate the trend of prices, buying and selling, importing, and exporting of goods using statistical methods and past data. Ya-Lun-Chou says, “It is not an exaggeration to say that today nearly every decision in business is made with the aid of statistical data and statistical methods.”

- **Statistics and Economics**

Statistical methods are very much useful to understand economic concepts, such as mandatory policy and public finance. In the modern world, economics is taught as an exact service which makes extensive use of statistics. Some of the important statistical techniques used in economic analysis are Times series, Index Numbers, Estimation theory and Tests of significance, stochastic models. According to Engberg "No Economist would attempt to arrive at a conclusion concerning the production or distribution of wealth without an exhaustive study of statistical data." In our country many state governments have a division called Department of Economics and Statistics for the analysis of Economic data of the state.

- **Statistics and Medicine**

In medical field, statistical methods are extensively used. If we look at the medical journals one can understand to what extent the statistical techniques play a key role. Medical statistics deals with the applications of statistical methods like tests of significance and confidence intervals to medicine and health science including epidemiology, public health. Modern statistical methods help the medical practitioners to understand how long a patient affected by a dreaded disease will survive and what are the factors that influence a patient to be alive or dead.

- **Statistics and Agriculture**

Experimentation and inference based on these experiments are the key features of general scientific methodology. Agricultural scientists conduct experiments and make inferences to decide whether the particular variety of crop gives a better yield than others or a particular type of fertilizer etc, there are several institutes where research is being done by making use of statistical methods like analysis of variance (ANOVA), factorial experiments etc., falls under the hut of Design of experiments.

- **Statistics and Industry**

Statistical methods play a vital role in any modern use of science and technology. Many statistical methods have been developed and applied in industries for various problems. For example, to maintain the quality of manufactured products the concept of statistical quality control is used. The quality in time domain study of mechanical, electrical, or electronic items the concept of 'Reliability' has emerged. Total quality management and six-sigma theories make use of statistical concepts.

- **Statistics and Information Technology**

Information Technology is the applications of computers and telecommunication equipment to store, retrieve, transmit and manipulate data. Now-a-days, several industries are involved in information technology and massive amounts of data are stored every day. These data are to be analyzed meaningfully so that the information contained in the data is used by the respective users. To address this issue, fields such as data mining, Machine learning have emerged. Data mining an interdisciplinary sub field of computer science is the computational process of discovering patterns in large data sets involving methods such as artificial intelligence and statistics. Persons trained in statistics with computing knowledge have been working as data analytics to analyze such huge data.

- **Statistics and Government**

Statistics provides statistical information to government to evolve policies, to maintain law and order, to promote welfare schemes and to other schemes of the government. In other words, statistical information is vital in overall governance of the state. For instance, statistics provide information to the government on population, agricultural production, industrial production, wealth, imports, exports, crimes, birth rates, unemployment, education, minerals and so on.

Limitation of Statistics

Although Statistics has wide field of application, it has some limitations. Some of these limitations are as follows.

- **Qualitative Aspect Ignored:**

The statistical methods don't study the nature of phenomenon which cannot be expressed in quantitative terms. Such phenomena cannot be a part of the study of statistics. These include health, riches, intelligence etc. It needs conversion of qualitative data into quantitative data. So, experiments are being undertaken to measure the reactions of a man through data. Now a days statistics is used in all the aspects of the life as well as universal activities.

- **It does not deal with individual items:**

It is clear from the definition given by Prof. Horace Sacrist, "By statistics we mean aggregates of facts and placed in relation to each other", that statistics deals with only aggregates of facts or items, and it does not recognize any individual item. Thus, individual terms as death of 6 persons in an accident, 85% results of a class of a school in a particular year, will not amount to statistics as they are not placed in a group of similar items. It does not deal with the individual items, however, important they may be.

- **It does not depict entire story of phenomenon:**

When even phenomena happen, that is due to many causes, but all these causes cannot be expressed in terms of data. So, we cannot reach at the correct conclusions. Development of a group depends upon many social factors like, parents' economic condition, education, culture, region, administration by government etc. But all these factors cannot be placed in data. So, we analyze only that data we find quantitatively and not qualitatively. So, results or conclusion are not 100% correct because many aspects are ignored.

- **Statistics can be misused**

Only the experts or statistician can handle statistical data properly. It is likely to be misused the Statistics by non-statistical persons in handling data and interpreting the result.

- **Statistical laws are not exact:**

Generally statistical laws are probabilistic in nature. Based on probability or interpolation, we can only estimate the production of paddy in 2008 but cannot make a claim that it would be exactly 100 %. Here only approximations are made.

- **Results are true only on average:**

As discussed above, here the results are interpolated for which time series or regression or probability can be used. These are not true. If average of two sections of students in statistics is same, it does not mean that all the 50 students is section A has got same marks as in B. There may be much variation between the two. So, we get average results.

- **To Many methods to study problems:**

In this subject we use so many methods to find a single result. Variation can be found by quartile deviation, mean deviation or standard deviations and results vary in each case.

Data Collection

Introduction

Statistical data are the basic ‘ingredients’ of Statistics on which statistician work. A set of numbers representing records of observations is termed statistical data. The need to collect data arises in every sphere of human activity. However, that ‘Garbage in garbage out’ applies in Statistics too. Hence adequate care must be taken in the collection of data. It is a poor practice to depend on whatever data available.

Information, especially facts or numbers collected for decision making is called data. Data may be numerical or categorical. Data may also be generated through a variable.

Variable

A variable is an entity that varies from a place to place, a person to person, a trial to trial and so on. For instance, the height is a variable; domicile is a variable since they vary from person to person.

- A variable is said to be quantitative if it is measurable and can be expressed in specific units of measurement (numbers).
- A variable is said to be qualitative if it is not measurable and cannot be expressed in specific units of measurement (numbers). This variable is also called categorical variable.

The variable height is a quantitative variable since it is measurable and is expressed in a number while the variable domicile is qualitative since it is not measured and is expressed as rural or urban. It is noted that they are free from units of measurement.

Classification of Data

The data that are unorganized or have not been arranged in any way are called raw data. The ungrouped data are often voluminous, complex to handle and hardly useful to draw any vital decisions. Hence, it is essential to rearrange the elements of the raw data set in a specific pattern. Further, it is important that such data must be presented in a condensed form and must be classified according to homogeneity for the purpose of analysis and interpretation. An arrangement of raw data in an order of magnitude or in a sequence is called **array**. Specifically, an arrangement of observations in an ascending or a descending order of magnitude is said to be an **ordered array.**

Classification is the process of arranging the primary data in a definite pattern and presenting in a systematic form. *Horace Secrist* defined classification as the process of arranging the data into sequences and groups according to their common characteristics or separating them into different but related parts. It is treated as the process of classifying the elements of observations or things into different groups or classes or sequences according to the resemblances and similarities of their character. It is also defined as the process of dividing the data into different groups or classes which are as homogeneous as possible within the groups or classes, but heterogeneous between themselves.

Objectives of Classification

- It explains the features of the data.
- It facilitates comparison with similar data.
- It strikes a note of homogeneity in the heterogeneous elements of the collected information.
- It explains the similarities which may exist in the diversity of data points.
- It is required to condense the mass data in such a manner that the similarities and dissimilarities are understood.
- It reduces the complexity of nature of data and renders the data to comprehend easily.
- It enables proper utilization of data for further statistical treatment.

Data collection process

There are five important questions to ask in the process of collecting data: What?

QUESTION	RELATED ACTIVITY
What data is to be collected?	Decide the relevant data of the study
How will the data be collected?	Choice of a data collection instrument
Who will collect the data?	Method of enquiry: Primary / Secondary
Where the data will be collected?	Decide the Population oftt the survey
When will the data be collected?	Fixing the time schedule

Data Measurement Scale

Measurement may be defined as the assignment of numbers to objects or events according to certain rules. There are generally four types of measurement scales, which are as follows.

a) Nominal scale

Nominal scale is used for measuring variables which are qualitative in nature. It is the first level of measurement where labels are assigned to the attributes of the variables in the form of number. Numbers are used as mere identifiers and do not hold any numerical value & no arithmetic operations can be drawn upon them. It only satisfies the 'Identity' property of scale of measurement. Nominal scale is the simplest scale & is also called as the 'Categorical scale' because it represents only the names or categories. It is also called as least powerful level of measurement. The only statistical analysis that can be performed on a nominal scale is frequency count. Mode is used as a measure of central tendency.

For example: -

- jersey number of players in cricket team, types of hair color, PAN number, Telephone number etc.
- Another example, what is your gender? – Male (1) or Female (2)

b) Ordinal scale

Ordinal scale is used for measuring variable which are qualitative in nature. It is the second level of measurement where labels are assigned to the variable in the form of numbers & they are arranged in a proper order. Not only the numbers but also the order of the variables is important. That's why it is called as ordinal scale. It satisfies the 'Identity' & 'Magnitude/Order' property.

*Ordinal scales measure non-numeric concepts like satisfaction, happiness, discomfort, beauty etc. By giving ranks. Median or mode are used as the measures of central tendency & spearman's rank correlation.

For example,

- Order- How much happy are you with our services?

Very happy-1

Happy-2

Neutral-3

Unhappy-4

Very unhappy-5

- Another example, Ranks of students in an academic test, health status (excellent, average, poor)

Here, the order is represented but the difference between the variable is not indicated.

c) Interval scale

Interval scale is used for measuring variables which are quantitative in nature. It is the third level of measurement where labels are assigned to the variables in the form of numbers & they are arranged in a proper order with equal differences between the values. Along with the numbers & order, the difference between the values is also known. That's why it is called an Interval scale. It is an extension of ordinal scale. (i. e., it possesses the property of identity, order & equal intervals). Arithmetic operations like addition & subtraction can be performed on the variables but not multiplication & division and hence, ratios can't be calculated. Interval scales don't have a true zero meaning negative values also exist. Like -10 degree Celsius temperature. Mean, median, mode is used as the measure of central tendency. And standard deviation and range are used as the measures of dispersion. For example, a temperature scale where difference between 60 & 70 degree Celsius is same as that of the difference between 20 &30 degree Celsius

d) Ratio scale

Ratio scale is used for measuring variables which are quantitative in nature. It is the fourth level of measurement which possesses all the attributes of an interval scale along with the property of absolute zero. Arithmetic operations like addition & subtraction can be performed on the variables along with multiplication & division. Here, the ratios can be calculated. That's why it is called a ratio scale. Like, weight of ram is double of that of Shyam. Ratio scales have a true zero meaning negative values don't exist. Like there cannot be a negative weight or negative length. It is the most powerful level of measurement. Mean, median, mode, harmonic mean, geometric mean are used as the measures of central tendency. And standard deviation and coefficient of variation are used as the measures of dispersion. For example, height, weight, length, distance etc.

Summary

Feature	**Nominal**	**Ordinal**	**Interval**	**Ratio**
Level of measurements	First	Second	Third	Fourth
Type of variable	Qualitative	Qualitative	Quantitative	Quantitative
Identity	Yes	Yes	Yes	Yes
Magnitude/order	No	Yes	Yes	Yes
Equal interval	No	No	Yes	Yes
Absolute zero	No	No	No	Yes
Central tendency	Mode	Median & mode	Mean, median & mode	Mean, median, mode, geometric & harmonic mean
Source of dispersion	……..	……	Standard deviation & range	Standard deviation & coefficient of variation
Arithmetic operation	……..	……..	Only addition & subtraction	Add, subtract, multiply & divide
Statistical tests	Non-parametric	Non-parametric	Parametric	Parametric

Types of Data

One of the major elements and basis of statistical research is data collection, where the most basic data that can be collected in this process is primary data. In other words, we can say that data is the basis of all statistical operations and primary data is the simplest of all data. Primary data is one of the 2 main types of data, with the second one being secondary data. These 2 data types have important uses in research.

Primary data

Primary data is a type of data that is collected by researchers directly from main sources through interviews, surveys, experiments, etc. Primary data are usually collected from the source—where the data originally originates from and are regarded as the best kind of data in research.

The sources of primary data are usually chosen and tailored specifically to meet the demands or requirements of research. Also, before choosing a data collection source, things like the aim of the research and target population need to be identified.

The various methods used to collect primary data are:

- Direct Method
- Indirect Method
- Questionnaire Method
- Local Correspondents Method
- Enumeration Method

1. Direct Method:

There are two methods under the direct method

(a) Personal Contact Method

As the name says, the investigator himself goes to the field, meets the respondents, and gets the required information. In this method, the investigator personally interviews the respondent either directly or through phone or through any electronic media. This method is suitable when the scope of investigation is small and greater accuracy is needed.

Merits:

- This method ensures accuracy because of personal interaction with the investigator.
- This method enables the interviewer to suitably adjust the situations with the respondent.

Limitations:

- When the field of enquiry is vast, this method is more expensive, time consuming and cumbersome.
- In this type of survey, there is chance for personal bias by the investigator in terms of asking 'leading questions.

(b) Telephone Interviewing

In the present age of communication explosion, telephones and mobile phones are extensively used to collect data from the respondents. This saves the cost and time of collecting the data with a good amount of accuracy.

2. Indirect Method:

The indirect method is used in cases where it is delicate or difficult to get the information from the respondents due to unwillingness or indifference. The information about the respondent is collected by interviewing the third party who knows the respondent well.

Instances for this type of data collection include information on addiction, marriage proposal, economic status, witnesses in court, criminal proceedings etc. The shortcoming of this method is genuineness and accuracy of the information, as it completely depends on the third party.

3. Questionnaire Method

A questionnaire contains a sequence of questions relevant to the study arranged in a logical order. Preparing a questionnaire is a very interesting and a challenging job and requires good experience and skill.

The general guidelines for a good questionnaire:

- The wording must be clear and relevant to the study
- Ability of the respondents to answer the questions to be considered
- Avoid jargons
- Ask only the necessary questions so that the questionnaire may not be lengthy.
- Arrange the questions in a logical order.
- Questions which hurt the feelings of the respondents should be avoided.
- Calculations are to be avoided.
- It must be accompanied by the covering letter stating the purpose of the survey and guaranteeing the confidentiality of the information provided.

Editing the preliminary questionnaire

Once a preliminary draft of the questionnaire has been designed, the researcher is obligated to critically evaluate and edit, if needed. This phase may seem redundant, given all the careful thoughts that went into each question. But recall the crucial role played by the questionnaire.

Pre Test

Once the rough draft of the questionnaire is ready, pretest is to be conducted. This practice of pretest often reveals certain short comings in the questions, which can be modified in the final form of the questionnaire. Sometimes, the questionnaire is circulated among the competent investigators to make suggestions for its improvement. Once this has been done and suggestions are incorporated, the final form of the questionnaire is ready for the collection of data.

Advantages:

- In a short span of time, vast geographical area can be covered.
- It involves less labor.

Limitations:

- This method can be used only for the literate population.
- Some of the mailed questionnaires may not be returned.
- Some of the filled questionnaires may not be complete.
- The success of this method depends on the nature of the questions and the involvement of the respondents.

4. Local Correspondents Method

In this method, the investigator appoints local agents or correspondents in different places. They collect the information on behalf of the investigator in their locality and transmit the data to the investigator or headquarters. This method is adopted by newspapers and government agencies. This method is economical and provides timely information on a continuous basis. It involves high degree of personal bias of the correspondents.

5. Enumeration method:

In this method, the trained enumerators or interviewers take the schedules themselves, contact the informants, get replies, and fill them in their own handwriting. Thus, schedules are filled by the enumerator whereas questionnaires are filled by the respondents. The enumerators are paid honorarium. This method is suitable when the respondents include illiterates. The success of this method depends on the training imparted to the enumerators.

Secondary data

Secondary data is collected and processed by some other agency, but the investigator uses it for his study. They can be obtained from published sources such as government reports, documents, newspapers, books written by economists or from any other source., for example websites. Use of secondary data saves time and cost. Before using the secondary data, scrutiny must be done to assess the suitability, reliability, adequacy, and accuracy of the data.

Merits:

- It saves time and cost.
- If specially trained persons collect it, the quality of secondary data is better.
- It helps to make primary data collection more specific since with the help of secondary data, we can make out what are the gaps and deficiencies and what additional information needs to be collected.
- It helps to improve the understanding of the problem.
- It provides a basis for comparison for the data that is collected by the researcher.

Limitations:

- Accuracy of secondary data is not known.
- Data may be outdated.

Precaution in using secondary data

The following are the main precautions that should be taken before using secondary data.	
(1) Reliable agency	● We must ensure the agency that has published the data should be reliable.
(2) Suitability for the purpose of an enquiry	● The Investigator must ensure that the data is suitable for the purpose of the present enquiry. ● The suitability of the data is determined by investigating the nature, objectives, time of collection, etc. of the secondary data.
(3) Adequacy and accuracy to avoid the impact of bias	● It is necessary to use adequate data to avoid biases and prejudices leading to incorrect conclusions.
(4) Method of collecting the data used	● The investigator should also ascertain as to what method was used in collecting the data. ● Sampling methods may be biased depending upon the mode of selection of samples. ● All these should be ascertained before making use of the secondary data.

Difference between Primary data and Secondary data

Parameters of Comparison	Primary Data	Secondary Data
Definition	It is the crude form of all the data.	It is a refined form of data.
Source	It can be collected using various methods like interviews, experiments, etc.	It can be obtained from the internet, journals, etc.
Authenticity	It is very authentic in relation to the topic concerned.	It may be biased. It depends on the biases of the researcher.
Cost of collection	It is very costly to collect such data.	It costs very little or nothing.
Purpose	The primary purpose of the data is to add new knowledge.	It is a manipulated form of data and just tells the same story from a different perspective.

Census Method

The census method is also called complete enumeration method. In this method, information is collected from every individual in the statistical population. Census of Nepal is one of the best examples. It is carried out once in every ten years. An enquiry is carried out, covering each and every house in Nepal. It focuses on demographic details. They are collected and published by the Central Bureau of Statistics Nepal.

Appropriateness of this method:

The complete enumeration method is preferable provided the population is small and not scattered. Otherwise, it will have the following disadvantages.

Disadvantages:

- It is more time consuming, expensive and requires more skilled and trained investigators.
- More errors creep in due to the volume of work.
- Complete enumeration cannot be used if the units in the population destructive in nature. For example, blood testing, testing whether the rice is cooked or not in a kitchen,
- When area of the survey is very large and there is less knowledge about the population, this method is not practicable. For example, the tiger population in Nepal, number trees in a forest cannot be enumerated using census method.

Sampling method:

In view of all these difficulties one has to resort to sampling methods for collecting the data.

Sample is small proportion of the population taken from the population to study the characteristics of the population. By observing the sample one can make inferences about the population from which it is taken.

Sampling is a technique adopted to select a sample. The sample must represent or exactly duplicate the characteristics of the population under study. In such case that sample is called as a representative sample. The sampling method used for selecting a sample is important in determining how closely the sample resembles the population, in determining.

Sampling unit is the basic unit to be sampled from the population which cannot be further subdivided for the purpose of sampling. Head of the house is the sampling unit for the household survey. In the study to know the average age of a class, student is the sampling unit.

Sampling frame to adopt a sampling procedure it is precisely about sampling necessary to prepare a list such that there exists, one to one that “one grain suffices correspondence between sampling units and numbers. Such a list or map is called sampling frame. A list of villages in a district, Student list of +1 and +2 students in the above said example, A list of houses in a household survey etc.

Sample size is the number of units in the sample.

Merits and Limitations of Sampling

The prime objective of the sampling is to get the representative sample which will provides the desired information about the population with maximum accuracy at a given cost.

Merits

- Cost: Expenditure on conducting the survey is less compared to complete enumeration.
- Time: The consumption of time is relatively less in a sample study than potentially generated voluminous data.
- Accuracy: It is practically proved that the results based on representative samples more reliable than the complete enumeration.
- In the case of destructive type situations, sampling method is the only way.

Limitations

- Accuracy depends on the honesty of the investigator
- There is possibility for sampling error.

DIAGRAMATIC AND GRAPHICAL REPRESENTATION

Introduction:

We need to arrange the raw data to make them understandable. For this we use the techniques of classification and tabulation that help in summarizing the collected data and presenting them in a systematic manner. However, these forms of presentation do not always prove to be interesting to the common man. One of the most convincing and appealing ways in which statistical results may be presented is through diagrams and graphs. Just one diagram is enough to represent a given data more effectively than thousand words. Moreover even a layman who has nothing to do with numbers can also understands diagrams. Evidence of this can be found in newspapers, magazines, journals, advertisement, etc. An attempt is made in this chapter to illustrate some of the major types of diagrams and graphs frequently used in presenting statistical data.

Diagrams:

A diagram is a visual form for presentation of statistical data, highlighting their basic facts and relationship. If we draw diagrams on the basis of the data collected they will easily be understood and appreciated by all. It is readily intelligible and save a considerable amount of time and energy.

Significance of Diagrams and Graphs:

Diagrams and graphs are extremely useful because of the following reasons.

1. They are attractive and impressive.
2. They make data simple and intelligible.
3. They make comparison possible
4. They save time and labour.
5. They have universal utility.
6. They give more information.
7. They have a great memorizing effect.

General rules for constructing diagrams:

The construction of diagrams is an art, which can be acquired through practice. However, observance of some general guidelines can help in making them more attractive and effective. The diagrammatic presentation of statistical facts will be advantageous provided the following rules are observed in drawing diagrams.

1. A diagram should be neatly drawn and attractive.
2. The measurement of geometrical figures used in diagram should be accurate and proportional.
3. The size of the diagrams should match the size of the paper.
4. Every diagram must have a suitable but short heading.
5. The scale should be mentioned in the diagram.
6. Diagrams should be neatly as well as accurately drawn with the help of drawing instruments.
7. Index must be given for identification so that the reader can easily make out the meaning of the diagram.
8. Footnote must be given at the bottom of the diagram.
9. Economy in cost and energy should be exercised in drawing diagram.

Types of diagrams:

In practice, a very large variety of diagrams are in use and new ones are constantly being added. For the sake of convenience and simplicity, they may be divided under the following heads:

1. One-dimensional diagrams(Bar diagrams)
2. Two-dimensional diagrams(Pie diagram)
3. Three-dimensional diagrams(Cubic diagram, cylinders, spheres, prisms)
4. Pictograms and Cartograms

One-dimensional diagrams:

In such diagrams, only one-dimensional measurement, i.e height is used and the width is not considered. These diagrams are in the form of bar or line charts and can be classified as

1. Line Diagram
2. Simple Bar Diagram
3. Sub-divided Bar Diagram
4. Percentage Bar Diagram
5. Multiple Bar Diagram

Line Diagram:

Line diagram is used in case where there are many items to be shown and there is not much of difference in their values. Such diagram is prepared by drawing a vertical line for each item according to the scale. The distance between lines is kept uniform. Line diagram makes comparison easy, but it is less attractive.

