MODULE-1: INTRODUCTION AND IMPORTANCE - MEANING OF STATISTICS AND THEIR FUNCTION

Learning objective

One can understand the importance, use and meaning of Statistics after going through this module.

STATISTICS

Meaning of Statistics

The word *'Statistics'* has come from the Latin word *'status'* ,the Italian word *'statista'* or the German word *'statistik'* , the French word *'statistique'*, each of which means political state.