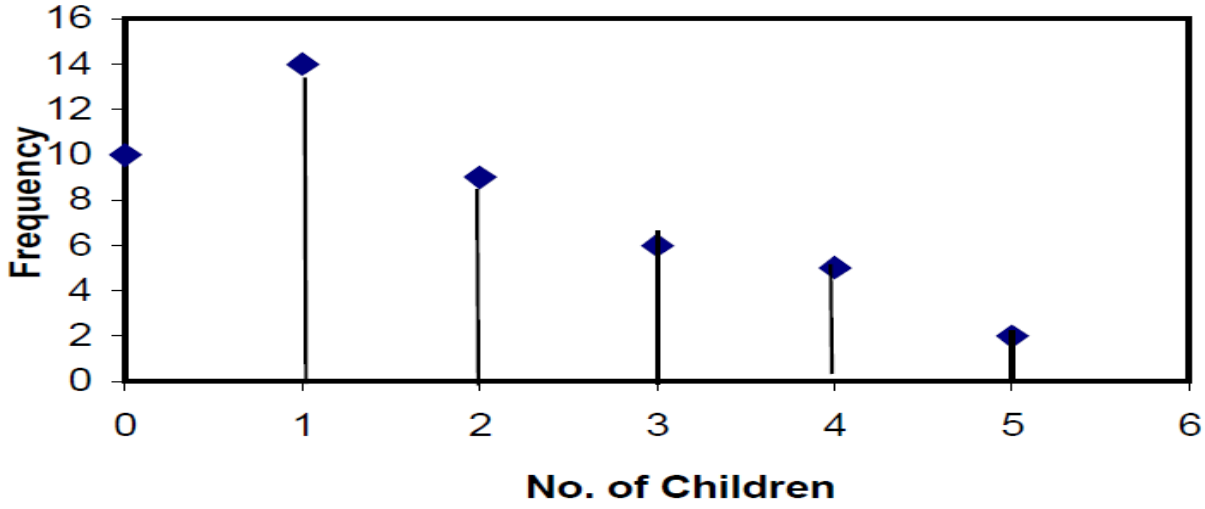
Example:

Show the following data by a line chart:

No. of children	0	1	2	3	4	5
Frequency	10	14	9	6	4	2

Solution:

Line Diagram



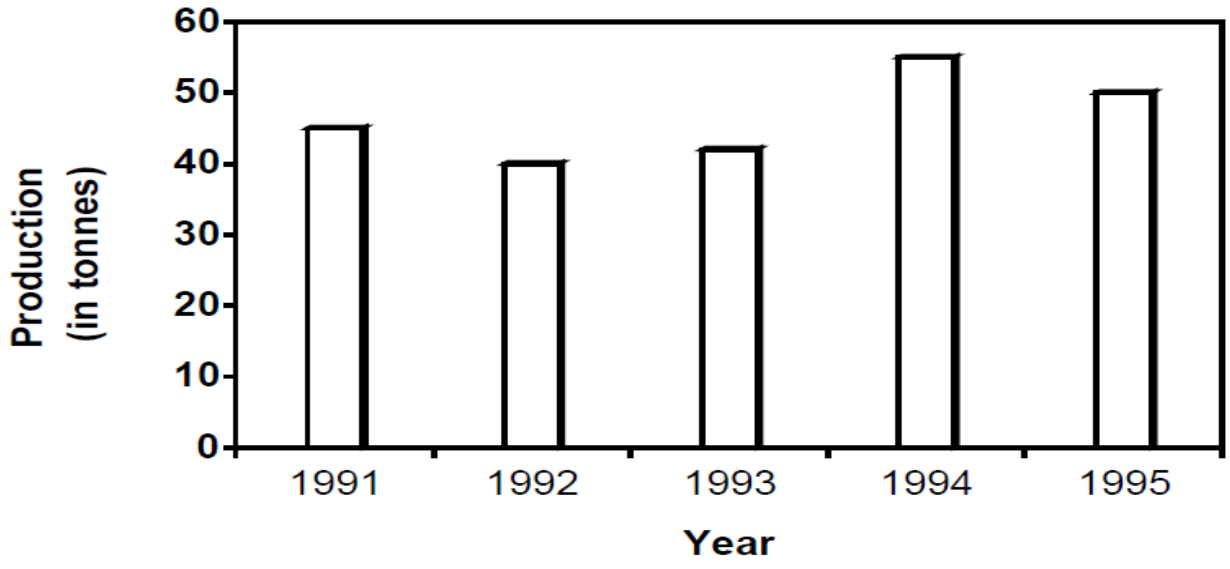
Simple Bar Diagram:
Simple bar diagram can be drawn either on horizontal or vertical base, but bars on horizontal base more common. Bars must be uniform width and intervening space between bars must be equal. While constructing a simple bar diagram, the scale is determined on the basis of the highest value in the series. To make the diagram attractive, the bars can be coloured. Bar diagram are used in business and economics. However, an important limitation of such diagrams is that they can present only one classification or one category of data. For example, while presenting the population for the last five decades, one can only depict the total population in the simple bar diagrams, and not its sex-wise distribution.

Example:
Represent the following data by a bar diagram.

Year	Production (in tones)
1991	45
1992	40
1993	42
1994	55
1995	50

Solution

Simple Bar Diagram



Sub-divided Bar Diagram:

In a sub-divided bar diagram, the bar is sub-divided into various parts in proportion to the values given in the data and the whole bar represent the total. Such diagrams are also called Component Bar diagrams. The sub divisions are distinguished by different colours or crossings or dottings. The main defect of such a diagram is that all the parts do not have a common base to enable one to compare accurately the various components of the data.

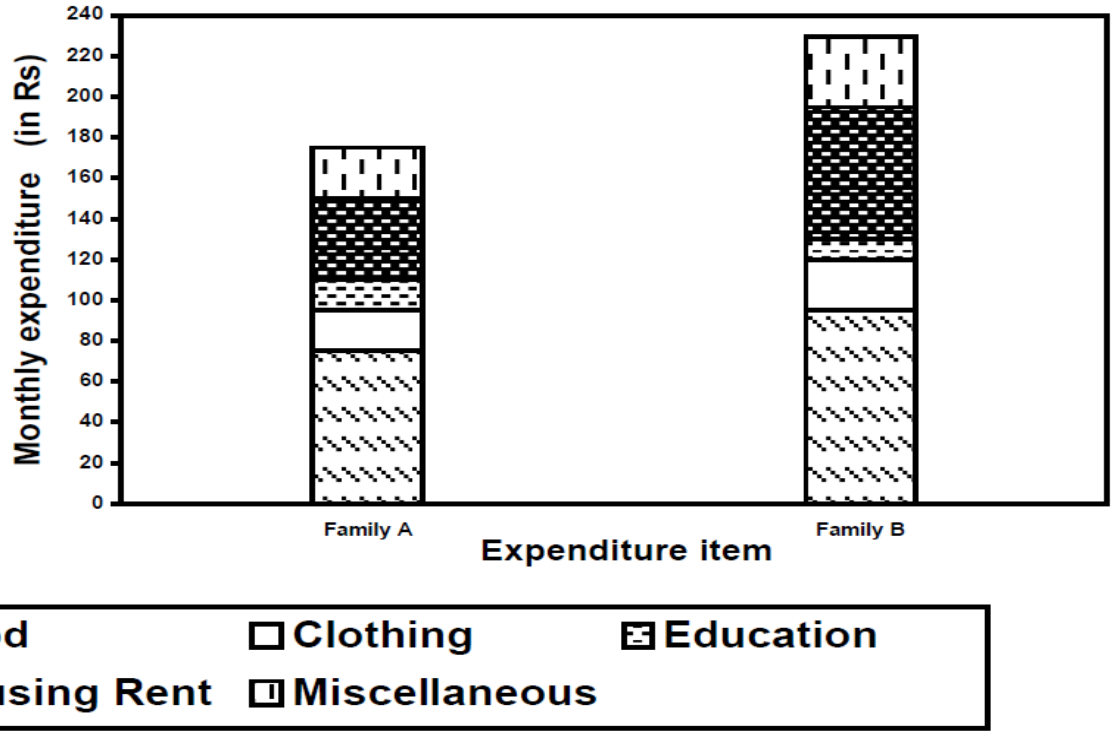
Example :

Represent the following data by a bar diagram.

Expenditure items	Monthly expenditure (in Rs.)	
	Family A	Family B
Food	75	95
Clothing	20	25
Education	15	10
Housing Rent	40	65
Miscellaneous	25	35

Solution

Sub-divided Bar Diagram



Percentage bar diagram:

This is another form of component bar diagram. Here the components are not the actual values but percentages of the whole. The main difference between the sub-divided bar diagram and percentage bar diagram is that in the former the bars are of different heights since their totals may be different whereas in the latter the bars are of equal height since each bar represents 100 percent. In the case of data having sub-division, percentage bar diagram will be more appealing than sub-divided bar diagram.

Example:

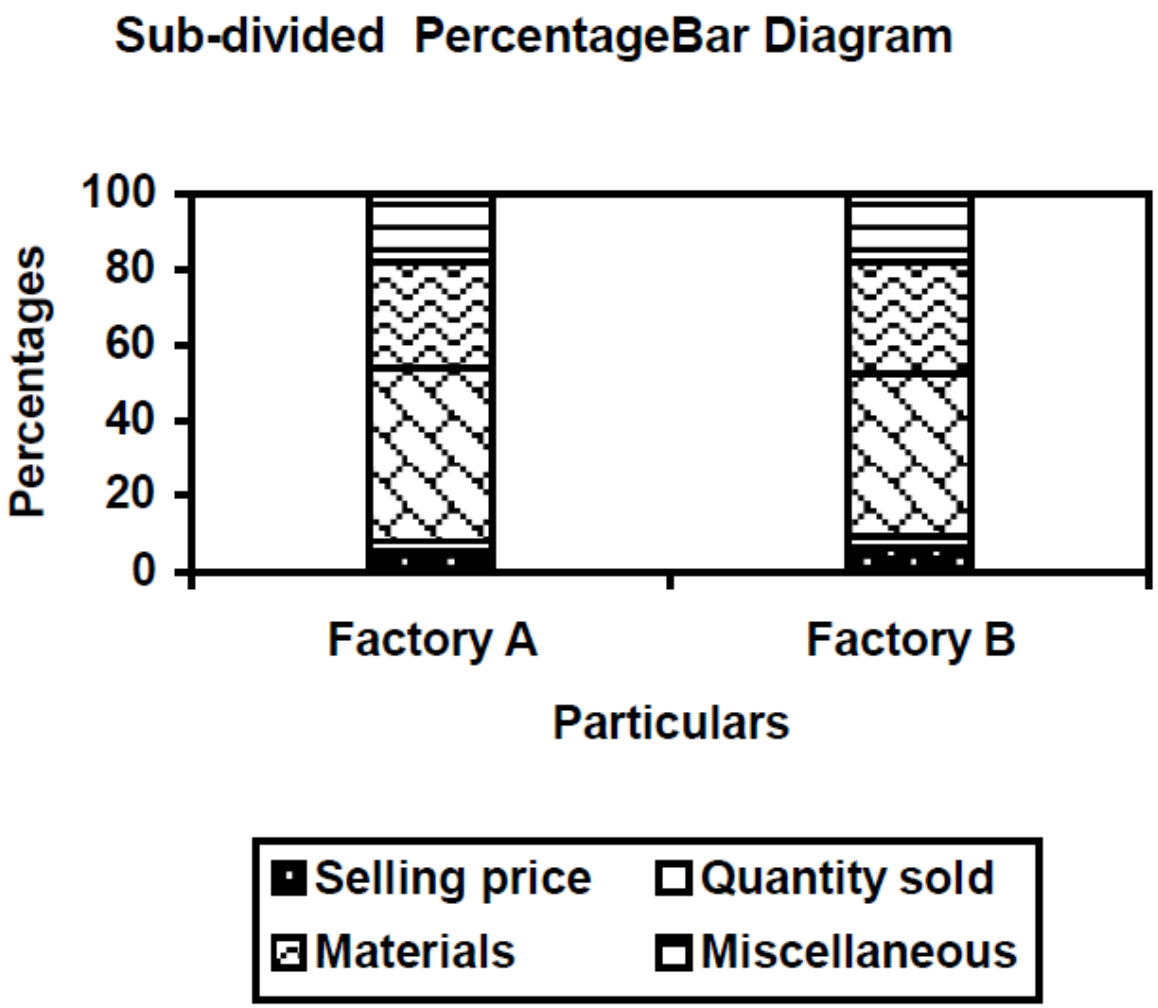
Represent the following data by a percentage bar diagram.

Particular	Factory A	Factory B
Selling Price	400	650
Quantity Sold	240	365
Wages	3500	5000
Materials	2100	3500
Miscellaneous	1400	2100

Solution:

Convert the given values into percentages as follows:

Particulars	Factory A		Factory B	
	Rs.	%	Rs.	%
Selling Price	400	5	650	6
Quantity Sold	240	3	365	3
Wages	3500	46	5000	43
Materials	2100	28	3500	30
Miscellaneous	1400	18	2100	18
Total	7640	100	11615	100



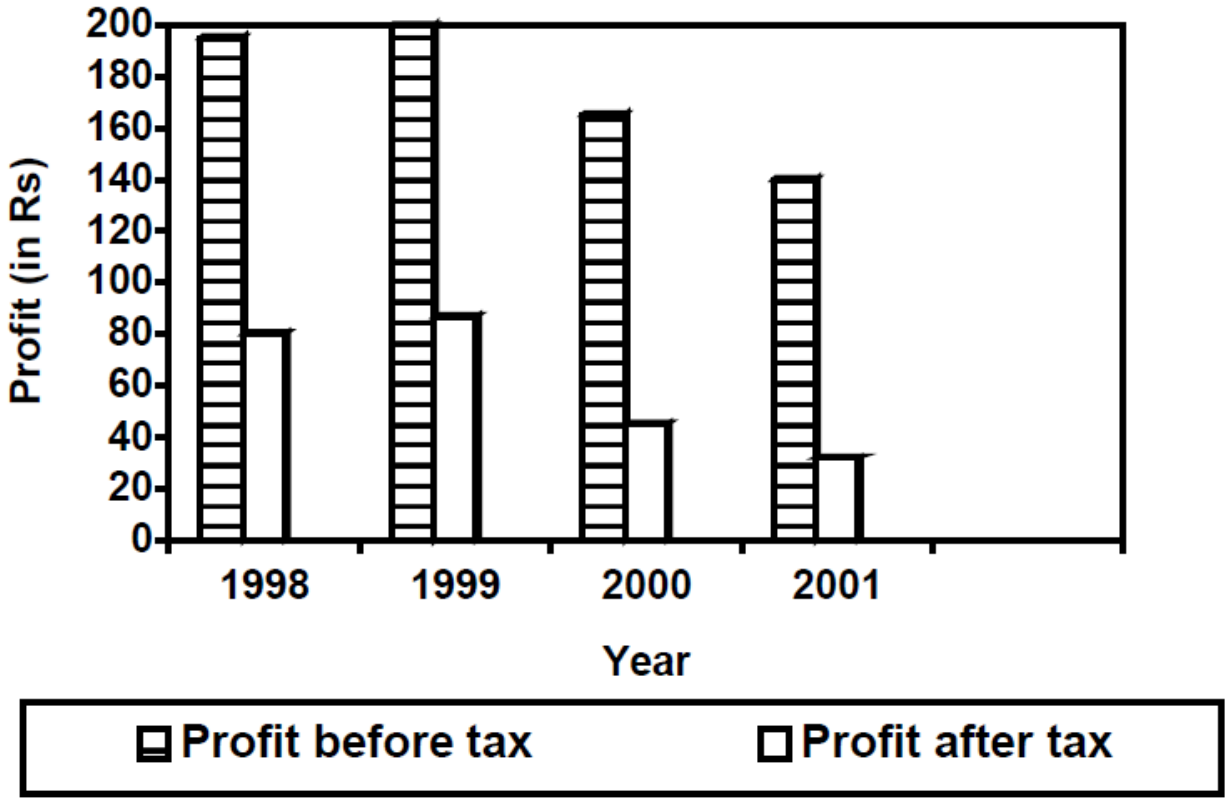
Multiple Bar Diagram:

Multiple bar diagram is used for comparing two or more sets of statistical data. Bars are constructed side by side to represent the set of values for comparison. In order to distinguish bars, they may be either differently coloured or there should be different types of crossings or dotting, etc. An index is also prepared to identify the meaning of different colours or dottings.

Year	Profit before tax (in lakhs of rupees)	Profit after tax (in lakhs of rupees)
1998	195	80
1999	200	87
2000	165	45
2001	140	32

Solution:

Multiple Bar Diagram



Pie Diagram or Circular Diagram:

Another way of preparing a two-dimensional diagram is in the form of circles. In such diagrams, both the total and the component parts or sectors can be shown. The area of a circle is proportional to the square of its radius. While making comparisons, pie diagrams should be used on a percentage basis and not on an absolute basis. In constructing a pie diagram the first step is to prepare the data so that various components values can be transposed into corresponding degrees on the circle.

The second step is to draw a circle of appropriate size with a compass. The size of the radius depends upon the available space and other factors of presentation. The third step is to measure points on the circle and representing the size of each sector with the help of a protractor.

Example:

Draw a Pie diagram for the following data of production of sugar in quintals of various countries.

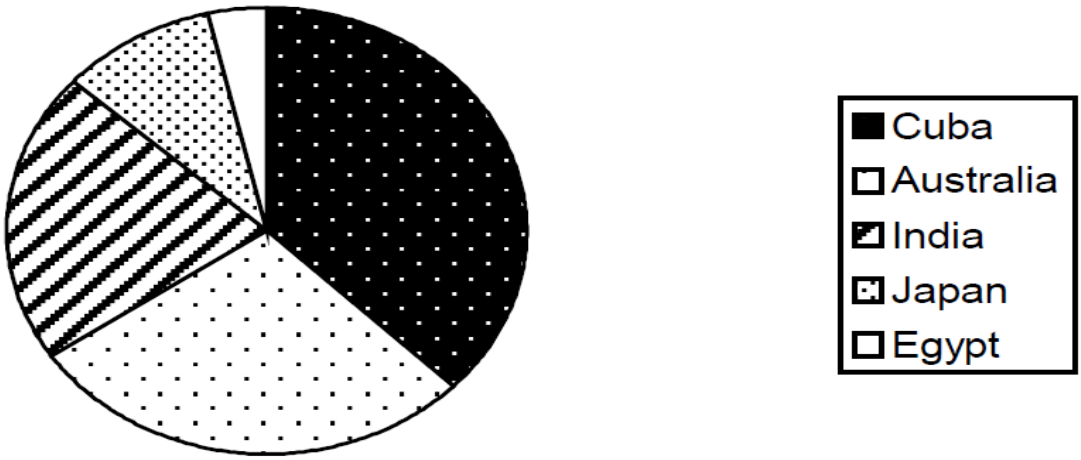
Country	Production of Sugar (in quintals)
Cuba	62
Australia	47
India	35
Japan	16
Egypt	6

Solution:

The values are expressed in terms of degree as follows.

Country	Production of Sugar	
	In Quintals	In Degrees
Cuba	62	134
Australia	47	102
India	35	76
Japan	16	35
Egypt	6	13
Total	166	360

Pie diagram



Graphs:

A graph is a visual form of presentation of statistical data. A graph is more attractive than a table of figure. Even a common man can understand the message of data from the graph. Comparisons can be made between two or more phenomena very easily with the help of a graph. However here we shall discuss only some important types of graphs which are more popular and they are

1.Histogram 2. Frequency Polygon 3.Frequency Curve 4. Ogive 5. Lorenz Curve

Histogram:

A histogram is a bar chart or graph showing the frequency of occurrence of each value of the variable being analyzed. In histogram, data are plotted as a series of rectangles. Class intervals are shown on the 'X-axis' and the frequencies on the 'Y-axis'. The height of each rectangle represents the frequency of the class interval. Each rectangle is formed with the other so as to give a continuous picture. Such a graph is also called staircase or block diagram.

However, we cannot construct a histogram for distribution with open-end classes. It is also quite misleading if the distribution has unequal intervals and suitable adjustments in frequencies are not made. While constructing histograms, we should know that exclusive type of classification is to be converted if we are given inclusive type of classification and we should form the class in case mid values are given.

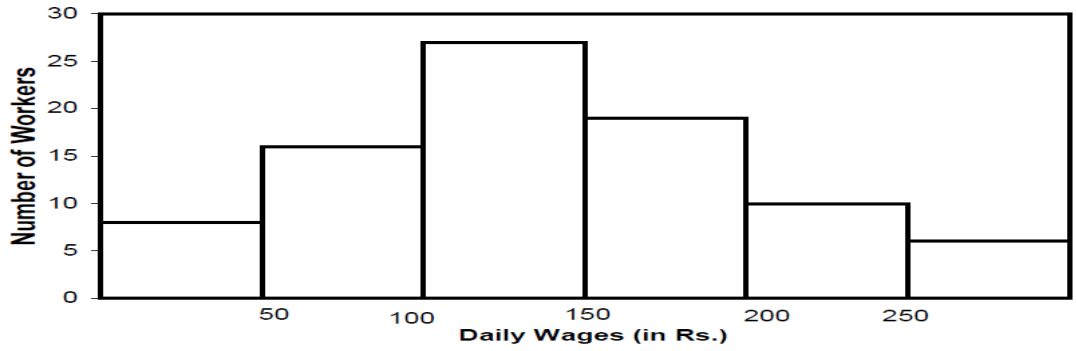
Example:

Draw a histogram for the following data.

Daily Wages	Number of Workers
0-50	8
50-100	16
100-150	27
150-200	19
200-250	10
250-300	6

Solution

Histogram



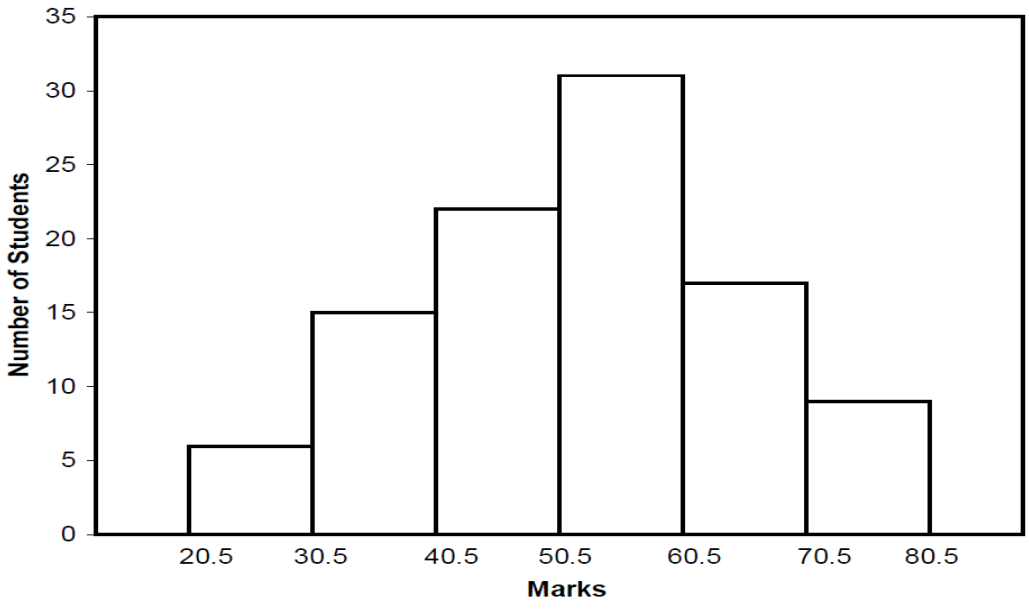
Example:
For the following data, draw a histogram.

Marks	Number of Students
21-30	6
31-40	15
41-50	22
51-60	31
61-70	17
71-80	9

Solution:
For drawing a histogram, the frequency distribution should be continuous. If it is not continuous, then first make it continuous as follows. Here Correction Factor = (31-30)/2 = 0.5

Marks	Number of Students
20.5-30.5	6
30.5-40.5	15
40.5-50.5	22
50.5-60.5	31
60.5-70.5	17
70.5-80.5	9

Histogram

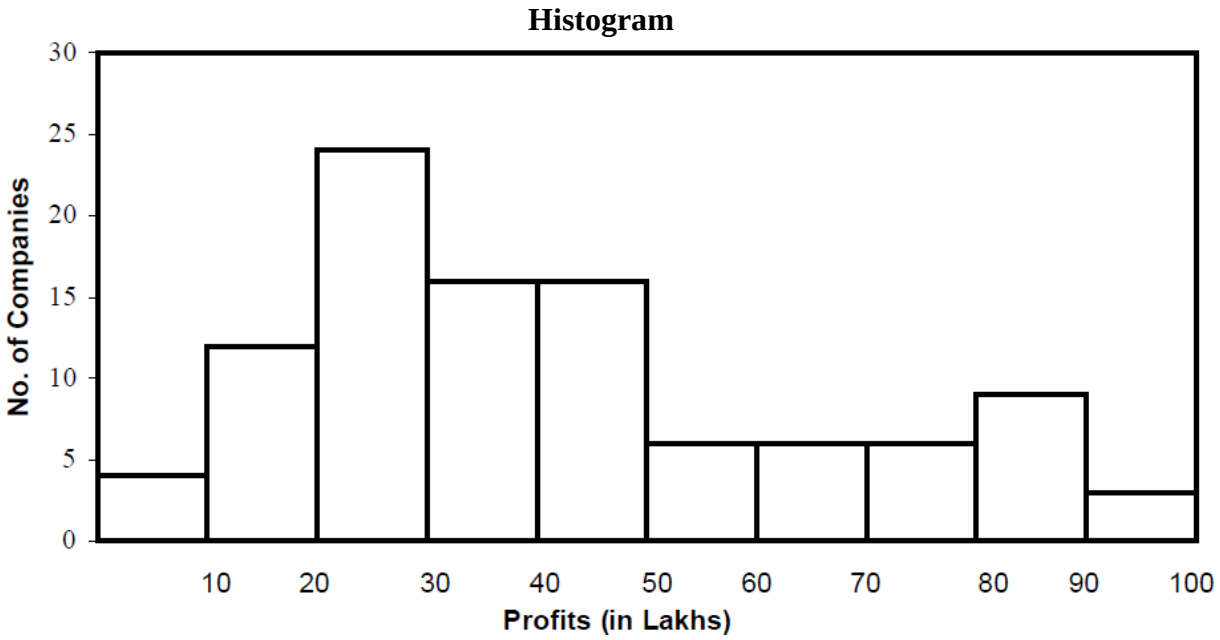


Example:
Draw a histogram for the following data.

Profits (in lakhs)	Number of Companies
0-10	4
10-20	12
20-30	24
30-50	32
50-80	18
80-90	9
90-100	3

Solution:
When the class intervals are unequal, a correction for unequal class intervals must be made. The frequencies are adjusted as follows: The frequency of the class 30-50 shall be divided by two since the class interval is in double. Similarly the class interval 50- 80 can be divided by 3. Then draw the histogram.

Profits (in lakhs)	Number of Companies
0-10	4
10-20	12
20-30	24
30-40	16
40-50	16
50-60	6
60-70	6
70-80	6
80-90	9
90-100	3



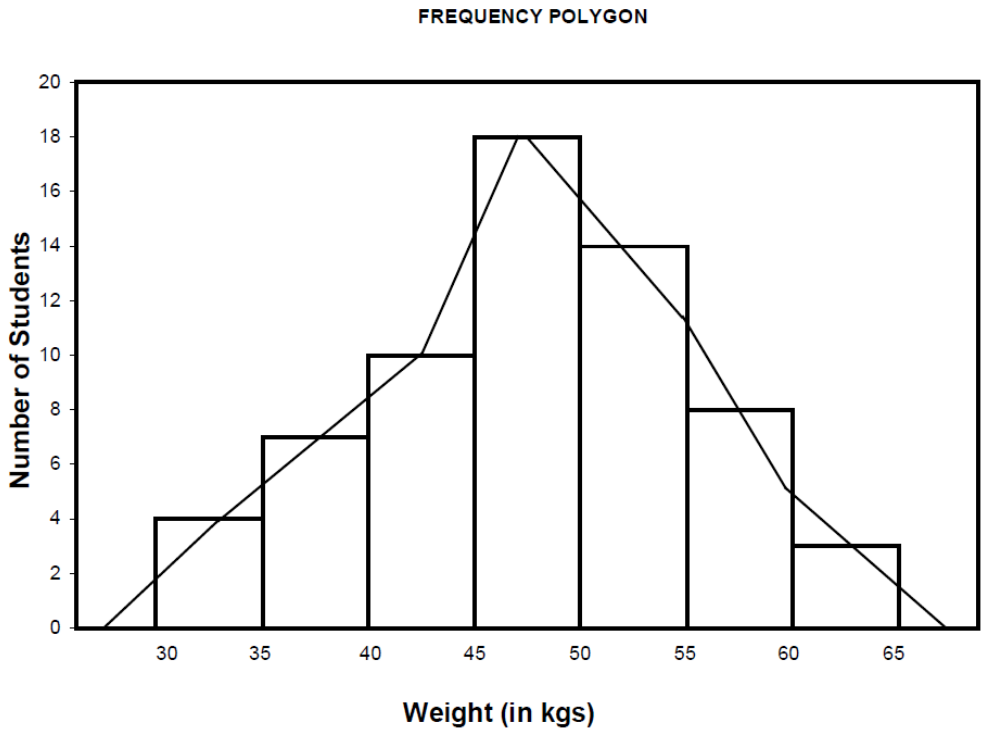
Frequency Polygon:

If we mark the midpoints of the top horizontal sides of the rectangles in a histogram and join them by a straight line, the figure so formed is called a Frequency Polygon. This is done under the assumption that the frequencies in a class interval are evenly distributed throughout the class. The area of the polygon is equal to the area of the histogram, because the area left outside is just equal to the area included in it.

Example:

Draw a frequency polygon for the following data.

Weight (in kg)	Number of Students
30-35	4
35-40	7
40-45	10
45-50	18
50-55	14
55-60	8
60-65	3



Frequency Curve:

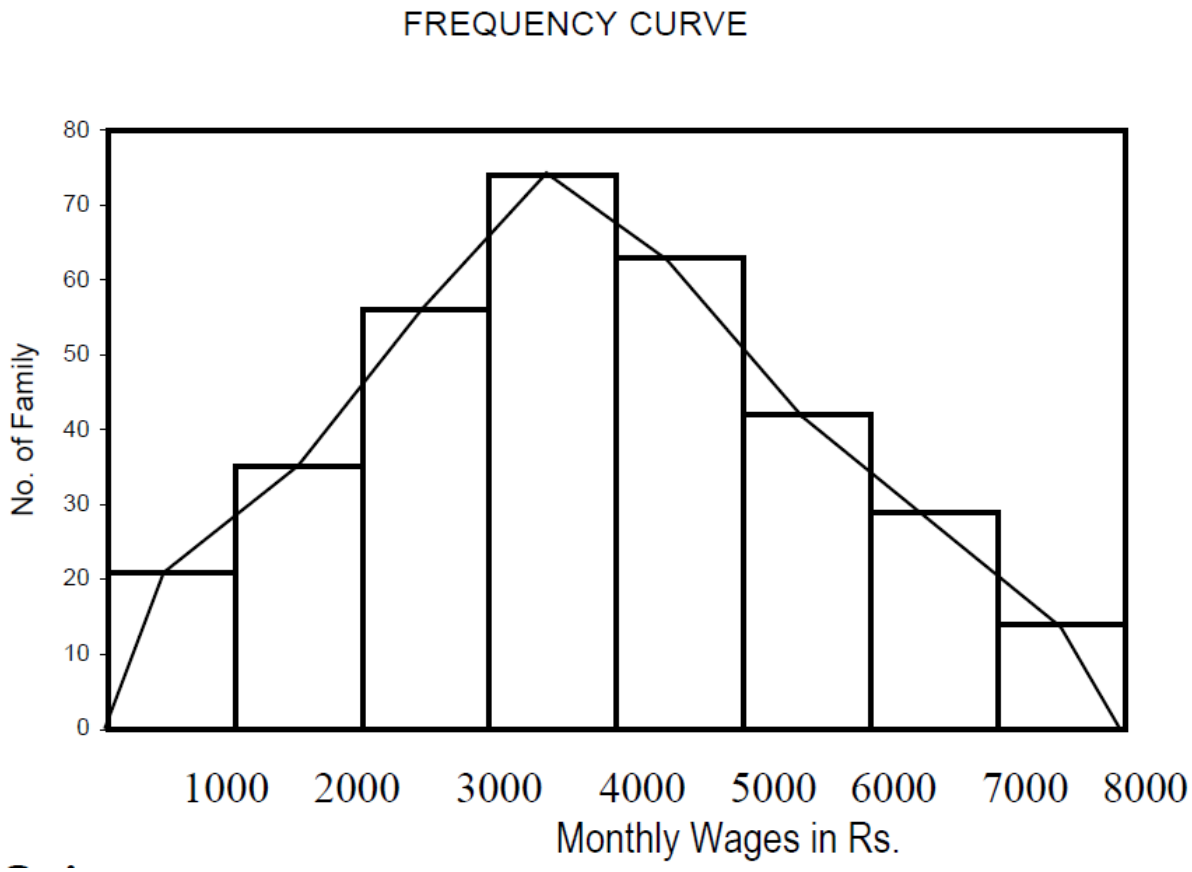
If the middle point of the upper boundaries of the rectangles of a histogram is corrected by a smooth freehand curve, then that diagram is called frequency curve. The curve should begin and end at the base line.

Example:

Draw a frequency curve for the following data.

Monthly Wages (in Rs.)	No. of family
0-1000	21
1000-2000	35
2000-3000	56
3000-4000	74
4000-5000	63
5000-6000	40
6000-7000	29
7000-8000	14

Solution



Ogives:

For a set of observations, we know how to construct a frequency distribution. In some cases we may require the number of observations less than a given value or more than a given value.
This is obtained by a accumulating (adding) the frequencies upto (or above) the give value. This accumulated frequency is called cumulative frequency. These cumulative frequencies are then listed in a table is called cumulative frequency table. The curve table is obtained by plotting cumulative frequencies is called a cumulative frequency curve or an ogive.

There are two methods of constructing ogive namely:
1. The ‘ less than ogive’ method
2. The ‘more than ogive’ method.

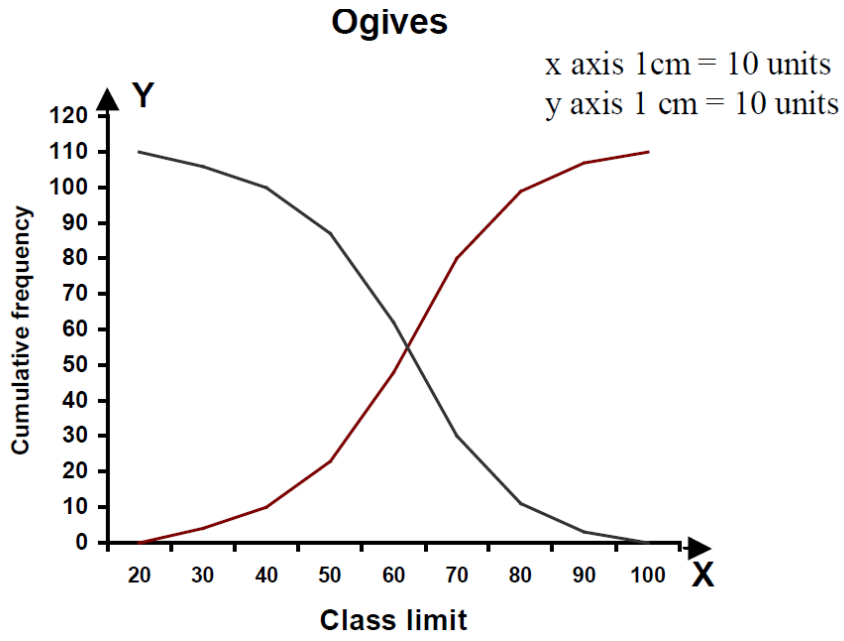
In less than ogive method we start with the upper limits of the classes and go adding the frequencies. When these frequencies are plotted, we get a rising curve. In more than ogive method, we start with the lower limits of the classes and from the total frequencies we subtract the frequency of each class. When these frequencies are plotted we get a declining curve.

Example:
Draw the Ogives for the following data.

Class interval	Frequency
20-30	4
30-40	6
40-50	13
50-60	25
60-70	32
70-80	19
80-90	8
90-100	3

Solution:

Class limit	Less than ogive	More than ogive
20	0	110
30	4	106
40	10	100
50	23	87
60	48	62
70	80	30
80	99	11
90	107	3
100	110	0



Lorenz Curve:

Lorenz curve is a graphical method of studying dispersion. It was introduced by Max.O.Lorenz, a great Economist and a statistician, to study the distribution of wealth and income. It is also used to study the variability in the distribution of profits, wages, revenue, etc.

It is specially used to study the degree of inequality in the distribution of income and wealth between countries or between different periods. It is a percentage of cumulative values of one variable in combined with the percentage of cumulative values in other variable and then Lorenz curve is drawn.

The curve starts from the origin (0,0) and ends at (100,100). If the wealth, revenue, land etc are equally distributed among the people of the country, then the Lorenz curve will be the diagonal of the square. But this is highly impossible.

The deviation of the Lorenz curve from the diagonal, shows how the wealth, revenue, land etc are not equally distributed among people.

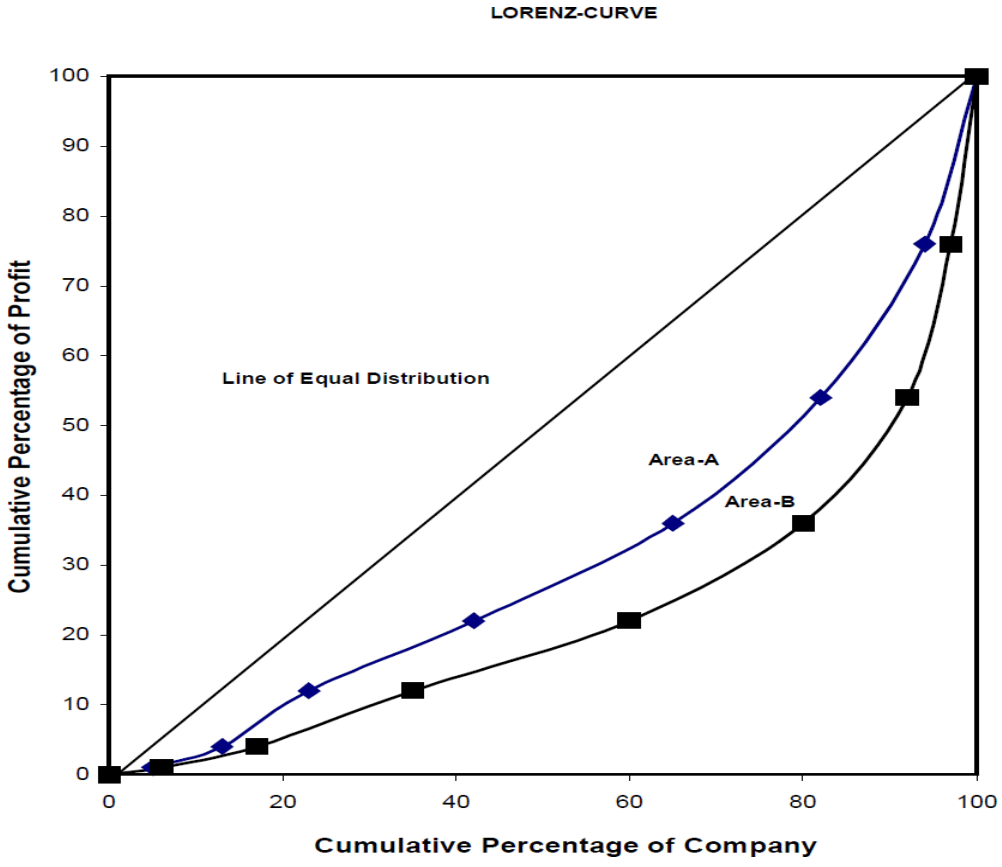
Example:

In the following table, profit earned is given from the number of companies belonging to two areas A and B. Draw in the same diagram their Lorenz curves and interpret them.

Profit earned (in thousands)	Number of Companies	
	Area A	Area B
5	7	13
26	12	25
65	14	43
89	28	57
110	33	45
155	25	28
180	18	13
200	8	6

Solution:

Profits			Area A			Area B		
In Rs.	Cumulative profit	Cumulative percentage	No. of companies	Cumulative number	Cumulative percentage	No. of companies	Cumulative number	Cumulative percentage
5	5	1	7	7	5	13	13	6
26	31	4	12	19	13	25	38	17
65	96	12	14	33	23	43	81	35
89	185	22	28	61	42	57	138	60
110	295	36	33	94	65	45	183	80
155	450	54	25	119	82	28	211	92
180	630	76	18	137	94	13	224	97
200	830	100	8	145	100	6	230	100



CHAPTER
MEASURES OF CENTRAL TENDENCY

Meaning:

In the study of a population with respect to one in which we are interested we may get a large number of observations. It is not possible to grasp any idea about the characteristic when we look at all the observations. So it is better to get one number for one group. That number must be a good representative one for all the observations to give a clear picture of that characteristic. Such representative number can be a central value for all these observations. This central value is called a measure of central tendency or an average. Measures of Central Tendency are also called the measures of location since they enable us to locate the position or place of the distribution. The general idea behind this measure of central tendency is that to look for a common measure that best describe or represents the characteristics of the entire group. This typical central value is a focal point around which large amount of data try to concentrate.

There are five averages. Among them mean, median and mode are called simple averages and the other two averages geometric mean and harmonic mean are called special averages. The meaning of average is nicely given in the following definitions.

"A measure of central tendency is a typical value around which other figures congregate."
"An average stands for the whole group of which it forms a part yet represents the whole."
"One of the most widely used set of summary figures is known as measures of location."

Requisites for a good or an ideal average:

The following properties should possess for a good average:

1. It should be rigidly defined.
2. It should be easy to understand and compute.
3. It should be based on all observations.
4. Its definition shall be in the form of a mathematical formula.
5. It should be suitable for further mathematical treatment.
6. It should be fluctuated least from sample to sample drawn from the same population.
7. It should be least affected by the extreme observations.
8. A good average should represent maximum characteristics of the data, its value should be nearest to the most items of the given series.

Types of Measures of Central Tendency

There are five types of the most commonly used measures of Central tendency which are as follows:

1. Arithmetic mean (A.M.)
   - (i) Simple Arithmetic mean
   - (ii) Weighted Arithmetic mean
2. Median (M_d)
3. Mode (M_o)
4. Geometric Mean (G.M.)
5. Harmonic Mean (H.M.)

Arithmetic mean or mean:

Arithmetic mean or simply the mean of a variable is defined as the sum of all observations divided by total number of observations. It is the most popular and widely applicable measures of central Tendency. It can be divided into two Categories:

- Simple Arithmetic mean
- Weighted Arithmetic mean

Simple Arithmetic mean:

The sum of all the observations divided by the number of observations is called simple arithmetic mean. In simple arithmetic mean, all the observations or items are equally important.

(a) **A.M for individual series:** Let x_1 x_2 ..... x_n be a variate values of the variable X, then their arithmetic mean (A.M.) is defined as

$$\text{A.M. } (\bar{x}) = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum x_i}{n} \text{ or simply } \frac{\sum x}{n} \text{ (**Direct method**)}$$

Where, $\sum xi = \sum x$= Sum of all observations.

n = Total number of observations.

Let, d = x – A then

$$\boxed{\text{A.M } (\bar{x}) = A + \frac{\sum d}{n}}\text{, (**Deviation Method**)}$$

Where, A = Assumed mean

d = Deviation from assumed mean A.

(b) **A.M for discrete series:** Let x_1 x_2 .....x_n be the variate values of the variable X with their respective frequencies f_1, f_2, ... , f_n then their A.M. is defined as

$$\text{A.M } (\bar{x}) = \frac{f_1 x_1 + f_2 x_2 + ... + f_n x_n}{f_1 + f_2 + ... + f_n} = \frac{\sum f_i x_i}{N} \text{ or simply } \frac{\sum fx}{N} \text{ (**Direct method**)}$$

Let d = x – A then

$$A.M.(\bar{X}) = A + \frac{\sum fd}{N} \text{ (**Deviation Method**)}$$

Where, A = Assumed mean

d = Deviation from the assumed mean A.

(c) **A.M for continuous series:** Let x_1 x_2 ..... x_n be the mid-value of the continuous variable x with their respective frequencies f_1, f_2, ... , f_n then

$$\text{A.M. } (\bar{x}) = \frac{f_1 x_1 + f_2 x_2 + ... + f_n x_n}{f_1 + f_2 + ... + f_n} = \frac{\sum f_i x_i}{N} \text{ or simply } \frac{\sum fx}{N} \text{ (**Direct method**)}$$

Where $\sum f$ = Total frequency = N

Let $d' = \frac{x - A}{h}$, Where h = Class - interval.

Then, A.M. $(\bar{x}) = A + \frac{\sum fd'}{N} \times h$ (**Step – deviation method**)

Weight Arithmetic Mean:

For calculating simple arithmetic mean, we suppose that all the values or the sizes of items in the distribution have equal importance. But, in practical life this may not be always true. In case some items are more important than others, a simple average computed is not representative of the distribution. Proper weightage has to be given to the various items. . For example, to have an idea of the change in cost of living of a certain group of persons, the simple average of the prices of the commodities consumed by them will not do because all the commodities are not equally important, e.g. rice, wheat and pulses are more important than tea, confectionery etc., It is the weighted arithmetic average which helps in finding out the average value of the series after giving proper weight to each group.

Let $x_1, x_2, \ldots\ldots, x_n$ be a variate values with their respective weight $w_1, w_2, \ldots\ldots, w_n$ then their weighted arithmetic mean is denoted by $\bar{x}_w$ and is defined as

$$\bar{x}_w = \frac{w_1 x_1 + w_2 x_2 + \ldots\ldots + w_n x_n}{w_1 + w_2 + \ldots\ldots + w_n}$$

$$\boxed{\bar{x}_w = \frac{\sum wx}{\sum w}}$$

Example: Calculate the mean for 2, 4, 6, 8, 10.

Solution:

Mean $= \bar{x} = \frac{2+4+6+8+10}{5} = 6$

Example: A student's marks in 5 subjects are 75, 68, 80, 92, and 56. Find his average mark using deviation method.

Solution:

Let X be the marks, and assumed mean (A) = 68

Calculation of Mean

X	**d = x – A (68)**
75	7
68	0
80	12
92	24
56	-12
Total	**31**

Here, n = 5, A = 68,

We know $\boxed{A.M\left(\bar{x}\right) = A + \frac{\sum d}{n}}$`

$= 68 + \frac{31}{5}$

$= 68 + 6.2$

$= 74.2$

Hence, mean marks = 74.2

Example: From the following data, determine the average age by using deviation method.

Age(years)	10	20	30	40	50	60	70
No. of People	20	15	12	10	4	8	6

Solution:

Let X be the age of the people and 40 be the assumed mean i.e. A.

Calculation of average age

Age (X)	**d = x –A(40)**	**f**	**fd**
10	–30	20	–600
20	–20	15	–300
30	–10	12	–120
40	0	10	0
50	10	4	40
60	20	8	160
70	30	6	180
		N = 75	$\sum$fd = –640

We know, $A.M.(\bar{X}) = A + \frac{\sum fd}{N}$

$\therefore$ A.M $(\bar{x}) = 40 + \frac{(-640)}{75} = 40 - 8.53$

A.M $\left(\bar{x}\right) = 31.47$

Hence, average age = 31.47 years

Example: Data below represents the wages (Rs) received by workers in construction site. Calculate the average wages using step deviation method.

Wages:	25-30	30-35	35-40	40-45	45-50	50-55	55-60	60-65	65-70
No. of workers:	10	13	18	21	24	28	20	11	8

Solution: Calculation of average wages using changing origin and scale of data

Wages (Rs)	**Mid-Value (X)**	**$d' = \frac{x-47.5}{5}$**	**f**	**fd'**
25-30	27.5	–4	10	–40
30-35	32.5	–3	13	–39
35-40	37.5	–2	18	–36
40-45	42.5	–1	21	–21

45-50	47.5	0	24	0
50-55	52.5	+1	28	+28
55-60	57.5	+2	20	+40
60-65	62.5	+3	11	+33
65-70	67.5	+4	8	+32
			N = 153	$\sum fd' = -3$

We know, A.M. $(\bar{x}) = A + \frac{\sum fd'}{N} \times h$, where, A = 47.5, h = 5,

$= 47.5 + \frac{(-3)}{153} \times 5 = 47.40$

Hence, mean wages is Rs 47.40.

Example : Given the following frequency distribution, calculate the arithmetic mean.

Wages (Rs)	64	63	62	61	60	59
Number of People	8	18	12	9	7	6

Solution:

Let X be the wages and f be the number of students.

x	**f**	**fx**	**d=x-A(62)**	**fd**
64	8	512	2	16
63	18	1134	1	18
62	12	744	0	0
61	9	549	-1	-9
60	7	420	-2	-14
59	6	354	-3	-18
	N=60	**3713**		**-7**

Direct method

$A.M.(\bar{X}) = \frac{\sum fx}{N} = \frac{3713}{60} = 61.88$

Short-cut method

$A.M.(\bar{X}) = A + \frac{\sum fd}{N} = 61.88$

Example: For a certain frequency table of number of accident and number of days which is only partly reproduced here, the average number of accidents was found to be 1.46.

Number of accident	Number of days
0	46
1	?
2	?
3	25
4	10
5	5

Calculate the missing frequencies. Where N = 200.

Solution:

Computation of missing frequencies

No. of accidents (X)	No. of days (f)	fx
0	46	0
1	f_1	f_1
2	f_2	$2f_2$
3	25	75
4	10	40
5	5	25
	$N = 86 + f_1 + f_2$	$\sum fx = 140 + f_1 + 2f_2$

$\therefore$ Here, $N = 86 + f_1 + f_2 = 114$

Or, $200 = 86 + f_1 + f_2$

Or, $f_1 + f_2 = 114$

Or, $f_1 = 114 - f_2$ ..... (i)

Now, $\bar{x} = \frac{\sum fx}{N} \Rightarrow 1.46 = \frac{140 + f_1 + 2f_2}{200}$

Or, $140 + f_1 + 2f_2 = 200 \times 1.46$

$\therefore$ $f_1 + 2f_2 = 152$ ..... (ii)

Putting the value of f_1 from (i) to (ii)

We get, $114 - f_2 + 2f_2 = 152$

Or $f_2 = 38$

Putting the value of f_2 in equation (i),

We get, $f_1 = 114 - 38 = 76$

Hence, $f_1 = 76$, $f_2 = 38$

Example: The mean of 200 items was 50. Later on it was discovered that two items were misread as 92 and 8 instead of 192 and 88. Find out the correct mean.

Solution: We have given,

Incorrect mean $(\bar{x}) = 50$, and Number of observations $(n) = 200$

We know, $\bar{x} = \frac{\sum x}{n} \Rightarrow \sum x = n\bar{x} = 200 \times 50$

$\sum x = 10000$

Now,

Correct $\sum x = 10000 + 192 + 88 - 92 - 8$

Correct $\sum x = 10180$

$\therefore$ Correct $\bar{x} = \frac{\text{Correct } \sum x}{n} = \frac{10180}{200}$

Correct $\bar{x} = 50.9$

Therefore, required correct mean is 50.9

Merits and Demerits of Arithmetic Mean:

Merits:

1. It is rigidly defined.
2. It is easy to understand and easy to calculate.
3. If the number of items is sufficiently large, it is more accurate and more reliable.
4. It is a calculated value and is not based on its position in the series.
5. It is possible to calculate even if some of the details of the data are lacking.
6. Of all averages, it is affected least by fluctuations of sampling.
7. It provides a good basis for comparison.
8. Arrangement of data is not required for computing Arithmetic mean.
9. It is suitable for further mathematical treatment.

Demerits:

1. It cannot be obtained by inspection nor located through a frequency graph.
2. It cannot be in the study of qualitative phenomena not capable of numerical measurement i.e. Intelligence, beauty, honesty etc.,
3. It is not a suitable measure of central value in case of highly skewed distribution.
4. It is very much affected by extreme values.
5. It cannot be calculated for open-end classes.

Harmonic mean (H.M) :

Harmonic mean of a set of observations is defined as the reciprocal of the arithmetic average of the reciprocal of the given values.

H.M for individual series: If $x_1, x_2 \ldots . x_n$ are n non-zero observations, then H.M. is given by

$$H.M. = \frac{n}{\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)}$$

H.M. for discrete series: Let $x_1, x_2, \ldots\ldots, x_n$ be non-zero variate values with their correspondings frequencies $f_1, f_2, \ldots\ldots, f_n$, then their H.M. is defined as

$$H.M. = \frac{N}{\sum_{i=1}^{n} f_i\left(\frac{1}{x_i}\right)}, \text{ where } \sum f = N$$

H.M. for continuous series:

Let $x_1, x_2, \ldots\ldots, x_n$ be the mid-values of a variable of the classes with their corresponding frequencies $f_1, f_2, \ldots\ldots, f_n$, then their H. M. is given by

$$H.M = \frac{N}{\sum f\frac{1}{x}}, \text{ where } \sum f = N$$

Example: From the given data, calculate H.M.

5, 10, 17, 24, 30

Solution:

x	**1/x**
5	0.2000
10	0.1000
17	0.0588
24	0.0417
30	0.0333
Total	**0.4338**

We know,

$$H.M. = \frac{n}{\sum\frac{1}{x}} = 5/0.4338 = 11.526$$

Example: Find the harmonic mean for the following data:

x:	5	10	15	20	25
f:	4	6	8	5	2

Calculation of H.M

X	**f**	$\mathbf{f\frac{1}{x}}$
5	4	0.800
10	6	0.600
15	8	0.534
20	5	0.250
25	2	0.080
	N = 25	$\sum f\frac{1}{x} = 2.264$

Here, n = 25, $\sum f\frac{1}{x} = 2.2644$

$$H.M. = \frac{N}{\sum f\frac{1}{x}} = \frac{25}{2.264}$$

$\therefore$ H.M. = 11.042

Example: Ages of some people of a village are given below. Calculate the harmonic mean of the age of these people.

Age(years)	20	21	22	23	24	25
Number of People	4	2	7	1	3	1

Solution:

Age(x)	**Number of People(f)**	$\frac{1}{x}$	$f\left(\frac{1}{x}\right)$
20	4	0.0500	0.2000
21	2	0.0476	0.0952
22	7	0.0454	0.3178
23	1	0.435	0.0435
24	3	0.0417	0.1251
25	1	0.0400	0.0400
Total	**N=18**		**0.8216**

We know,

$$\text{H.M.} = \frac{N}{\sum f\frac{1}{x}}, \text{ where } \sum f = N$$

=21.9

Merits of H.M. :

1. It is rigidly defined.
2. It is defined on all observations.
3. It is amenable to further algebraic treatment.
4. It is the most suitable average when it is desired to give greater weight to smaller observations and less weight to the larger ones.

Demerits of H.M. :

1. It is not easily understood.
2. It is difficult to compute.
3. It is only a summary figure and may not be the actual item in the series
4. It gives greater importance to small items and is therefore, useful only when small items have to be given greater weightage.

Geometric mean:

G.M. for individual series:

The geometric mean of a series containing n observations is the n^{th} root of the product of the values. If $x_1, x_2, \ldots, x_n$ are n observations then

$$G.M. = \sqrt[n]{x_1.x_2 \ldots\ldots x_n} = (x_1.x_2 \ldots\ldots x_n)^{\frac{1}{n}}$$

$$log(G.M.) = \frac{1}{n} log(x_1.x_2 \ldots\ldots x_n)$$

$$= \frac{1}{n}(\log x_1 + \log x_2 + \cdots .. + log\, x_n)$$

$$= \frac{\sum log\, x_i}{n}$$

$$G.M. = Antilog\ \frac{\sum log\, x_i}{n}$$

G.M. for discrete series:

Let $x_1, x_2, \ldots\ldots, x_n$ be the variate values with their respective frequency $f_1, f_2, \ldots\ldots, f_n$ then their G.M. is given by

$$G.M. = Antilog\ \frac{\sum f\ logx}{N}, \text{ where } \sum f = N$$

G.M. for continuous series:

Let $x_1, x_2, \ldots\ldots, x_n$ be the mid - values of the classes with their respective frequencies $f_1, f_2, \ldots\ldots, f_n$ then their G.M. is given by

$$G.M. = Antilog\ \frac{\sum f\ logx}{N}, \text{ where } \sum f = N$$

Example: Calculate the geometric mean of the following series of monthly income of a batch of families 180, 250, 490, 1400, and 1050.

Solution:

Let X be the monthly income

x	**Log x**
180	2.2553
250	2.3979
490	2.6902
1400	3.1461
1050	3.0212
Total	**13.5107**

We know

$$G.M. = Antilog\ \frac{\sum log\, x_i}{n}$$

$$= Antilog\ \frac{13.5107}{5}$$

$$= Antilog\ 2.7021$$

$$= 503.6$$

Example: Find the geometric mean of the following data:

Age	3	6	8	11	13
Number of Students	4	8	5	3	2

Solution:

Let X be the age of the students.

Calculation of Geometric mean

x	f	log x	f log x
3	4	0.4771	1.9084
6	8	0.7781	6.2248
8	5	0.9030	4.5150
11	3	1.0414	3.1242
13	2	1.1140	2.2280
Total	N = 22		$\sum$ f logx = 18.0004

Here, N = 22, $\sum$f. log x = 18.0004

G.M. = Antilog $\left[\frac{\sum f.\ \log x}{N}\right]$ = Antilog $\left[\frac{18.0004}{22}\right]$ = Antilog (0.8182) = 6.578

$\therefore$ G.M. = 6.578 years

Merits of Geometric mean:

1. It is rigidly defined
2. It is based on all items
3. It is very suitable for averaging ratios, rates and percentages
4. It is capable of further mathematical treatment.
5. Unlike AM, it is not affected much by the presence of extreme values

Demerits of Geometric mean:

1. It cannot be used when the values are negative or if any of the observations is zero
2. It is difficult to calculate particularly when the items are very large or when there is a frequency distribution.
3. It brings out the property of the ratio of the change and not the absolute difference of change as the case in arithmetic mean.
4. The GM may not be the actual value of the series.

Combined Mean:

If the arithmetic averages and the number of items in two or more related groups are known, the combined or the composite mean of the entire group can be obtained by

Combined mean = $\bar{x}_{12} = \frac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2}$

Where, n_1 and n_2 are the number of items in two groups and $\bar{x}_1$ & $\bar{x}_2$ are the arithmetic averages of items in two groups.

The advantage of combined arithmetic mean is that, we can determine the over, all mean of the combined data without going back to the original data.

Example: Find the combined mean for the data given below

n_1=20, n_2 = 30, $\bar{x}_1$ = 4 and $\bar{x}_2$ = 3.

Solution:

Combined mean = $\bar{x}_{12}$

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

$= \frac{20\times4+30\times3}{20+30}$

$= 3.4$

Positional Averages:

These averages are based on the position of the given observation in a series, arranged in an ascending or descending order. The magnitude or the size of the values does matter as was in the case of arithmetic mean. It is because of the basic difference that the median and mode are called the positional measures of an average.

Median:

The median is that value of the variate which divides the group into two equal parts, one part comprising all values greater, and the other, all values less than median.

Ungrouped or Raw data or Individual data:

Arrange the given values in the increasing or decreasing order. If the number of values are odd, median is the middle value. If the number of values are even, median is the mean of middle two values.

By formula, Median = Value of $(n+1)/2^{th}$ item

When odd number of values are given:

Example: Find median for the following data 25, 18, 27, 10, 7, 30, 42, 20, 53

Solution:

Arranging the data in the increasing order, we get 7, 10, 18, 20, 25, 27, 30, 42, 53

The middle value is the 5^{th} item i.e., 25 is the median.

Using formula, Median = Value of $(n+1)/2^{th}$ item

= Value of $(9+1)/2^{th}$ item

= Value of 5^{th} item

= 25

When even number of values are given:

Example: Find the median for the following data 4, 8, 12, 30, 18, 10, 2, 22

Solution:

Arranging the data in the increasing order, we get 2, 4, 8, 10, 12, 18, 22, 30

Here median is the mean of the middle two items (i.e.) mean of (10, 12) i.e. = (10+12)/2= 11

Using formula, Median = Value of $(n+1)/2^{th}$ item

= Value of $(8+1)/2^{th}$ item

= Value of 4.5^{th} item

= Mean of 4^{th} and 5^{th} item

= (10+12)/2

= 11

Example*:* The following table represents the marks obtained by a batch of 10 students in certain class tests in statistics and Economics.

Serial No.	1	2	3	4	5	6	7	8	9	10
Marks (Statistics)	53	55	52	32	30	60	47	46	35	28
Marks (Economics)	57	45	24	31	25	84	43	80	32	72

Indicate in which subject is the level of knowledge higher?

Solution*:*
For such question, median is the most suitable measure of central tendency. The marks in the two subjects are first arranged in increasing order as follows:

Serial No.	1	2	3	4	5	6	7	8	9	10
Marks (Statistics)	28	30	32	35	46	47	52	53	55	60
Marks (Economics)	24	25	31	32	43	45	57	72	80	84

Using formula, Median = Value of $(n+1)/2^{th}$ item
= Value of $(10+1)/2^{th}$ item
= Value of 5.5^{th} item
= Mean of 5^{th} and 6^{th} item

So,
Median for statistics = (46+47)/2
= 46.5
Median for Economics = (43+45)/2
=44

Since the median for Statistics is greater than the median for Economics, the level of knowledge in Statistics is higher than that in Economics.

Grouped Data:
In a grouped distribution, values are associated with frequencies. Grouping can be in the form of a discrete frequency distribution or a continuous frequency distribution. Whatever may be the type of distribution, cumulative frequencies have to be calculated to know the total number of items.

Cumulative frequency (c.f.):
Cumulative frequency of each class is the sum of the frequency of the class and the frequencies of the pervious classes, i.e. adding the frequencies successively, so that the last cumulative frequency gives the total number of items.

Discrete Series:
- Find cumulative frequencies.
- Find (N+1)/2

- See in the cumulative frequencies the value just greater than (N+1)/2
- Then the corresponding value of x is median.

Example: The following data pertaining to the number of members in a family. Find median size of the family.

Number of Members (x)	1	2	3	4	5	6	7	8	9	10	11	12
Frequency(f)	1	3	5	6	10	13	9	5	3	2	2	1

Solution:

X	**F**	**c. f.**
1	1	1
2	3	4
3	5	9
4	6	15
5	10	25
6	13	38
7	9	47
8	5	52
9	3	55
10	2	57
11	2	59
12	1	60
Total	**60**	

Using formula, Median = Size of $(N+1)/2^{th}$ item
= Size of $(60+1)/2^{th}$ item
= Size of 30.5^{th} item

The cumulative frequency just greater than 30.5 is 38 and the value of x corresponding to 38 is 6. Hence the median size is 6 members per family.

Continuous Series:

The steps given below are followed for the calculation of median in continuous series.
Step1: Find cumulative frequencies.
Step2: Find N/2
Step3: See in the cumulative frequency the value first greater than N/2

Then the corresponding class interval is called the Median class. Then apply the formula

$$Median(M_d) = l + \frac{\frac{N}{2} - m}{f} \times c$$

Where, l = Lower limit of the median class
m = cumulative frequency preceding the median class
c = width of the median class
f =frequency in the median class.
N=Total frequency.

Note: If the class intervals are given in inclusive type convert them into exclusive type.

Example: The following table gives the frequency distribution of 325 workers of an industry, according to their average monthly income in a certain year.

Income group (in Rs)	Number of Workers
Below 100	1
100 – 150	20
150 – 200	42
200 – 250	55
250 – 300	62
300 – 350	45
350 – 400	30
400 – 450	25
450 – 500	15
500 – 550	18
550 – 600	10
600 and above	2
Total	325

Calculate median income.

Solution:

Income group (in Rs)	Number of Workers	c. f.
Below 100	1	1
100 – 150	20	21
150 – 200	42	63
200 – 250	55	118
250 – 300	62	180
300 – 350	45	225
350 – 400	30	255
400 – 450	25	280
450 – 500	15	295
500 – 550	18	313
550 – 600	10	323
600 and above	2	325
Total	325	

Here, N/2 = 325/2 = 162.5, so the median class is 250 – 300.
l = 250, m = 118, c = 50, f = 62
Hence,

$$\text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

= 250 + (162.5-118) x 50/62 = 285.89

Example: Calculate median from the following data

Value	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-50
Frequency	5	8	10	12	7	6	3	2

Solution:

To convert the inclusive class interval into exclusive, correction factor = (10-9)/2=0.5

So, we get the exclusive frequency distribution as follows:

Value	**Frequency**	**Exclusive class interval**	**c.f.**
5-9	5	4.5-9.5	5
10-14	8	9.5-14.5	13
15-19	10	14.5-19.5	23
20-24	12	19.5-24.5	35
25-29	7	24.5-29.5	42
30-34	6	29.5-34.5	48
35-39	3	34.5-39.5	51
40-44	2	39.5-44.5	53
	53		

Here, N/2 = 53/2 = 26.5. So, Median class = 19.5-24.5

l= 19.5, m = 23, f = 12, c = 5

So,

$$Median(M_d) = l + \frac{\frac{N}{2} - m}{f} \times c$$ = 19.5+ (26.5-23) x 5/12 = 20.96

Example: The following table shows the wage distribution of person in particular region:

Wage Below (Rs.):	10	20	30	40	50	60	70	80
No. of persons (00):	2	5	9	12	14	15	15.5	15.6

Find the median wage.

Solution:

Calculation of median

Wage (X)	**c. f.**	**f**
Below 10	2	2
10-20	5	3
20-30	9	4
30-40	12	3
40-50	14	2
50-60	15	1
60-70	15.5	0.5
70-80	15.6	0.1
		N = 15.6

Here, $\frac{N}{2} = \frac{15.6}{2} = 7.8$

The c. f. just greater than 7.8 is 9. Therefore, median lies in the class 20-30.

Then, $l = 20$, h = 10, c.f. = 5, f = 4

$$\therefore \quad \text{Md.} = l + \frac{\frac{N}{2} - \text{c.f}}{\text{f}} \times \text{h}$$

$$= 20 + \frac{7.8 - 5}{4} \times 10$$

= Rs. 27

∴ Median Wage = Rs. 27

Example: Compute median for the following data.

Mid-value	5	15	25	35	45	55	65	75
Frequency	7	10	15	17	8	4	6	7

Solution: Here, correction factor = (15-5)/2 = 5. Hence we change the given distribution with mid value as follows:

Mid-value	**Class Interval**	**Frequency**	**c.f.**
5	0-10	7	7
15	10-20	10	17
25	20-30	15	32
35	30-40	17	49
45	40-50	8	57
55	50-60	4	61
65	60-70	6	67
75	70-80	7	74
Total		**74**	

N/2 = 74/2 = 37, so median class is 30 - 40.

l= 30, m = 32, f = 17, c = 10

So,

$$Median(M_d) = l + \frac{\frac{N}{2} - m}{f} \times c$$ = 19.5+ (37-32) x 10/17 = 32.94

Graphic method for Location of median:

Median can be located with the help of the cumulative frequency curve or ‘ ogive’ . The procedure for locating median in a grouped data is as follows:

Step1: The class boundaries, where there are no gaps between consecutive classes, are represented on the horizontal axis (x-axis).

Step2: The cumulative frequency corresponding to different classes is plotted on the vertical axis (y-axis) against the upper limit of the class interval (or against the variate value in the case of a discrete series.)

Step3: The curve obtained on joining the points by means of freehand drawing is called the ‘ ogive’ . The ogive so drawn may be either a (i) less than ogive or a (ii) more than ogive.

Step4: The value of N/2 or (N+1)/2 is marked on the y-axis, where N is the total frequency.

Step5: A horizontal straight line is drawn from the point N/2 or (N+1)/2 on the y-axis parallel to x-axis to meet the ogive.

Step6: A vertical straight line is drawn from the point of intersection perpendicular to the horizontal axis.

Step7: The point of intersection of the perpendicular to the x-axis gives the value of the median.

Remarks :

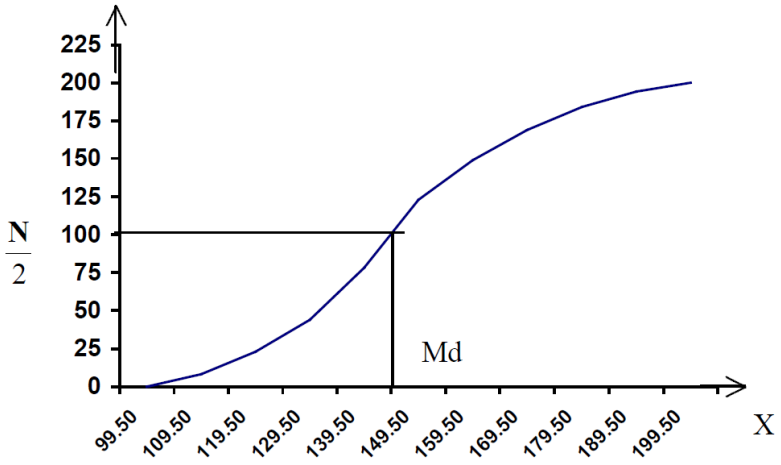
1. From the point of intersection of ‘ less than’ and ‘more than’ ogives, if a perpendicular is drawn on the x-axis, the point so obtained on the horizontal axis gives the value of the median.
2. If ogive is drawn using cumulated percentage frequencies, then we draw a straight line from the point intersecting 50 percent cumulated frequency on the y-axis parallel to the x-axis to intersect the ogive. A perpendicular drawn from this point of intersection on the horizontal axis gives the value of the median.

Example: Draw an ogive of ‘ less than’ type on the data given below and hence find median.

Weight(lbs)	Number of persons
100-109	8
110-119	15
120-129	21
130-139	34
140-149	45
150-159	26
160-169	20
170-179	15
180-189	10
190-199	6

Solution:

Class interval	No of persons	True class interval	Less than c.f
100-109	8	99.5-109.5	8
110-119	15	109.5-119.5	23
120-129	21	119.5-129.5	44
130-139	34	129.5-139.5	78
140-149	45	139.5-149.5	123
150-159	26	149.5-159.5	149
160-169	20	159.5-169.5	169
170-179	15	169.5-179.5	184
180-189	10	179.5-189.5	194
190-199	6	189.5-199.5	200

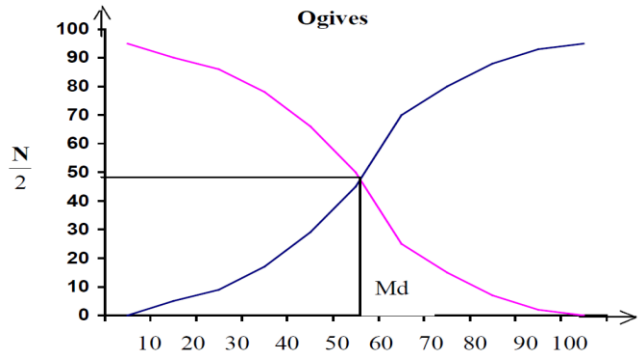


Example: Draw an ogive for the following frequency distribution and hence find median.

Marks	Number of students
0-10	5
10-20	4
20-30	8
30-40	12
40-50	16
50-60	25
60-70	10
70-80	8
80-90	5
90-100	2

Solution:

Class boundary	Cumulative Frequency	
	Less than	More than
0	0	95
10	5	90
20	9	86
30	17	78
40	29	66
50	45	50
60	70	25
70	80	15
80	88	7
90	93	2
100	95	0



Merits of Median :

1. Median is not influenced by extreme values because it is a positional average.
2. Median can be calculated in case of distribution with open end intervals.
3. Median can be located even if the data are incomplete.
4. Median can be located even for qualitative factors such as ability, honesty etc.

Demerits of Median :

1. A slight change in the series may bring drastic change in median value.
2. In case of even number of items or continuous series, median is an estimated value other than any value in the series.
3. It is not suitable for further mathematical treatment except its use in mean deviation.
4. It is not taken into account all the observations.

Quartiles:

The quartiles divide the total number of observations into four equal parts. There are three quartiles. The second quartile divides the distribution into two halves and therefore is the same as the median. The first (lower) quartile (Q_1) covers the first one-fourth (25%), the third (upper) quartile (Q3) covers the three-fourth (75%) of the series. Thus, Q1 < Q2 < Q3.

Quartiles for individual series:

Let $x_1, x_2, \ldots, x_n$ be n observations arranged in ascending order of their magnitude then the quartiles, can be computed as

Qi = Value of $i\left(\frac{n+1}{4}\right)^{th}$ items, where i = 1, 2, and 3.

Example: Compute quartiles for the data given below 25,18,30, 8, 15, 5, 10, 35, 40, 45

Solution : arranging the given data in ascending order,

5, 8, 10, 15, 18,25, 30,35,40, 45

$$Q_1 = \left(\frac{n+1}{4}\right)^{\text{th}} \text{item}$$

$$= \left(\frac{10+1}{4}\right)^{\text{th}} \text{item}$$

$$= (2.75)^{\text{th}} \text{item}$$

$$= 2^{\text{nd}} \text{item} + \left(\frac{3}{4}\right) (3^{\text{rd}} \text{item} - 2^{\text{nd}} \text{item})$$

$$= 8 + \frac{3}{4} (10-8)$$

$$= 8 + \frac{3}{4} \times 2$$

$$= 8 + 1.5$$

$$= 9.5$$

$$Q_3 = 3\left(\frac{n+1}{4}\right)^{th} \text{item}$$

$$= 3 \times (2.75)^{\text{th}} \text{item}$$

$$= (8.25)^{\text{th}} \text{item}$$

$$= 8^{\text{th}} \text{item} + \frac{1}{4} [9^{\text{th}} \text{item} - 8^{\text{th}} \text{item}]$$

$$= 35 + \frac{1}{4} [40-35]$$

$$= 35+1.25=36.25$$

Discrete Series :

Step1: Find cumulative frequencies.

Step2: Find 1(N+1)/4

Step3: See in the cumulative frequencies , the value just greater than 1(N+1)/4, then the corresponding value of x is Q_1

Step4: Find 3(N+1)/4
Step5: See in the cumulative frequencies, the value just greater than 3(N+1)/4, then the corresponding value of x is Q_3

Example: Compute quartiles for the data given bellow.

X	5	8	12	15	19	24	30
f	4	3	2	4	5	2	4

Solution:

x	f	c.f ←
5	4	4
8	3	7
12	2	9
15	4	13
19	5	18
24	2	20
30	4	24
Total	24	

$$Q_1 = \left(\frac{N+1}{4}\right)^{th} \text{item} = \left(\frac{24+1}{4}\right) = \left(\frac{25}{4}\right) = 6.25^{\text{th}} \text{ item}$$

$$Q_3 = 3\left(\frac{N+1}{4}\right)^{th} \text{item} = 3\left(\frac{24+1}{4}\right) = 18.75^{\text{th}} \text{ item} \therefore Q_1 = 8;\ Q_3 = 24$$

Continuous series :
Step1: Find cumulative frequencies
Step2: Find N/4
Step3: See in the cumulative frequencies, the value just greater than N/4, then the corresponding class interval is called first quartile class.
Step4: Find 3(N/4), then see in the cumulative frequencies the value just greater than 3(N/4), then the corresponding class interval is called 3rd quartile class. Then apply the respective formulae

$$Q_1 = l_1 + \frac{\frac{N}{4} - m_1}{f_1} \times c_1$$

$$Q_3 = l_3 + \frac{3\left(\frac{N}{4}\right) - m_3}{f_3} \times c_3$$

Where l_1 = lower limit of the first quartile class
f_1 = frequency of the first quartile class
c_1 = width of the first quartile class
m_1 = c.f. preceding the first quartile class
l_3 = 1ower limit of the 3rd quartile class

f_3 = frequency of the 3rd quartile class
c_3 = width of the 3rd quartile class
m_3 = c.f. preceding the 3rd quartile class

Example: The following series relates to the marks secured by students in an examination.

Marks	No. of students
0-10	11
10-20	18
20-30	25
30-40	28
40-50	30
50-60	33
60-70	22
70-80	15
80-90	12
90-100	10

Find the lower and upper quartiles

Solution:

C.I.	f	cf
0-10	11	11
10-20	18	29
20-30	25	54
30-40	28	82
40-50	30	112
50-60	33	145
60-70	22	167
70-80	15	182
80-90	12	194
90-100	10	204
	204	

N/4 = 204/4 = 51; 3(N/4) = 3x51 = 153, So, lower quartile is

$$Q_1 = l_1 + \frac{\frac{N}{4} - m_1}{f_1} \times c_1$$

$$= 20 + \frac{51-29}{25} \times 10 \quad = 20 + 8.8 = 28.8$$

Upper quartile is

$$Q_3 = l_3 + \frac{3\left(\frac{N}{4}\right) - m_3}{f_3} \times c_3$$

$$= 60 + \frac{153-145}{22} \times 12 \quad = 60 + 4.36 = 64.36$$

Deciles:

The deciles divide the total number of observations into ten equal parts. There are nine deciles. These are D_1, D_2, ……., D_9. These are called first decile, second decile, …..ninth decile. Thus $D_1 < D_2$ ……$< D_9$.

Deciles for individual series:

Let $x_1, x_2, \ldots, x_n$ be n observations arranged in ascending order of their magnitude then the quartiles, can be computed as

Di = Value of $i\left(\frac{n+1}{10}\right)^{th}$ items, where i = 1, 2, ……, 9.

Example: Compute D_5 for the data 5, 24, 36, 12, 20, 8

Solution : Arranging the given values in the increasing order, we get 5, 8, 12, 20, 24, 36

D_5 = value of $5(n+1)/10^{th}$ item

= value of $5(6+1)/10^{th}$ item

= value of $(3.5)^{th}$ item

= 3^{rd} item + 0.5 × [4^{th} item – 3^{rd} item]

= 12 + 0.5 x [20 – 12] = 12+ 4 = 16

Deciles for Grouped data :

Example: Calculate D_3 and D_7 for the data given below

Class Interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency :	5	7	12	16	10	8	4

Solution:

C.I	f	c.f
0-10	5	5
10-20	7	12
20-30	12	24
30-40	16	40
40-50	10	50
50-60	8	58
60-70	4	62
	62	

D_3 item $= \left(\frac{3N}{10}\right)^{th}$ item

$= \left(\frac{3\times 62}{10}\right)^{th}$ item

$= (18.6)^{th}$ item

which lies in the interval 20-30

$\therefore D_3 = l + \frac{3\left(\frac{N}{10}\right) - m}{f} \times c$

$= 20 + \frac{18.6 - 12}{12} \times 10$

$= 20 + 5.5 = 25.5$

D_7 item $= \left(\frac{7 \times N}{10}\right)^{th}$ item

$= \left(\frac{7 \times 62}{10}\right)^{th}$ item

$= \left(\frac{434}{10}\right)^{th}$ item $=$ $(43.4)^{th}$ item

which lies in the interval(40-50)

$$D_7 = l + \frac{\left(\frac{7N}{10}\right) - m}{f} \times c$$

$$= 40 + \frac{43.4 - 40}{10} \times 10$$

$$= 40 + 3.4 = 43.4$$

Percentiles : The percentile values divide the distribution into 100 parts each containing 1 percent of the cases. The percentile (P_k) is that value of the variable up to which lie exactly k% of the total number of observations.

Relationship : $P_{25} = Q_1$; $P_{50} = D_5 = Q_2$ = Median and $P_{75} = Q_3$

Percentile for Raw Data or Ungrouped Data :

Example: Calculate P_{15} for the data 5, 24 , 36 , 12 , 20 , 8

Arranging the given values in the increasing order, we get 5, 8, 12, 20, 24, 36

P_{15} = Value of $15(n+1)/100^{th}$ observation= Value of $15(6+1)/100^{th}$ observation

= Value of $(1.05)^{th}$ observation = 1st item + 0.05 (2nd item – 1st item)

= 5 + 0.05 (8-5) = 5 + 0.15 = 5.15

Percentile for grouped data:

Example: Find P_{53} for the following frequency distribution.

Class interval	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency	5	8	12	16	20	10	4	3

Solution:

Class Interval	Frequency	C.f
0-5	5	5
5-10	8	13
10-15	12	25
15-20	16	41
20-25	20	61
25-30	10	71
30-35	4	75
35-40	3	78
Total	78	

$$P_{53} = l + \frac{\frac{53N}{100} - m}{f} \times c$$

$$= 20 + \frac{41.34 - 41}{20} \times 5$$

$$= 20 + 0.085 = 20.085.$$

Mode:

The mode is the value in the distribution which occurs most frequently. It is an actual value, which has the highest concentration of items in and around it. According to Croxton and Cowden "The mode of a distribution is the value at the point around which the items tend to be most heavily concentrated. It may be regarded at the most typical of a series of values".

Its importance is very great in marketing studies where a manager is interested in knowing about the size, which has the highest concentration of items. For example, in placing an order for shoes or ready-made garments the modal size helps because this sizes and other sizes around in common demand.

For ungrouped or Raw Data:

For ungrouped data or a series of individual observations, mode is often found by mere inspection.

Example*:* 3, 5, 11, 18, 11, 16, 9, 11, 3, 11

$\therefore$ Mode = $M_0 = 11$ since it occurs 4 times i.e. maximum times of occurrence.

In some cases the mode may be absent while in some cases there may be more than one mode.

Example: 1. 12, 10, 15, 24, 30 (no mode)

2. 7, 10, 15, 12, 7, 14, 24, 10, 7, 20, 10

$\therefore$ the modes are 7 and 10

Discrete distribution:

See the highest frequency and corresponding value of X is mode.

Continuous distribution:

See the highest frequency then the corresponding value of class interval is called the modal class. Then apply the formula.

$$\text{Mode} = \text{Mo} = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h,$$

Where,

l = Lower limit of the modal class

f_1 = frequency of the modal class

f_0 = frequency of the class preceding the modal class

f_2 = frequency of the class succeeding the modal class

h = size of the modal class

Example: The following table gives the distribution of 100 families according to their age:

Age(years):	0-10	10-20	20-30	30-40	40-50
No. of families:	14	a	27	b	15

Find the values of a and b if mode is 24.

Solution:

Age	Frequency
0-10	14
10-20	A

20-30	27
30-40	B
40-50	15
	N = 56 + a + b

Given that N = 100

$\therefore$ a + b = 44 ..... (i)

And b = 44 – a

Mo = 24 which lies in the class 20-30.

Here, l = 20, h = 10, $f_1 = 27$, $f_0 = a$, $f_2 = b$, $f_2 = 44 - a$

Now,

$$Mo = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

$$\text{Or, } 24 = 20 + \frac{(27 - a)}{54 - a - (44 - a)}$$

$$\text{Or, } 4 = \frac{(27 - a)}{10} \times 10$$

$\therefore$ a = 23 , b = 44 – 23 = 21

Determination of Modal class :

For a frequency distribution modal class corresponds to the maximum frequency. But in any one (or more) of the following cases

- If the maximum frequency is repeated
- If the maximum frequency occurs in the beginning or at the end of the distribution
- If there are irregularities in the distribution

The modal class is determined by the method of grouping.

Method of grouping

Steps for Calculation:

We prepare a grouping table with 6 columns.

1. In column I, we write down the given frequencies.
2. Column II is obtained by combining the frequencies two by two.
3. Leave the 1st frequency and combine the remaining frequencies two by two and write in column III.
4. Column IV is obtained by combining the frequencies three by three.
5. Leave the 1st frequency and combine the remaining frequencies three by three and write in column V.
6. Leave the 1st and 2nd frequencies and combine the remaining frequencies three by three and write in column VI.

Mark the highest frequency in each column. Then form an analysis table to find the modal class. After finding the modal class, use the formula to calculate the modal value.

Example: Calculate mode for the following frequency distribution.

Class interval	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency	9	12	15	16	17	15	10	13

Solution: Grouping Table

C I	f	2	3	4	5	6
0- 5	9					
5-10	12	21				
10-15	15		27	36		
15-20	16	31			**43**	
20-25	**17**		**33**			**48**
25-30	15	**32**		**48**		
30-35	10		25		42	38
35-40	13	23				

Analysis Table

Columns	**0-5**	**5-10**	**10-15**	**15-20**	**20-25**	**25-30**	**30-35**	**35-40**
1					1			
2					1	1		
3				1	1			
4				1	1	1		
5		1	1	1				
6			1	1	1			
Total		1	2	4	5	2		

The maximum occurred corresponding to 20-25, and hence it is the modal class.

$$\text{Mode} = \text{Mo} = l + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times C$$

Here $l = 20$; $\Delta_1 = f_1 - f_0 = 17 - 16 = 1$

$\Delta_2 = f_1 - f_2 = 17 - 15 = 2$

$$\therefore M_0 = 20 + \frac{1}{1+2} \times 5$$

$$= 20 + 1.67 = 21.67$$

Graphic Location of mode:

Steps:

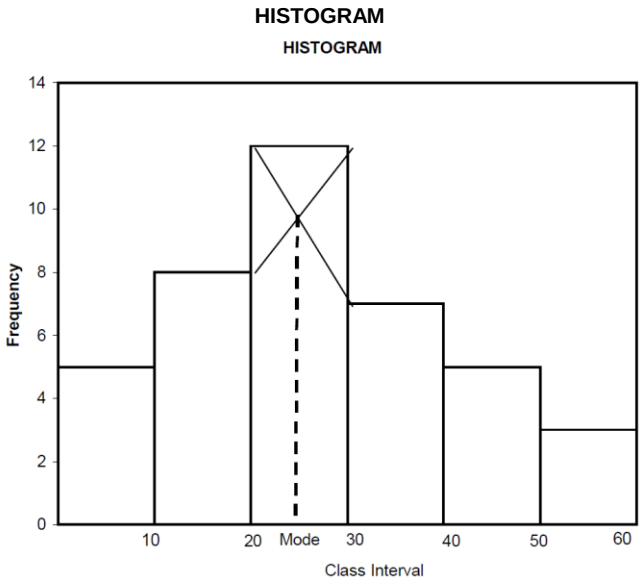
1. Draw a histogram of the given distribution.
2. Join the rectangle corner of the highest rectangle (modal class rectangle) by a straight line to the top right corner of the preceding rectangle. Similarly the top left corner of the highest rectangle is joined to the top left corner of the rectangle on the right.
3. From the point of intersection of these two diagonal lines, draw a perpendicular to the x -axis.
4. Read the value in x-axis gives the mode.

Example:

Locate the modal value graphically for the following frequency distribution.

Class interval	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	5	8	12	7	5	3

Solution:



Merits of Mode:

1. It is easy to calculate and in some cases it can be located mere inspection
2. Mode is not at all affected by extreme values.
3. It can be calculated for open-end classes.
4. It is usually an actual value of an important part of the series.
5. In some circumstances it is the best representative of data.

Demerits of mode:

1. It is not based on all observations.
2. It is not capable of further mathematical treatment.
3. Mode is ill-defined generally, it is not possible to find mode in some cases.
4. As compared with mean, mode is affected to a great extent, by sampling fluctuations.
5. It is unsuitable in cases where relative importance of items has to be considered.

EMPIRICAL RELATIONSHIP BETWEEN AVERAGES

In a symmetrical distribution the three simple averages mean = median = mode. For a moderately asymmetrical distribution, the relationship between them are brought by Prof. Karl Pearson as

mode = 3median - 2mean.

Example: If the mean and median of a moderately asymmetrical series are 26.8 and 27.9 respectively, what would be its most probable mode?

Solution: Using the empirical formula

Mode = 3 median – 2 mean = $3 \times 27.9 - 2 \times 26.8 = 30.1$

Example: In a moderately asymmetrical distribution the values of mode and mean are 32.1 and 35.4 respectively. Find the median value.

Solution: Median = [2mean+mode]/2 = $[2 \times 35.4 + 32.1]/2 = 34.3$

Example: Following are the daily wages of workers in a textile. Find the median. Ans: 468.75

Wages (in Rs.)	Number of workers
less than 100	5
less than 200	12
less than 300	20
less than 400	32
less than 500	40
less than 600	45
less than 700	52
less than 800	60
less than 900	68
less than 1000	75

$= 400 + 68.75 = 468.75$

Example: Find median for the data given below.(Ans: 43.75)

Marks	Number of students
Greater than 10	70
Greater than 20	62
Greater than 30	50
Greater than 40	38
Greater than 50	30
Greater than 60	24
Greater than 70	17
Greater than 80	9
Greater than 90	4

Example : Following is the distribution of persons according to different income groups. Calculate arithmetic mean. Ans=31

Income Rs(100)	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Number of persons	6	8	10	12	7	4	3

Example: Calculate mode for the following :

C- I	f
0-50	5
50-100	14
100-150	40
150-200	91
200-250	150
250-300	87
300-350	60
350-400	38
400 and above	15

MEASURES OF DISPERSION

Introduction :

The measure of central tendency serve to locate the center of the distribution, but they do not reveal how the items are spread out on either side of the center. This characteristic of a frequency distribution is commonly referred to as dispersion. In a series all the items are not equal. There is difference or variation among the values. The degree of variation is evaluated by various measures of dispersion. Small dispersion indicates high uniformity of the items, while large dispersion indicates less uniformity. For example consider the following marks of two students.

Student I	Student II
68	85
75	90
65	80
67	25
70	65

Both have got a total of 345 and an average of 69 each. The fact is that the second student has failed in one paper. When the averages alone are considered, the two students are equal. But first student has less variation than second student. Less variation is a desirable characteristic.

Characteristics of a good measure of dispersion:

An ideal measure of dispersion is expected to possess the following properties

1.It should be rigidly defined

2. It should be based on all the items.

3. It should not be unduly affected by extreme items.

4. It should lend itself for algebraic manipulation.

5. It should be simple to understand and easy to calculate

Absolute and Relative Measures :

There are two kinds of measures of dispersion, namely

1.Absolute measure of dispersion

2.Relative measure of dispersion.

Absolute measure of dispersion indicates the amount of variation in a set of values in terms of units of observations. For example, when rainfalls on different days are available in mm, any absolute measure of dispersion gives the variation in rainfall in mm. On the other hand relative measures of dispersion are free from the units of measurements of the observations. They are pure numbers. They are used to compare the variation in two or more sets, which are having different units of measurements of observations. The various absolute and relative measures of dispersion are listed below.

Absolute measure	**Relative measure**
1. Range	1.Co-efficient of Range
2.Quartile deviation	2.Co-efficient of Quartile deviation
3.Mean deviation	3. Co-efficient of Mean deviation
4.Standard deviation	4.Co-efficient of variation

Range and coefficient of Range:

Range:

This is the simplest possible measure of dispersion and is defined as the difference between the largest and smallest values of the variable.

In symbols, Range = L – S. Where L = Largest value, S = Smallest value.
In individual observations and discrete series, L and S are easily identified. In continuous series, the following two methods are followed.

Method 1:

L = Upper boundary of the highest class
S = Lower boundary of the lowest class.

Method 2:

L = Mid value of the highest class.
S = Mid value of the lowest class.

Co-efficient of Range :

$$\text{Co-efficient of Range} = \frac{L-S}{L+S}$$

Example: Find the value of range and its co-efficient for the following data.
7, 9, 6, 8, 11, 10, 4
Solution: L=11, S = 4.
Range = L – S = 11- 4 = 7

$$\text{Co-efficient of Range} = \frac{L-S}{L+S} = \frac{11-4}{11+4} = \frac{7}{15} = 0.4667$$

Example: Calculate range and its co efficient from the following distribution.

Size:	60-63	63-66	66-69	69-72	72-75
Number:	5	18	42	27	8

Solution: L = Upper boundary of the highest class = 75
S = Lower boundary of the lowest class = 60
Range = L – S = 75 – 60 = 15

$$\text{Co-efficient of Range} = \frac{L-S}{L+S} = \frac{75-60}{75+60} = \frac{15}{135} = 0.1111$$

Merits and Demerits of Range :

Merits:

1. It is simple to understand.
2. It is easy to calculate.
3. In certain types of problems like quality control, weather forecasts, share price analysis, etc., range is most widely used.

Demerits:
1. It is very much affected by the extreme items.
2. It is based on only two extreme observations.
3. It cannot be calculated from open-end class intervals.
4. It is not suitable for mathematical treatment.
5. It is a very rarely used measure.

Quartile Deviation and Co efficient of Quartile Deviation :

Quartile Deviation (Q.D) :

Definition: Quartile Deviation is half of the difference between the first and third quartiles. Hence, it is called Semi Inter Quartile Range. In Symbols,

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

Among the quartiles Q_1, Q_2 and Q_3, the range $Q_3 - Q_1$ is called inter quartile range and

$\frac{Q_3 - Q_1}{2}$, Semi inter quartile range.

Co-efficient of Quartile Deviation :

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

Example: Find the Quartile Deviation for the following data:
391, 384, 591, 407, 672, 522, 777, 733, 1490, 2488

Solution: Arrange the given values in ascending order. We get
384, 391, 407, 522, 591, 672, 733, 777, 1490, 2488.
Position of Q_1 is (n+1)/4 = (10+1/)4 = 2.75th item

Q_1 = 2nd value + 0.75 (3rd value – 2nd value)
= 391 + 0.75 (407 – 391)
= 391 + 0.75 × 16
= 391 + 12
= 403

Position Q_3 is 3(n+1)/4 = 3x2.75 = 8.25th item
Q_3 = 8th value + 0.25 (9th value – 8th value)
= 777 + 0.25 (1490 – 777)
= 777 + 0.25 (713)
= 777 + 178.25 = 955.25
Q.D = $(Q_3 - Q_1)/2$= 276.125

Example: Weekly wages of labours are given below. Calculate Q.D and Coefficient of Q.D.

Weekly Wage (Rs.) :	100	200	400	500	600
No. of Weeks :	5	8	21	12	6

Solution:
Position of Q_1 is (N+1)/4 = (52+1/)4 = 13.25th item
Q_1 = 13th value + 0.25 (14th Value – 13th value)
= 13th value + 0.25 (400 – 200)
= 200 + 0.25 (400 – 200)

= 200 + 0.25 (200)
= 200 + 50 = 250

Position Q_3 is $3(N+1)/4 = 3 \times 13.25 = 39.75^{th}$ item

$Q_3 = 39^{th}$ value + 0.75 (40^{th} value – 39^{th} value)

= 500 + 0.75 (500 – 500)

= 500 + 0.75 ×0

= 500

$Q.D = (Q_3 - Q_1)/2 = (500 – 250)/2 = 125$

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

$$= \frac{500 - 250}{500 + 250}$$

$$= \frac{250}{750} = 0.3333$$

Example: For the date given below, give the quartile deviation and coefficient of quartile deviation.

X :	351 – 500	501 – 650	651 – 800	801–950	951–1100
f :	48	189	88	4	28

Solution :

x	f	True class Intervals	Cumulative frequency
351- 500	48	350.5- 500.5	48
501- 650	189	500.5- 650.5	237
651- 800	88	650.5- 800.5	325
801- 950	47	800.5- 950.5	372
951- 1100	28	950.5- 1100.5	400
Total	N = 400		

$$Q_1 = l_1 + \frac{\frac{N}{4} - m_1}{f_1} \times c_1$$

$N/4 = 400/4 = 100$

Q_1 Class is 500.5 – 650.5, $l_1 = 500.5$, $m_1 = 48$, $f_1 = 189$, $c_1 = 150$

$$\therefore Q_1 = 500.5 + \frac{100 - 48}{189} \times 150$$

$$= 500.5 + \frac{52 \times 150}{189}$$

$$= 500.5 + 41.27$$

$$= 541.77$$

$$Q_3 = l_3 + \frac{3\frac{N}{4} - m_3}{f_3} \times c_3$$

3N/4 = 3 x 100 = 300

Q3 Class is 650.5 – 800.5, $l_3 = 650.5$, $m_3 = 237$, $f_3 = 88$, $C_3 = 150$

$$\therefore Q_3 = 650.5 + \frac{300 - 237}{88} \times 150$$

$$= 650.5 + \frac{63 \times 150}{88}$$

$$= 650.5 + 107.39$$

$$= 757.89$$

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

$$= \frac{757.89 - 541.77}{2}$$

$$= \frac{216.12}{2}$$

$$= 108.06$$

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

$$= \frac{757.89 - 541.77}{757.89 + 541.77}$$

$$= \frac{216.12}{1299.66} = 0.1663$$

Mid-range: It is the average of minimum value and maximum value

Mid-

Merits and Demerits of Quartile Deviation

Merits :

1. It is Simple to understand and easy to calculate
2. It is not affected by extreme values.
3. It can be calculated for data with open end classes also.

Demerits:

1. It is not based on all the items. It is based on two positional values Q_1 and Q_3 and ignores the extreme 50% of the items
2. It is not amenable to further mathematical treatment.
3. It is affected by sampling fluctuations.

Standard Deviation and Coefficient of variation:

Standard Deviation :

Karl Pearson introduced the concept of standard deviation in 1893. It is the most important measure of dispersion and is widely used in many statistical formulae. Standard deviation is also called Root-Mean Square Deviation. The reason is that it is the square–root of the mean of the squared deviation from the arithmetic mean. It provides accurate result. Square of standard deviation is called Variance.

Definition:

It is defined as the positive square-root of the arithmetic mean of the Square of the deviations of the given observation from their arithmetic mean. The standard deviation is denoted by the Greek letter σ (sigma)

Calculation of Standard deviation-Individual Series :

There are two methods of calculating Standard deviation in an individual series.

a) Deviations taken from Actual mean

b) Deviation taken from Assumed mean

Example: Calculate the standard deviation from the data 14, 22, 9, 15, 20, 17, 12, 11

Solution: Deviations from actual mean.

Values (X)	$X - \overline{X}$	$(X - \overline{X})^2$
14	-1	1
22	7	49
9	-6	36
15	0	0
20	5	25
17	2	4
12	-3	9
11	-4	16
120		140

$$\overline{X} = \frac{120}{8} = 15$$

$$\sigma = \sqrt{\frac{\Sigma(x - \overline{x})^2}{n}}$$

$$= \sqrt{\frac{140}{8}}$$

$$= \sqrt{17.5} = 4.18$$

Example: The table below gives the marks obtained by 10 students in statistics. Calculate standard deviation.

Student Nos :	1	2	3	4	5	6	7	8	9	10
Marks :	43	48	65	57	31	60	37	48	78	59

Solution: Standard deviation from assumed mean

Nos.	Marks (x)	d=X-A (A=57)	d^2
1	43	-14	196
2	48	-9	81
3	65	8	64
4	57	0	0
5	31	-26	676
6	60	3	9
7	37	-20	400
8	48	-9	81
9	78	21	441
10	59	2	4
n = 10		$\sum d=-44$	$\sum d^2=1952$

$$\sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$$

$$= \sqrt{\frac{1952}{10} - \left(\frac{-44}{10}\right)^2}$$

$$= \sqrt{195.2 - 19.36}$$

$$= \sqrt{175.84} = 13.26$$

Example: calculate the standard deviation from the following data: 4, 6, 8, 14, 18

Solution:

X	x^2
4	16
6	36
8	64
14	6
18	324
$\sum x = 50$	$\Sigma x^2 = 50$

We know standard deviation(σ) = $\sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$ =5.21

For discrete series,

Standard deviation(σ) = $\sqrt{\frac{\Sigma f x^2}{N} - \left(\frac{\Sigma f x}{N}\right)^2}$

For Continuous series,

Standard deviation(σ) = $\sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2}$, where x is mid value of class interval.

Variance: Square of the standard deviation is Variance. It is denoted by σ^2

Combined Standard Deviation:

If a series of N_1 items has mean $\overline{X}_1$ and standard deviation σ_1 and another series of N_2 items has mean $\overline{X}_2$ and standard deviation σ_2, we can find out the combined mean and combined standard deviation by using the formula.

$$\overline{X}_{12} = \frac{N1\,\overline{X}_1 + N_2\overline{X}_2}{N_1 + N_2}$$

and

$$\sigma_{12} = \sqrt{\frac{N_1\,\sigma_1^2 + N_2\,\sigma_2^2 + N_1 d_1^2 + N_2 d_2^2}{N_1 + N_2}}$$

Where $d_1 = \overline{X}_1 - \overline{X}_{12}$

$d_2 = \overline{X}_2 - \overline{X}_{12}$

Merits and Demerits of Standard Deviation:

Merits:

1. It is rigidly defined and its value is always definite and based on all the observations and the actual signs of deviations are used.
2. As it is based on arithmetic mean, it has all the merits of arithmetic mean.
3. It is the most important and widely used measure of dispersion.
4. It is possible for further algebraic treatment.
5. It is less affected by the fluctuations of sampling and hence stable.
6. It is the basis for measuring the coefficient of correlation and sampling.

Demerits:

1. It is not easy to understand and it is difficult to calculate.
2. It gives more weight to extreme values because the values are squared up.
3. As it is an absolute measure of variability, it cannot be used for the purpose of comparison.

Coefficient of Variation :

The Standard deviation is an absolute measure of dispersion. It is expressed in terms of units in which the original figures are collected and stated. The standard deviation of heights of students cannot be compared with the standard deviation of weights of students, as both are expressed in different units, i.e heights in centimeter and weights in kilograms. Therefore the standard deviation must be converted into a relative measure of dispersion for the purpose of comparison. The relative measure is known as the coefficient of variation. The coefficient of variation is obtained by dividing the standard deviation by the mean and multiply it by 100. It is always expressed as percentage.

symbolically,

Coefficient of Variation (C.V.) = $\frac{\sigma}{\bar{x}} \times 100\%$

If we want to compare the variability of two or more series, we can use C.V. The series or groups of data for which the C.V. is greater indicate that the group is more variable, less stable, less uniform, less consistent, less equitable or less homogeneous. If the C.V. is less, it indicates that the group is less variable, more stable, more uniform, more consistent, more equitable or more homogeneous.

Example : In two factories A and B located in the same industrial area, the average weekly wages (in rupees) and the standard deviations are as follows:

Factory	Average	Standard Deviation	No. of workers
A	34.5	5	476
B	28.5	4.5	524

Given $N_1 = 476$, $\overline{X}_1 = 34.5$, $\sigma_1 = 5$; $N_2 = 524$, $\overline{X}_2 = 28.5$, $\sigma_2 = 4.5$

1. Total wages paid by factory A = 34.5 × 476 = Rs.16.422

Total wages paid by factory B = 28.5 × 524 = Rs.14,934.

Therefore factory A pays out larger amount as weekly wages.

2. C.V. of distribution of weekly wages of factory A and B are

Coefficient of Variation (C.V.) for A = $\frac{\sigma}{\bar{x}} \times 100\% = \frac{5}{34.5} \times 100\% = 14.49\%$

Coefficient of Variation (C.V.) for B = $\frac{\sigma}{\bar{x}} \times 100\% = \frac{4.5}{28.5} \times 100\% = 15.79\%$

Factory B has greater variability in individual wages, since C.V. of factory B is greater than C.V of factory A.

Very short questions

1. Describe the type I and II error. (Describe error in testing of hypothesis)

Type I error:

The rejection of null hypothesis, when it is true, is called type I error. The probability of a type I error is denoted by α. This α is also known as size of the critical region.

α=P (type I error)

=P (reject H_0 when H_0 is true)

Type II error

If false null hypothesis H_0 is accepted, it is said to be type II error. The probability of this type of error is denoted by β.

β=P (type II error)

= P (accept H_0 when H_0 is false)

2. Explain qualities of a good estimator.

A good estimator is one which is as close to the true value of the parameter as possible. A good estimator has following four properties.

- **Unbiasedness:** If the expected value of sample statistic is equal to parametric value, then the estimator is said to be unbiased, otherwise biased.
  i.e., E(t) =θ
- **Consistency:** if the sample size increases, the value of sample statistic (t_n) becomes very nearer to the value of population parameter θ then it is called consistent estimator.
  i.e., $t_n \rightarrow \theta$ as $n \rightarrow$□
- **Sufficiency:** an estimator containing all the information contained in the sample regarding the parameter θ is a sufficient estimator.
- **Efficiency:** let t_1 and t_2 be two consistent estimators of parameter θ. Then the estimator t_1 is said to be more efficient estimator than t_2 if variance of t_1 is less than variance of t_2.
  i.e., var (t_1) < var (t_2).

3. What are the advantages of stem and leaf display?

- Concise representation of data
- Shows range, minimum and maximum, gaps and cluster, and outlier easily.
- Can handle extremely large data sets.

4. Write down the area properties of normal distribution.

The area properties of normal distribution are as follows:

- The area under the normal probability curve between the ordinates at X= μ ± σ is 0.6826.
  i.e., P (μ-σ < X < μ + σ) = 0.6826. This is same as, the interval μ ± σ covers 68.26% of observations.
- The area under the normal probability curve between the ordinates at X= μ ± 2σ is 0.9544.

i.e., P ($\mu-2\sigma < X < \mu + 2\sigma$) = 0.9544. This is same as, the interval $\mu \pm 2\sigma$ covers 95.44% of observations.

- The area under the normal probability curve between the ordinates at $X= \mu \pm 3\ \sigma$ is 0.9974. i.e., P ($\mu-3\sigma < X < \mu + 3\sigma$) = 0.9974. This is same as, the interval $\mu \pm 3\sigma$ covers 99.74% of observations.

5. State the necessary conditions to apply binomial distribution.

Binomial distribution is widely used probability distribution of discrete random variable. The binomial distribution holds under the following conditions:

- A Binomial experiment consists of n trials repeated under the same conditions.
- Each trail has only two outcomes – Success and Failure.
- The repeated trials are independent.
- The probability of success remains constant for each trial.

6. State the necessary conditions that Poisson distribution is obtained from binomial distribution.

Poisson distribution is a limiting case of binomial distribution under the following conditions:

- The number of trials n is indefinitely large i.e., $n \rightarrow \square$.
- The probability of success for each trial is very small i.e., $p \rightarrow 0$.
- The mean np is finite constant i.e., $np=\lambda$ (finite)
- Events are independent.

Under these four conditions, the binomial probability function tends to probability function of the Poisson distribution given by

$P(X=x) = p(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$, x= 0, 1, 2, …

7. Define descriptive and inferential statistics.

Descriptive Statistics

In the descriptive Statistics, the Data is described in a summarized way. The summarization is done from the sample of the population using different parameters like Mean or standard deviation. Descriptive Statistics are a way of using charts, graphs, and summary measures to organize, represent, and explain a set of Data.

- Data is typically arranged and displayed in tables or graphs summarizing details such as histograms, pie charts, bars, or scatter plots.

Inferential Statistics

In the Inferential Statistics, we try to interpret the Meaning of descriptive Statistics. After the Data has been collected, analyzed, and summarized we use Inferential Statistics to describe the Meaning of the collected Data.

- Inferential Statistics are intended to test hypotheses and investigate relationships between variables and can be used to make population predictions.
- Inferential Statistics are used to draw conclusions and inferences, i.e., to make valid generalizations from samples.

8.Describe the difference between descriptive and inferential statistics.

Descriptive Statistics	Inferential Statistics
Concerned with the describing the target population.	Make inferences from the sample and generalize them to the population.
Organize, analyze, and present the data in a meaningful manner.	Compares, test and predicts future outcomes.
Final results are shown in form of charts, tables and graphs.	Final result is the probability scores.
Describes the data which is already known.	Tries to make conclusion about the population that is beyond the data available.
Tools- measures of central tendency and measure of dispersion.	Tools- hypothesis tests, analysis of variance etc.

9. Write down the methods of collecting primary data.

The various methods with which the primary data can be collected are as follows:

- **Direct personal interview:** In this method, the investigator personally interviews the respondent either directly or through phone or through any electronic media. This method is suitable when the scope of investigation is small and greater accuracy is needed.
- **Indirect oral interview:** The indirect method is used in cases where it is delicate or difficult to get the information from the respondents due to unwillingness or indifference. The information about the respondent is collected by interviewing the third party who knows the respondent well.
- **Local correspondents' method:** In this method, the investigator appoints local agents or correspondents in different places. They collect the information on behalf of the investigator in their locality and transmit the data to the investigator or headquarters.
- **Mailed questionnaire method:** In this method, a questionnaire is prepared. These questionnaires are addressed to individual informants and sent by post. They are requested to answer the questions and post back to the investigator.
- **Schedule sent through enumerators:** In this method, the trained enumerators or interviewers take the schedules themselves, contact the informants, get replies, and fill them in their own handwriting. This method is suitable when the respondents include illiterates.

10. Define secondary data and discuss different sources of secondary data.

Secondary data is collected and processed by some other agency, but the investigator uses it for his study. They can be obtained from published sources such as government reports, documents, newspapers, books written by economists or from any other source. Secondary data can be classified under the following two headings:

1) Published sources:

- Reports and publications of ministers, departments of the government.
- Reports and publications of reputed INGO's such as UBDP, ADB, WHO, IMF etc.
- Reports and publications of reliable NGO's, journals etc.

2) Unpublished sources

- Records maintained by government offices.
- Records maintained by research institutions, research scholars etc.
- Records updated by the department institutions for their internal purpose.

11. What are the limitation of statistics?

The limitation of statistics are as follows:

- Statistics does not deal with individuals.
- Statistics does not study qualitative phenomena.
- Statistical laws are not exact.
- Statistics can be misused.

12. Discuss the important of statistics on business and industry.

Statistics focuses on the study and manipulation of data, as well as gathering, documenting, reviewing, analyzing, and drawing conclusions from it. Statistical methods are widely used in business and trade solutions such as financial analysis, market research and manpower planning. Every business establishment irrespective of the type must adopt statistical techniques for its growth. They estimate the trend of prices, buying and selling, importing, and exporting of goods using statistical methods and past data. In any business enterprise, statistical methods can be used for three major purposes. Among them are:

- **Operational planning:** This may be done for special projects or for the recurring activities of a firm over a given period of time.

- **Setting standards:** For instance, setting out standards for the size of employment, sales volume, product quality specifications, and production output.

- **Control:** This involves comparing actual production with a norm or target established earlier. When production falls short of the target, it provides remedial measures so that such a lapse does not happen again.

13. What do you mean by data? What are the differences between primary and secondary data?

Data is a collection of discrete or continuous values that convey information, describing the quantity, quality, fact, statistics, other basic units of meaning, or simply sequences of symbols that may be further interpreted.

Difference between Primary data and Secondary data

Parameters of Comparison	**Primary Data**	**Secondary Data**
Definition	It is the crude form of all the data.	It is a refined form of data.
Source	It can be collected using various methods like interviews, experiments, etc.	It can be obtained from the internet, journals, etc.
Authenticity	It is very authentic in relation to the topic concerned.	It may be biased. It depends on the biases of the researcher.
Cost of collection	It is very costly to collect such data.	It costs very little or nothing.
Purpose	The primary purpose of the data is to add new knowledge.	It is a manipulated form of data and just tells the same story from a different perspective.

14. What are the differences between population and sample?

Basic for comparison	Population	Sample
Meaning	Population refers to the collection of all elements possessing common characteristics, that comprises universe.	Sample means a subgroup of the members of population chosen for participation in the study
Includes	Each and every unit of the group	Only a handful units of population
Characteristics	Parameter	Statistic
Data collection	Complete enumeration or census	Sample survey or sampling
Focus on	Identifying the characteristics	Making inferences about population.

15. Define mutually exclusive and independent events.

Mutually exclusive events: Two or more events are said to be mutually exclusive, when the occurrence of any one event excludes the occurrence of other event. Mutually exclusive events cannot occur simultaneously. E.g., Thus, if a coin is tossed, either the head can be up, or tail can be up; but both cannot be up at the same time.

Independent event: A set of events is said to be independent, if the occurrence of any one of them does not, in any way, affect the Occurrence of any other in the set. For example, when we toss a coin twice, the result of the second toss will in no way be affected by the result of the first toss.

16. Define the term sample space and exhaustive number of events.

Sample space: The set or aggregate of all possible outcomes is known as sample space. For example, when we roll a die, the possible outcomes are 1, 2, 3, 4, 5, and 6; one and only one face come upwards. Thus, all the outcomes— 1, 2, 3, 4, 5 and 6 are sample space. And each possible outcome or element in a sample space called sample point.

Exhaustive number of events: The total number of possible outcomes of a random experiment is called exhaustive events. The group of events is exhaustive, as there is no other possible outcome. Thus, tossing a coin, the possible outcome is head or tail; exhaustive events are two. Similarly throwing a die, the outcomes are 1, 2, 3, 4, 5 and 6. In case of two coins, the possible number of outcomes are 4 i.e. (2^2), i.e., HH, HT TH and TT.

17. What is confidence interval? What do you understand by 95% confidence level?

The **confidence interval** is the range of values that you expect your estimate to fall between a certain percentage of the time if you run your experiment again or re-sample the population in the same way. It is given by,

C.I. = $\bar{X} \pm Z_\alpha$ * S.E. ($\bar{X}$)

95% confidence level means, you are confident that 95 out of 100 times the estimate will fall between the upper and lower values specified by the confidence interval.

18. Describe the importance of sampling.

- It helps to collect vital information more quickly and it helps to make estimates of the characteristics of the total population in a shorter time.
- Sampling can save time and money.
- Sampling techniques often increases the accuracy of the data. With small samples it become easier to check the accuracy of the data.
- If the population is too large, or hypothetical sampling is the only method to be used.

19. What are the difference between Null hypothesis and Alternative hypothesis?

Difference between Null and Alternative hypothesis

Null hypothesis	**Alternative hypothesis**
A null hypothesis represents the hypothesis that there is **"no relationship" or "no association" or "no difference" between two variables.**	An alternative hypothesis is the opposite of the null hypothesis where **we can find some statistical importance or relationship between two variables.**
In case of null hypothesis, researcher tries to invalidate or reject the hypothesis.	In an alternative hypothesis, the researcher wants to show or prove some relationship between variables.
It is an assumption that specifies a possible truth to an event where there is **absence of an effect.**	It is an assumption that describes an alternative truth where there is **some effect** or some difference.
Null hypothesis is a statement that **signifies no change**, no effect, and no differences between variables.	Alternative hypothesis is a statement that **signifies some change, some effect, and some differences** between variables.
If null hypothesis is true, any discrepancy between observed data and the hypothesis is only due to chance.	If alternative hypothesis is true, the observed discrepancy between the observed data and the null hypothesis is not due to chance.
A null hypothesis is denoted as H_0.	An alternative hypothesis is denoted as H_1 or H_A.
Example of null hypothesis: There is no association between use of oral contraceptive and blood cancer. $H_0: \mu = 0$	**Example of an alternative hypothesis:** There is no association between use of oral contraceptive and blood cancer. $H_A: \mu \neq 0$

Measures of Central Tendency

1.Following are the marks obtained by 10 students in Statistics.

75, 79, 80, 81, 84, 85, 88, 90, 92, 95

Calculate mean marks of these 10 students.

2.Arithmetic mean of 98 items is 50. Two items 60 and 70 were left out at the time of calculations. What is the correct mean of all the items?

3.Calculate arithmetic mean for the following frequency distribution.

X	5	10	15	20	25
f	2	4	7	3	1

4.From the following data find the missing items, if the mean of the distribution is 115.86

Item	110	112	113	117	-	125	128	130
No. of hours	25	17	13	15	14	8	6	2

5.Calculate arithmetic mean for the following data.

Marks	0-10	10-20	20-30	30-40	40-50
No. of students	4	6	10	20	10

6.From the following data compute the mean marks of all the students of 50 schools in a city.

Marks obtained	10-15	15-20	20-25	25-30	30-35	35-40
No. of schools	4	5	9	15	10	7
Average no. of students in a school	100	150	200	300	250	200

7.The following data given the weekly wages of the worker in a firm. Calculate the average weekly wage per worker

Wages group (Rs.)	80-100	100-120	120-140	140-160	160-180	180-200
Total hours	168	170	225	272	126	91
Average no. of hours worked per worker	12	10	9	8.5	7	6.5

8.Find the median of the following set of observations.

i. 60, 70, 50, 80, 90, 100, 110
ii. 70, 80, 60, 90, 120, 140

9. Find the median of the pop quiz marks shown below:

Marks	0	1	2	3	4	5
Frequency	1	1	5	3	2	1

10. The following is the income distribution of the persons:

Income (00 Rs.)	50-80	80-100	100-110	110-120	120-130	130-150	150-180	180-200
No. of persons	30	127	140	240	176	135	20	3

Find the median incomes.

11. Find the median wage of a labor from the following table

Wages (Rs.)	Above 0	Above 10	Above 20	Above 30	Above 40	Above 50	Above 60	Above 70
No. of labors	650	500	425	375	300	275	250	100

(Ans: M_d =Rs 36.67)

12. The expenditure of 1000 families is given as below:

Expenditure (Rs.)	40-59	60-79	80-99	100-119	120-139
No. of families	50	-	500	-	50

The median for the distribution is Rs 87. Calculate the missing frequencies.

(Ans: 263, 137)

13. Find lower and upper quartiles from the given data.

20, 18, 15, 16, 19, 25, 12, 14, 22

14. Find first and third quartiles from the given data.

X	1	2	3	4	5	6	7
f	2	5	7	10	4	3	2

15. Draw less than ogive from the following data and hence locate median, first quartile, seventh decile, and 30^{th} percentile.

0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
2	3	6	10	12	15	10	6	4

16. From the following table showing the marks distribution in a certain class, determine:

a) Limits of the marks for the middle 50% of students.
b) The lowest marks obtained by top 40% students.
c) The minimum pass marks if 40% students failed the exam.

Marks	0-20	20-40	40-60	60-80	80-100
Students	12	18	36	24	10

17. Compute mode from the following distribution.

X	50	100	150	200	250	300	350	400
f	5	14	40	91	150	87	60	38

18. Calculate the modal size in the following distribution:

Size (inches)	Below 10	10-12	12-14	14-16	16-18	18-20
Demand	3	15	27	20	3	2

Will the median fall under the same size?

The Five-Number summary

A five-number summary consists of minimum observation, lower quartile, median, upper quartile, maximum observation and provides a way of determine the shape of the distribution, that is, to see if there is symmetry or not in data.

If the data are perfectly symmetrical, the relationship among the various measures in the five-number summary will be as expressed below:

- The distance from the smallest value, i.e. $X_{smallest}$ to median is equal to the distance from median to the largest value, i.e., $X_{largest}$.
- The distance from the smallest value i.e. $X_{smallest}$ to first quartile (Q_1) is equal to the distance from upper quartile (Q_3) to the largest value, i.e., $X_{largest}$.

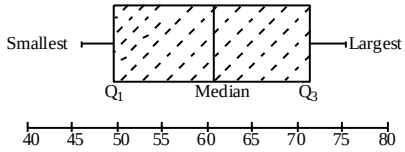
For asymmetrical distributions, the relationship among the various measures in the five-number summary will be as expressed below:

- In right-skewed distributions, the distance from median to the smallest value, i.e. $X_{smallest}$ is smaller than the distance from the largest value, i.e., $X_{largest}$ to median.
- In right-skewed distributions, the distance from upper quartile (Q_3) to the largest value, i.e., $X_{largest}$ is greater than the distance from the smallest value i.e. $X_{smallest}$ to first quartile (Q_1).
- In left-skewed distributions, the distance from the smallest value, i.e. $X_{smallest}$ to median is greater than the distance from the largest value, i.e., $X_{largest}$ to median.
- In right-skewed distributions, the distance from upper quartile (Q_3) to the largest value, i.e., $X_{largest}$ is greater than the distance from the smallest value i.e. $X_{smallest}$ to first quartile (Q_1).

Box-and-whisker plot

A box and whisker plot is a graphical presentation of the data that displays a five number summary of a data set based on the minimum, lower quartile, median, upper quartile, and maximum. The vertical line drawn within the box represents the median. The vertical line at the left side of the box represents the location of Q_1 and the vertical line in the right side of the box represents the location of Q_3. Hence the box contains the middle 50% of the values in the distribution of the given data set. The lower 25% of the data are represented by the line (known as whisker) connecting the left side of the box to the location of the smallest value. Similarly, the

upper 25% of the data are represented by the line (known as whisker) connecting the right side of the box to the location of the largest value as shown in the following box and whisker plot.



Example: The data represent the amount of grams of carbohydrates in a serving of breakfast cereal 1, 15, 23, 29, 19, 22, 21, 20, 15, 25, 17. Construct a box and whisker plot for the carbohydrate amounts.

Solution: For the calculation of Q_1, median, and Q_3 arrange the data in ascending order to their magnitude as follows

11 15 15 17 19 20 21 22 23 25 29

Five number summary (Smallest, Q_1, median, Q_3 and the largest) values can be represented through box - and - whisker plot.

The smallest value is 11 and the largest value is 29.

For first quartile (Q_1),

$$Q_1 = \text{value of } \left(\frac{n+1}{4}\right)^{th} \text{ item} = \text{value of } \left(\frac{11+1}{4}\right) \text{item} = \text{value of } 3^{rd} \text{ item}$$

$\therefore$ $Q_1 = 15$, Where the number of observations (n) = 11

For Median,

$$\text{Median } (M_d) = \text{Value of } \left(\frac{n+1}{2}\right)^{th} \text{ item} = \text{Value of } \left(\frac{n+1}{4}\right)^{th} \text{ item} = \text{Value of } 6^{th} \text{ item}$$

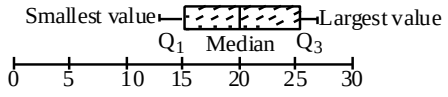
$\therefore$ $M_d = 20$

Similarly, for 3^{rd} quartile (Q_3)

$$Q_3 = \text{value of } \left[\frac{3(n+1)}{4}\right]^{th} \text{ item} = \text{value of } \left[\frac{3(11+1)}{4}\right]^{th} \text{ item} = \text{value of } 9^{th} \text{ item}$$

$\therefore$ $Q_3 = 23$

The five number summary is (11, 15, 20, 23, 29). Now, the box - and - whisker plot for the given data of amount of carbohydrates in a serving of breakfast is as follows:



Example: In Pokhara, saving banks are permitted to sell a form life insurance called Saving Bank Life insurance. The approval process consists of under writing, which includes a review of the application, a medical information bureau check, possible requests for additional medical information and medical exams a policy complication stage where the policy pages are generated and sent to the bank for delivery. The ability to deliver approved policies to customers in a timely manner is critical to the profitability of this service to the bank. During a period of 1 month, a random sample of 27 approved policies was selected and the total processing time in days was recorded with the following results.

73	19	16	64	28	28	31	90	60	56
22	18	45	48	17	17	7	91	63	50
51	51	31	56	69	16	17			

i) Construct a box and whisker plot.

ii) Are the data skewed? If So, how?

Solution: (i) Let arrange the data into ascending order of their magnitude as follows:

16, 16, 17, 17, 18, 19, 22, 28, 28, 31, 31, 45, 48, 50, 51, 56, 60, 63, 64, 69, 74, 90, 91, 92

In order to construct box - and - whisker plot, let us first compute the five number summary (i.e. smallest, Q_1, median, Q_3, largest) as follows:

Here, The smallest value (S) = 16

The largest value (L) = 92

For quartile, Q_1,

$$Q_1 = \text{Value of } \left(\frac{n+1}{4}\right)^{th} \text{ item} = \text{Value of } \left(\frac{27+1}{4}\right)^{th} = \text{Value of 7th item}$$

$\therefore$ $Q_1 = 18$

Median, Md,

$$Md = \text{Value of } \left(\frac{n+1}{2}\right)^{th} \text{ item} = \text{Value of } 14^{th} \text{ item}$$

$Md = 45$

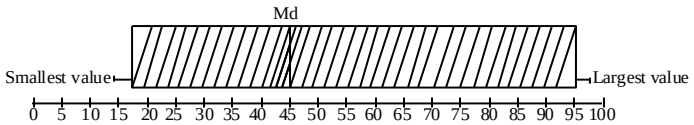
$$Q_3 = \text{Value of } \left[\frac{3(n+1)}{4}\right]^{th} \text{ item} = \text{Value of } \left[\frac{3(27+1)}{4}\right]^{th} \text{ item}$$

$= \text{Value of } 21^{th} \text{ item} = 63$

$\therefore$ $Q_3 = 63$

Hence, the five number summary is given by (16, 18, 45, 63, 92)

This five number summary can be presented in the box – and – whiser put as follows.



(ii) The distance from the smallest value to the median = 45 – 16 = 29

The distance from median to largest value = 92 – 45 = 47

The distance from the smallest value to Q_1 = 18 – 16 = 29

The distance from median to largest value = 92 – 45 = 47

The distance from the smallest value to Q_1 = 18 – 16 = 2

The distance from Q_3 to the largest value = 92 – 63 = 29

Here, the distance from the smallest value to the median is less than distance from the median to the largest value (i.e 29 < 49), Similarly, the distance from the smallest value to Q_1 is also less than the distance from Q_3 to the largest value (i.e 2 < 29). This shows that the data is right skewed.

However if we compute $(M_d - Q_1) > (Q_3 - M_d)$ i.e 27 > 18 which is indicative of left skewed distribution. This contradictory information clearly indicates the given data set is not uniformly distributed.

Probability

Introduction

The concept of probability is difficult to define in precise terms. In ordinary language, the word probable means likely or chance. The probability theory is an important branch of mathematics. Generally, the word, probability, is used to denote the happening of a certain event, and the likelihood of the Occurrence of that event, based on past experiences. By looking at the clear sky, one will say that there will not be any rain today. On the other hand, by looking at the cloudy sky or overcast sky, one will say that there will be rain today. In the earlier sentence, we aim that there will not be rain, and, in the latter, we expect rain. On the other hand, a mathematician says that the probability of rain is 0 in the first case and that the probability of rain is 1 in the second case. In between 0 and 1, there are fractions denoting the chance of the event occurring.

If a coin is tossed, the coin falls dawn. The coin has two sides: head and tail. On tossing a coin, the coin may fall down either with the head up or tail up. A coin, on reaching the ground, will not stand on its edge or rather, we assume; so, the probability of the coin coming down is 1. The probability of the head coming up is 50% and the tail coming up is 50%; in other words, we can say the probability of the head or the tail coming up is 1/2, 1/ 2 Ince 'head' and 'tail' share equal chances. The probability that it will come down head or tail is unity.

Some useful terms

Before discussing the theory of probability, let us have an understanding of the following terms:

Random Experiment:

If an experiment or trial can be repeated under the same conditions, any number of times and it is possible to count the total number of outcomes, but individual result i.e., individual outcome is not predictable. Suppose we toss a coin. It is not possible to predict exactly the outcomes. The outcome may be either head up or tail up. Thus, an action or an operation which can produce any result or outcome is called a random experiment.

Event and Trials:

Any possible outcome of a random experiment is called an event. Performing an experiment is called trial and outcomes are termed as events. An event whose occurrence is inevitable when a certain random experiment is performed, is called a sure event or certain event. At the same time, an event which can never occur when a certain random experiment is performed is called an impossible event. The events may be simple or composite. An event is called simple if it corresponds to a single possible outcome. For example, in rolling a die, the chance of getting 2 is a simple event. Further in tossing a die, chance of getting event numbers (1, 3, 5) are compound event.

Sample space:

The set or aggregate of all possible outcomes is known as sample space. For example, when we roll a die, the possible outcomes are 1, 2, 3, 4, 5, and 6; one and only one face come upwards. Thus, all the outcomes— 1, 2, 3, 4, 5 and 6 are sample space. And each possible outcome or element in a sample space called sample point.

Mutually exclusive events or cases:

Two events are said to be mutually exclusive if the occurrence of one of them excludes the possibility of the occurrence of the other in a single observation. The occurrence of one event prevents the occurrence of the other event. As such, mutually exclusive events are those events, the occurrence of which prevents the possibility of the other to occur. All simple events are mutually exclusive. Thus, if a coin is tossed, either the head can be up, or tail can be up; but both cannot be up at the same time.

Similarly, in one throw of a die, an even and odd number cannot come up at the same time. Thus, two or more events are considered mutually exclusive if the events cannot occur together.

Equally likely events:

The outcomes are said to be equally likely when one does not occur more often than the others. That is, two or more events are said to be equally likely if the chance of their happening is equal. Thus, in a throw of a die the coming up of 1, 2, 3, 4, 5 and 6 is equally likely. For example, head and tail are equally likely events in tossing an unbiased coin.

Exhaustive events

The total number of possible outcomes of a random experiment is called exhaustive events. The group of events is exhaustive, as there is no other possible outcome. Thus, tossing a coin, the possible outcomes are head or tail; exhaustive events are two. Similarly throwing a die, the outcomes are 1, 2, 3, 4, 5 and 6. In case of two coins, the possible number of outcomes are 4 i.e. (2^2), i.e., HH, HT TH and TT. In case of 3 coins, the possible outcomes are 23=8 and so on. Thus, in a throw of n" coin, the exhaustive number of cases is 2^n.

Independent events:

Events are said to be independent if the happening of one event does not affect the happening of the other event. In tossing of a coin, the occurrence of head in first tossing is independent of the occurrence of head in the second tossing.

Dependent events:

Events are said to be dependent if the occurrence of one event affects the probability of the occurrence of the other event.

For example, the probability of drawing a king from a pack of 52 cards is 4/52; the card is not put back; then the probability of drawing a king again is 3/51. Thus, the outcome of the first event

affects the outcome of the second event, and they are dependent. But if the card is put back, then the probability of drawing a king is 4/52 and is an independent event.

Favorable Cases

The number of outcomes which result in the happening of a desired event are called favorable cases to the event. For example, in drawing a card from a pack of cards, the cases favorable to "getting a diamond" are 13 and to "getting an ace of spade" is only one. Take another example, in a single throw of a dice the number of favorable cases of getting an odd number are three -1,3 &5.

MEASUREMENT OF PROBABILITY

The origin and development of the theory of probability dates back to the seventeenth century. Ordinarily speaking the probability of an event denotes the likelihood of its happening. A value of the probability is a number ranges between 0 and 1. Different schools of thought have defined the term probability differently. The various schools of thought which have defined probability are discussed briefly.

Classical Approach (Priori Probability)

The classical approach is the oldest method of measuring probabilities and has its origin in gambling games. According to this approach, the probability is the ratio of favorable events to the total number of equally likely events. If we toss a coin, we are certain that the head or tail will come up. The probability of the coin coming down is 1, of the head coming up is 1/2 and of the tail coming up is 1/2.

$$P = \frac{\text{Number of favorable cases}}{\text{total number of cases}}$$

If an event can occur in '*a*' way and fail to occur in 'b' ways and these are equally to occur, then the probability of the event occurring, $\frac{a}{a+b}$is denoted by P. Such probabilities are also known as unitary or theoretical or mathematical probability. P is the probability of the event happening and q is the probability of its not happening.

$P = \frac{a}{a+b}$ and $q = \frac{b}{a+b}$

Hence $p + q = \frac{a}{a+b} + \frac{b}{a+b} = 1$

Therefore, $p + q = 1$

Probability can be expressed either as ratio, fraction or percentage, such as $\frac{1}{2}$ or 0.5 or 50%.

Limitations of Classical Approach:

- We cannot apply this method when the total number of cases cannot be calculated.
- When the outcomes of a random experiment are not equally likely, this method cannot be applied.

Relative Frequency Theory of probability:

Classical approach is useful for solving problems involving game of chances—throwing dice, coins, etc. but if applied to other types of problems it does not provide answers. For instance, if a man jumps from a height of 300 feet, the probability of his survival will, not be 50%, since survival and death are not equally alike.

Similarly, the prices of shares of a Joint Stock Company have three alternatives i.e., the prices may remain constant, or prices may go up or prices may go down. Thus, the classical approach fails to answer questions of these type.

If we toss a coin 20 times, the classical probability suggests that we; should have heads ten times. But in practice it may not be so. This empirical approach suggests, that if a coin is tossed a large number of times, say, 1,000 times, we can expect 50% heads and 50% tails. Vor Miscs explained, "If the experiment be repeated a large number of times under essentially identical conditions, the limiting value of the ratio of the number of times the event **A** happens to the total, number of trials of the experiments as the number of trials increases indefinitely, is called the probability of the occurrence of **A**".

Thus, $P(A) = \lim_{n\to\infty} \frac{m}{n}$

The happening of an event is determined on the basis of past experience or on the basis of relative frequency of success in the past.

(i) the relative frequency obtained on the basis of past experience can be shown to come very close to the classical probability. For example, as said earlier, a coin is tossed for 6 times, we may not get exactly 3 heads and 3 tails. But the coin is tossed for larger number of times, say 10,000 times, we can expect heads and tails very close to 50%.

(ii) There are certain laws, according to which the 'occurrence' or 'non-occurrence of the events take place. Posterior probabilities, also called Empirical Probabilities are based on experiences of the past and on experiments conducted. Thus, relative frequency can be termed as a measure of probability, and it is calculated on the basis of empirical or statistical findings. For instance, if a machine produces 100 articles in the past, 2 particles were found to be defective, then the probability of the defective articles is 2/100 or 2%.

Subjective approach

This approach of probability is completely based on the personal belief of personal discretion of a person. Since different persons may assign different probabilities, one cannot arrive at objective conclusions using probabilities assigned by this subjective method.

Probability

1. A survey of 50 students at a certain college about the number of extracurricular activities resulted in the data shown.

No. of activities	0	1	2	3	4	5
frequency	8	20	12	6	3	1

a. Let A be the event that a student participates in at least 1 activity. Find P(A)

b. Let B be the event that a student participates in 3 or more activities. Find P(B)

c. What is the probability that a student participates in exactly 2 activities?

d. What is the probability that a student participates at most 2 activities?

2. What is the chance that a non-leap year should have fifty-three Sundays?

3. If two coins are tossed once, what is the probability of getting

a) both heads b) at least one head?

4. A card is drawn from a well shuffled pack of playing cards. Find the probability that it is

a) either a diamond or a king b) black or queen.

5. Mr. A and Mr. B both are interested to attend a seminar in central department of statistics, TU. The chance of attending a seminar by Mr. A is 0.6 and that by Mr. B is 0.3. They both can also attend the seminar. What is the probability that at least any one of them will attend the seminar?

6. The probability that an integrated circuit chip will have defecting etching is 0.12, the probability that it will have a crack defect is 0.29, and the probability that it has both defect is 0.07.

a) What is the probability that a newly manufactured chip will have at least one of two defects?
b) What is the probability that a newly manufactured chip will have neither any kind of defect?

6. The salesman has a 60% chance of making a sale to each customer. The behavior of successive customers is independent. If two customer A and B enter, what is the probability that the salesman will make a sale to A or B?

7. A bag contains 7 red, 12 white and 4 green balls. Three balls are drawn randomly. What is the probability that,

a) 3 balls are all white
b) 3 balls are one of each color?
c) 3 balls are same color?

8. A bag contains 3 red, 4 white and 9 black balls. Three balls are drawn randomly. What is the probability that,

a) all are black
b) all are of different color

9. There are 3 economists, 4 engineers, 2 statisticians and 1 doctor. A committee of 4 from among them is be formed. Find the probability that the committee

a) consists one of each kind
b) has at least one economist
c) has the doctor as a member and three others.

10. During a study of an auto accident, the highway safety council found that 60% of all accidents occur at night, 52% are alcohol-related, and 37% occur at night and are alcohol-related.

a) What is the probability that an accident was alcohol-related, given that it occurred at night?
b) What is the probability that an accident occurred at night, given that it was alcohol-related?

11. There are three machines A, B, and C producing 1000, 2000 and 3000 articles per hour respectively. These machines are known to be producing 1%, 2% and 3% defectives respectively. One article is selected at random form an hour production of the three machines and found to be defective. What is the probability that the article is produced from:

a) Machine A
b) Machine B?

12. In a certain factory, machines I, II, and III are all producing springs of the same length. Of their production, machine I, II and III produce 2%, 1% and 3% defective springs respectively. Of the total production of spring in the factory, machine I produces 35%, machine II produces 25% and machine III produces 40%. If one spring is selected at random from the total spring produced in a day, find

a) The probability that it is defective
b) The conditional probability that it was produced by machine III.

Binomial Distribution

In probability theory and statistics, the **binomial distribution** is the discrete probability distribution that gives only two possible results in an experiment, either **Success or Failure**. For example, if we toss a coin, there could be only two possible outcomes: heads or tails, and if any test is taken, then there could be only two results: pass or fail. This distribution is also called a binomial probability distribution.

There are two parameters n and p used here in a binomial distribution. The variable 'n' states the number of times the experiment runs, and the variable 'p' tells the probability of any one outcome.

Definition: A discrete random variable X is said to follow binomial distribution if it takes only the positive integer's values i.e., 0, 1, 2, …, n and probability mass function is as follows:

$P(X=x) = p(x) = {}^{n}C_{x}\ p^{x}q^{n-x}$

$= \frac{n!}{x!(n-x)!}\ p^{x}\ q^{n-x}$, for x= 0, 1, 2, …, n

= 0, otherwise

Where, p is probability of success and q=1-p, x the number of success(say) in a series of 'n' independent trials ($x \geq 0$)

Conditions for Binomial Distribution

We get the Binomial Distribution under the following experimental conditions:

- The number of trials `*n*' is finite
- The trials are independent of each other
- The probability of success `*p*' is same for each trial
- Each trial must result in a success or a failure.

Characteristics of Binomial Distribution

- Binomial distribution is a discrete distribution i.e., *X* can take values 0, 1, 2, ... *n* where `*n*' is finite.

- Constants of the distributions are:

Mean = np; Variance = npq; Standard deviation = √npq

$$\text{Skewness} = \frac{q-p}{\sqrt{npq}} \; ; \quad \text{Kurtosis} = \frac{1-6pq}{npq}$$

- It may have one or two modes.
- If X~B(n_1, p) and Y~B(n_2,p) and that X and Y are independent then X+Y~B(n_1+n_2,p)
- If `*n*' independent trials are repeated *N* times the expected frequency of `*x*' successes are $N \times {}^{n}C_x \, p^x \, q^{n-x}$
- If $p = 0.5$, the distribution is symmetric.

Some examples of Binomial Distribution

1)Comment on the following 'The mean of binomial distribution is 5 and its variance is 9'.

Solution:

Given mean np = 5 and variance npq = 9

$$\therefore \frac{\textit{Value of variance}}{\textit{Value of mean}} = \frac{npq}{np} = \frac{9}{5} \therefore q = \frac{9}{5} > 1 \qquad \text{is not possible}$$

as $0 \leq q \leq 1$ and hence the given statement is wrong.

2)Eight coins are tossed simultaneously. Find the probability of getting at least six heads.

Solution:

Here $n=8 \quad p = P(\text{head}) = \frac{1}{2} \qquad q = 1 - \frac{1}{2} = \frac{1}{2}$

Trials satisfy conditions of Binomial distribution

Hence $P(X = x) = {}_nC_x\, p^x\, q^{n-x} \quad x = 0, 1, 2, \ldots n$

$= {}_8C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{8-x} \quad x = 0,1,2,\ldots.8$

$= {}_8c_x \left(\frac{1}{2}\right)^{x+8-x}$

$= {}_8C_x \left(\frac{1}{2}\right)^8$

$\therefore P(X = x) = \frac{{}_8C_x}{256}$

P (getting atleast six heads)

$= P(x \geq 6)$

$= P(x = 6) + P(x = 7) + P(x = 8)$

$= \frac{{}_8C_6}{256} + \frac{{}_8C_7}{256} + \frac{{}_8C_8}{256}$

$= \frac{28}{256} + \frac{8}{256} + \frac{1}{256}$

$= \frac{37}{256}$

3)Ten coins are tossed simultaneously. Find the probability of getting (i) at least seven heads (ii) exactly seven heads (iii) at most seven heads.

Solution:

X denote the number of heads appear

$$P(X = x) \quad = \quad {}^{n}C_{x}\, p^{x}\, q^{n-x}\,,\ x = 0,1,2,\ldots n$$

Given: $p = P(\text{head}) = \frac{1}{2}$ $q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}$ and $n = 10$

$$\therefore X \sim B\left(10, \frac{1}{2}\right)$$

$$= 10C_{x}\left(\frac{1}{2}\right)^{x}\left(\frac{1}{2}\right)^{10-x}$$

$$P(X = x) = \frac{10C_{x}}{1024}$$

(i) P (atleast seven heads)

$$P(X \geq 7) = P(x = 7) + P(x = 8) + P(x = 9) + P(x = 10)$$

$$= \frac{10C_{7}}{1024} + \frac{10C_{8}}{1024} + \frac{10C_{9}}{1024}\frac{10C_{10}}{1024}$$

$$= \frac{120}{1024} + \frac{45}{1024} + \frac{10}{1024} + \frac{1}{1024}$$

$$= \frac{176}{1024}$$

(ii) P (exactly 7 heads)

$$P(x=7) = \frac{10C_{7}}{1024} = \frac{120}{1024}$$

(iii) P (atmost 7 heads)

$$= P(x \leq 7) = 1 - P(x > 7)$$

$$= 1 - \{P(x = 8) + P(x = 9) + P(x = 10)\}$$

$$= 1 - 1 - \left\{\frac{10C_{8}}{1024} + \frac{10C_{9}}{1024} + \frac{10C_{10}}{1024}\right\}$$

$$= 1 - \frac{56}{1024}$$

$$= \frac{968}{1024}$$

4)With usual notation find p for Binomial random variable X if $n = 6$ and $9\ P\ (x = 4) = P\ (x = 2)$

Solution:

$$P\ (X = x) = {}_nC_x\ p^x\ q^{n-x}\ ,\ x = 0,1,2,\ldots n$$

$$X \sim B\ (6, p) \Rightarrow P(X = x) = {}_6C_x p^x\ q^{(6-x)}$$

Also $\quad 9 \times P\ (X = 4) = P\ (X = 2)$

$$\Rightarrow \quad 9 \times {}_6C_4 p^4 q^2 = {}_6C_2\ p^2\ q^4$$

$$\Rightarrow \quad 9\ p^2 = q^2$$

$$\Rightarrow \quad 3p = q \qquad \text{as } p, q > 0$$

$$3p = 1 - p$$

$$4p = 1$$

$$\Rightarrow \quad p = \frac{1}{4} = 0.25$$

5)A Binomial distribution has parameters n=5 and p=1/4. Find the Skewness and Kurtosis.

Solution:

Here we are given n=5 and $p = \frac{1}{4}$

$$\text{Skewness} = \frac{q - p}{\sqrt{npq}}$$

$$= \frac{\frac{3}{4} - \frac{1}{4}}{\sqrt{5 \times \frac{1}{4} \times \frac{3}{4}}}$$

$$= \frac{\frac{2}{4}}{\sqrt{\frac{15}{16}}}$$

$$= \frac{2}{\sqrt{15}}$$

Finding: The distribution is positively skewed.

Kurtosis

$$\text{Kurtosis} = \frac{1-6pq}{npq}$$

$$= \frac{1-6\times\frac{1}{4}\times\frac{3}{4}}{\frac{15}{16}}$$

$$= \frac{\frac{-2}{16}}{\frac{15}{16}}$$

$$= \frac{-2}{15}$$

$$= -0.1333$$

Finding: The distribution is Platykurtic.

6)In a Binomial distribution with 7 trials, $P(X=3) = P(X=4)$ Check whether it is a symmetrical distribution?

Solution:

A Binomial distribution is said to be symmetrical if $p = q = \frac{1}{2}$

Given: $P(X=3) = P(X=4)$

$X \sim B(n, p)$

$P(X = x) = {}^{n}C_{x}\, p^{x}\, q^{n-x}$, $x = 0,1,2,\ldots n$

${}^{n}C_{3}\, p^{3} q^{n-3} = {}^{n}C_{4}\, p^{4}\, q^{n-4}$

${}^{7}C_{3}\, p^{3} q^{4} = {}^{7}C_{4}\, p^{4} q^{3}$ note that ${}^{7}C_{3} = {}^{7}C_{4}$

On simplifying, we have

$$q = p$$

$$1-p = p$$

$$1 = 2p$$

$$p = \frac{1}{2}$$

$$q = \frac{1}{2}$$

Hence the given Binomial distribution is symmetrical.

7)From a pack of 52 cards 4 cards are drawn one after another with replacement. Find the mean and variance of the distribution of the number of kings.

Solution:

Success *X*=event of getting king in a draw

p=probability of getting king in a single trial

$$p = \frac{4}{52}$$

$$= \frac{1}{13}$$

This is constant for each trial.

Hence, it is a binomial distribution with $n=4$ and $p = \frac{1}{13}$

$$\text{Mean} = np = 4\times\frac{1}{13} = \frac{4}{13}$$

$$\text{Variance} = npq = \frac{4}{13}\times\frac{12}{13} = \frac{48}{169}$$

8)In a street of 200 families, 40 families purchase the Hindu newspaper. Among the families a sample of 10 families is selected, find the probability that

i. Only one family purchase the news paper

ii. No family purchasing

iii. Not more than one family purchase it

Solution:

X ~ B (n, p)

$P(X = x) = nC_x p^x q^{n-x}$, x= 0,1,2,...n

Let *X* denote the number of families purchasing Hindu Paper

p = Probability of their family purchasing the Hindu

$$p = \frac{40}{200} = \frac{1}{5}$$

$$q = \frac{4}{5}$$

$$n = 10$$

(i) Only one family purchase the Hindu

$$P(X=1) = nC_1\, p^1\, q^{n-1}$$

$$= 10C_1 \times \frac{1}{5} \times \left(\frac{4}{5}\right)^9$$

$$= 2 \times \left(\frac{4}{5}\right)^9$$

(ii) No family purchasing the Hindu

$$P(X=0) = 10C_0 (\tfrac{1}{5})^0 (\tfrac{4}{5})^{10}$$

$$= (\frac{4}{10})^{10}$$

(iii) Not more than one family purchasing The Hindu means that $X \leq 1$

$$P(X \leq 1) = P[x=0] + P[x=1]$$

$$= 10C_0 (\tfrac{1}{5})^0 (\tfrac{4}{5})^{10} + 10C_1 (\tfrac{1}{5})^1 (\tfrac{4}{9})^9$$

$$= (\frac{4}{5})^{10} + 10 \times (\tfrac{1}{5})^1 (\tfrac{4}{5})^9$$

$$= (\frac{4}{5})^9 \left[(\tfrac{4}{5}) + 2\right]$$

$$= (\frac{4}{5})^9 (\tfrac{14}{5})$$

9)In a tourist spot, 80% of tourists are repeated visitors. Find the distribution of the numbers of repeated visitors among 4 selected peoples visiting the place. Also find its mode or the maximum visits by a visitor.

Solution:

Let the random variable *X* denote the number of repeated visitors.

X ~ B (n, p)

P (X = x) = $nC_x p^x q^{n-x}$, x= 0,1,2,...n

It is a Binomial Distribution with n=4

$$p = \frac{80}{100} = \frac{4}{5} \qquad q = 1-p = 1 - \frac{4}{5} = \frac{1}{5}$$

$$P(x=0) = 4C_0 \left(\frac{4}{5}\right) \left(\frac{1}{5}\right)^4 = \left(\frac{1}{625}\right)$$

$$P(x=1) = 4C_1 \left(\frac{4}{5}\right)\left(\frac{1}{5}\right)^3 = \frac{16}{625}$$

$$P(x=2) = 4C_2 \left(\frac{4}{5}\right)^2\left(\frac{1}{5}\right)^1 = \left(\frac{96}{625}\right)$$

$$P(x=3) = 4C_3 \left(\frac{4}{5}\right)^3\left(\frac{1}{5}\right)^1 = \left(\frac{256}{625}\right)$$

$$P(x=4) = 4C_4 \left(\frac{4}{5}\right)^4 \left(\frac{1}{5}\right)^0 = \left(\frac{256}{625}\right)$$

The probability distribution is given below.

$X=x$	0	1	2	3	4
$P(X=x)$	$\frac{1}{625}$	$\frac{16}{625}$	$\frac{96}{625}$	$\frac{256}{625}$	$\frac{256}{625}$

10)In a college, 60% of the students are boys. A sample of 4 students of the college, is taken, find the minimum number of boys should it have so that probability up to that number is ≥ 1/2.

Solution:

It is given that 60% of the students at the college are boys and the selection probability for a boy is 60% or 0.6 As we are taking four samples, the number of trials $n = 4$. The selection process is independent.

$X \sim B\ (n,\ p)$

$P\ (X = x) = \quad {}_nC_x\ p^x\ q^{n-x}$, $x = 0,1,2,\ldots n$

Let X be the number of boys so that $P(X \le x) \ge \frac{1}{2}$

If $x = 0 \quad P(X \le 0) = {}_4C_0 (\frac{3}{5})^0\ (\frac{2}{5})^4 = \frac{16}{625} < \frac{1}{2}$

$x=1 \qquad P(X \le 1) = P\ (x = 0) + P\ (x=1)$

$$= {}_4C_0 \left(\frac{3}{5}\right)^0 \left(\frac{2}{5}\right)^4 + {}_4C_1 \left(\frac{3}{5}\right)^1 \left(\frac{2}{5}\right)^3$$

$$= \frac{16}{625} + \frac{96}{625} = \frac{112}{625} < \frac{1}{2}$$

$x=2 \qquad P(X \le 2) = P(x=0) + P(x=1) + P(x=2)$

$$= \frac{112}{625} + P(x=2) = \frac{112}{625} + {}_4C_2 \left(\frac{3}{5}\right)^2 \left(\frac{2}{5}\right)^2$$

$$= \frac{112}{625} + \frac{216}{625} = \frac{328}{625} > \frac{1}{2}$$

Therefore, the sample should contain a minimum of 2 boys.

Binomial Distribution

1. Ten coins are tossed simultaneously. Find the probability of obtaining,

   a) No head
   b) Exactly 6 heads
   c) At least one head
   d) Not more than three heads i.e. at most 3 heads

2. A multiple choice test has 5 questions. There are 4 choices for each question. A student who has not studied for the test decides to answer all questions randomly. What is the probability that he will get:

   a) Five questions, correct?
   b) At least four questions, correct?

3. In a local hospital 48% of all babies born are males. On a particular day five babies are born. What is the probability that:

   a) None of them are male
   b) Two of them are male
   c) At least one is male

4. Out of 9000 families 4 children each, how many families would you expect to have 3 boys and 1 girl, the birth of male child and female child is assumed equal?

5. If the mean of a binomial distribution is 0.4 and its standard deviation is 0.6, find the probability of at least one success.

6. In a binomial distribution with 6 independent trials the probabilities of 3 and 4 successes are found to be 0.2457 and 0.0819. find the parameter ‘p’ of the binomial distribution.

7. It is known that on an average the probability that Shyam will win the game is 40%. What is the probability that out of 8 games Shyam will win between 2 and 6 games? (including 2 and 6)

8. Forty five percent of the Nepalese workers have been gone abroad are illegal. If in a sample of six, Nepalese workers who have gone abroad, what is the probability that

   a) Three are illegal
   b) All are legal
   c) At least one is legal

Fitting of Binomial Distribution

If n independent trials are repeated N times and satisfying the condition of binomial distribution, then theoretical or expected frequencies of x successes is given by

$$f(x)= N\ {}^{n}C_{x}\ p^{x}\ q^{n-x}, \text{ where } x= 0, 1, 2, \ldots., n$$

The following steps are generally followed to fit binomial distribution.

- First of all, we compute the mean of given frequency distribution by using the following formula $\bar{X} = \frac{\sum fX}{N}$

  If p is known (given), it is not necessary to find mean.
- Equate the value of mean with np to find the values of p and q,

  i.e., $\bar{X} = np$
- Calculate the expected frequency by using the formula

  $f(x)= N\ {}^{n}C_{x}\ p^{x}\ q^{n-x}$, where x= 0, 1, 2, …., n

Practical problems

1. Fit the binomial distribution for the following data:

No. of heads	0	1	2	3	4	Total
Frequency	28	62	46	20	4	160

2. Fit the binomial distribution for the following data

X	0	1	2	3	4	5	Total
f	1	4	10	31	26	13	85

Estimation

1. A random sample of size 36 from a finite population consisting 101 units. If the population standard deviation is 12.6, find the standard error od sample mean when the sample is drawn

a) with replacement b) without replacement

2. If sample mean is 20, population standard deviation is 3 and sample size is 64, find the interval estimate of the population mean. (α =5%)

3. Systolic blood pressure of a sample of 400 males was taken. Sample mean blood pressure was found to be 128 mm and standard deviation 13.05 mm. Find 95% confidence limits of blood pressure within which the population mean would lie?

4. From a population of 540, a sample of 60 individuals is taken. From this sample, the mean is found to be 6.2 and standard deviation 1.368,

a) Find the estimated standard error of the mean
b) Construct a 95% confidence interval for the mean.

5. A random sample of 500 oranges was taken from a large consignment and it was observed that 65 were found to be bad. Find the standard error of proportion of bad oranges.

6. Out of 300 households in a town 123 have T.V. sets. Find 95% confidence limits to the value of the proportion of the households with T.V. sets in the whole town.

7. 400 oranges are taken from a large consignment and 50 of them are found to be bad. Estimate the percentage of the bad oranges in the consignment and assign 95% limits within which the percentage lies.

8. In a survey, 200 people were asked to identify their major sources of news information; 110 stated that their major source was television news.

a) Construct a 95% confidence interval for the proportion of the people in the population who consider television their major source of news information.
b) What happens to the width of a confidence interval as the confidence level is increased?
c) What happens to the width of a confidence interval as the sample size is increased?

Sample Size

1. In a study of time and motion in a factory, the supervisor estimated the S.D. to be 0.95 seconds. If you want to be 95% confident that the error will not exceed 0.01 second. What should be the size of the sample to estimate population mean?

2. The mean and standard deviation of a random sample of 49 were found to be 100 and 10 respectively. If the investigator wants to be 95% confident that the error in estimate of a population mean should not exceed 2 how many additional observations are required?

3. Potato chips distributor wants to estimate average monthly sales of its product. If the standard deviation is Rs. 12, find the sample size if the maximum error is not more than Rs. 3 with 99% level of confidence.

4. It is desired to estimate the proportion of junior executives who change their first job within the first five years. This proportion is to be estimate within 3% error and 0.95 degree of confidence is to be used. A random study conducted several years ago revealed that 30% of such junior executives changed their first job within 5 years.

a) How large a sample is required to update the study?
b) How large should the sample be if no such previous estimates are available?

Poisson distribution

A discrete random variable X is said to have a Poisson distribution, with parameter $\lambda>0$, if it has a probability mass function given by

$$P(X=x) = \begin{cases} \frac{e^{-\lambda}\lambda^{x}}{x!}; & x=0,1,2,..... \\ 0 & otherwise \end{cases}$$

Conditions:

- The number of trails '*n*' is indefinitely large i.e., $n\rightarrow \infty$
- The probability of a success '*p*' for each trial is very small i.e., $p\rightarrow 0$
- $np=\lambda$ is finite
- Events are Independent

Characteristics of Poisson Distribution

- Poisson distribution is a discrete distribution i.e., X can take values 0, 1, 2, …
- p is small, q is large, and n is indefinitely large i.e., $p\rightarrow 0$ $q\rightarrow 1$ and $n\rightarrow 3$ and np is finite
- Values of constants: (a) Mean = λ = variance (b) Standard deviation = $\sqrt{\lambda}$ (c) Skewness = $1/\sqrt{\lambda}$ (iv) Kurtosis = $1/\lambda$
- It may have one or two modes
- If X and Y are two independent Poisson variates, X+Y is also a Poisson variate.
- If X and Y are two independent Poisson variates, X-Y need not be a Poisson variate.
- Poisson distribution is positively skewed.
- It is leptokurtic.

1)If 2% of electric bulbs manufactured by a certain company are defective find the probability that in a sample of 200 bulbs (i) less than 2 bulbs are defective (ii) more than 3 bulbs are defective. [$e^{-4} = 0.0183$]

Solution:

Let X denote the number of defective bulbs

$X \sim P(\lambda)$

$$\therefore \quad P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad x = 0, 1, 2, \ldots \infty$$

Given $p = P$ (a defective bulb) $= 2\% = \frac{2}{100} = 0.02$

$$n = 200$$

$$\therefore \quad \lambda = np = 200 \times 0.02 = 4$$

$$\therefore \quad P(X = x) = \frac{e^{-4}4^x}{x!}, \quad x = 0, 1, 2, \ldots \infty$$

(i) P (less than 2 bulbs are defective)

$$= P(X < 2)$$

$$= P(x = 0) + P(x = 1)$$

$$= \frac{e^{-4}.4^0}{0!} + \frac{e^{-4}.4^1}{1!}$$

$$= e^{-4}(1 + 4)$$

$$= 0.0183 \times 5$$

$$= 0.0915$$

(ii) P (more than 3 defectives)

$$= P(X > 3)$$

$$= 1 - P(X \leq 3)$$

$$= 1 - \{P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3)\}$$

$$= 1 - \left\{\frac{e^{-4}.4^0}{0!} + \frac{e^{-4}.4^1}{1!} + \frac{e^{-4}.4^2}{2!}\frac{e^{-4}.4^3}{3!}\right\}$$

$$= 1 - e^{-4}\{1 + 4 + 8 + 10.667\}$$

$$= 1 - 0.0183 \times 23.667$$

$$= 0.567$$

2)In a Poisson distribution $3P(X=2) = P(X=4)$. Find its parameter 'λ

Solution:

The *pmf* of Poisson distribution is $P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!}$, $x = 0, 1, 2, \ldots\infty$,

Given $3\,P(X=2) = P(X=4)$

$$3.\frac{e^{-\lambda}\lambda^2}{2!} = \frac{e^{-\lambda}.\lambda^4}{4!}$$

$$\lambda^2 = \frac{3\times 4!}{2!} = 36$$

$$\therefore \lambda = 6 \text{ as } \lambda > 0$$

3)Find the skewness and kurtosis of a Poisson variate with parameter 4.

Solution:

$$\lambda = 4$$

$$\text{Skewness} = \frac{1}{\sqrt{\lambda}} = \frac{1}{\sqrt{4}} = \frac{1}{2}$$

$$\text{Kurtosis} = \frac{1}{\lambda} = \frac{1}{4}$$

4)If there are 400 errors in a book of 1000 pages, find the probability that a randomly chosen page from the book has exactly 3 errors.

Solution:

Let X denote the number of errors in pages

$X \sim P(\lambda)$

$$\therefore P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad x = 0, 1, 2, \ldots\infty$$

The average number of errors per page $= \frac{400}{1000}$

i.e., $\lambda = \frac{400}{1000} = 0.4$

$$P(X=3) = \frac{e^{-\lambda}\lambda^3}{3!} = \frac{e^{-0.4}(0.4)^3}{3\times 2\times 1}$$

$$= \frac{0.6703\times 0.064}{6}$$

$$= 0.00715$$

5)If X is a Poisson variate with $P(X=0) = 0.2725$, find $P(X=1)$

Solution:

$X \sim P(\lambda)$

$\therefore P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$ $\quad x = 0, 1, 2, \ldots \infty$

$P(X=0) = 0.2725$

$\frac{e^{-\lambda}\lambda^0}{0!} = 0.2725$

$e^{-\lambda} = 0.2725$

$\lambda = 1.3$ (from the table of values of e^{-m})

$$P(x=1) = \frac{e^{-\lambda}\lambda^1}{1!} = \frac{e^{-1.3} \times 1.3^1}{1!}$$

$$= 0.2725 \times 1.3$$

$$= 0.3543$$

6)The probability of safety pin manufactured by a firm to be defective is 0.04. (i) Find the probability that a box containing 100 such pins have one defective pin. (ii) Among 200 such boxes, how many boxes will have no defective pin

Solution:

Let X denote the number boxes with defective pins

$X \sim P(\lambda)$

$\therefore \quad P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$ $\quad x = 0, 1, 2, \ldots \infty$

$p = 0.04$

$n = 100$

$\lambda = np = 4$

(i) $P(X=1) = \frac{e^{-\lambda}\lambda}{1!} = e^{-4}(4) = 0.0183 \times 4$

$= 0.0732$

(ii) $P(X=0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-4} = 0.0183$

Number of boxes having no defective pin $= 200 \times 0.0183$

$= 3.660$

$= 4$

Poisson Distribution

1. If the prices of new cars increase an average of four times every 3 years, find the probability of

   a) No price hikes in a randomly selected period of 3 years.
   b) Two price hikes.
   c) Five or more price hikes.

2. The number of accidents that occurs on an assembly line with an average of three accidents per week, what is the probability that:

   a) A particular week will be accident free.
   b) Exactly five accidents will occur in a week.
   c) At least three accidents in a week.

3. The number of accidents in a year attributed to taxi driver in a city, follows Poisson distribution with mean 3. Out of 1000 taxi drivers, find approximately the number of drivers with

   a) No accident in a year.
   b) More than 3 accidents in a year.

4. The probability of getting no misprint in a page of book is e^{-4}. What is the probability that a page contains more than 2 misprints?

5. Calculate mean and variance of a Poisson variable X, if P(X=4) = P(X=5).

6. The standard deviation of a Poisson distribution is 2. Find the probability that X=3.

7. If mean and variance of a distribution is 2, find the P(X<3).

Poisson is a Good Approximation to Binomial

Poisson approximation to binomial when np<10, n≥20 and p≤0.05

Now, $P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!}$

$= \frac{e^{-np}(np)^x}{x!}$, x=0, 1, 2, ….

$\lambda = np$ = mean

1. If 3% of electric bulbs manufactured by a company are defective, use Poison distribution to find the probability that in a sample of 100 bulbs

   a) None is defective
   b) 3 bulbs are defective.

2. In a certain manufacturing process 5% of the tools produced turn out to be defective. Find the probability that in a sample of 40 tools, at most 2 will be defective.

3. Assuming that the probability of a fatal accident in a factory during the year is 1/120. Calculate the probability that in a factory employing 30 workers there will be

a) Exactly 2 fatal accidents in a year.
b) At least 2 fatal accidents in a year.
c) At most 2 fatal accidents in a year.

Fitting of Poisson Distribution

If a series of trials is repeated N times and satisfied the condition of Poisson distribution, then expected or theoretical frequency for x successes is given by,

$f(x) = N\ P(X=x)$

$= N \frac{e^{-\lambda}\lambda^{x}}{x!}$, x=0, 1, 2, ….

The following steps are generally followed to fit Poisson distribution.

1. First of all, we calculate the mean of given frequency distribution by using following formula.

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

if λ is known (given), it is not necessary to find mean.

2. The computed mean is equating to the parameter of Poisson distribution (λ) i.e. $\bar{X} = \lambda$.

3. Calculate the expected frequencies by using the formula

$f(x) = N \frac{e^{-\lambda}\lambda^{x}}{x!}$, x=0, 1, 2, ….

Practical problem

1. Fit a Poisson distribution to the following frequency distribution.

No. of heads	0	1	2	3	4	Total
Frequency	28	62	46	20	4	160

2. Fit a Poisson distribution to the following frequency distribution.

X	0	1	2	3	4	5	Total
f	1	4	10	31	26	13	85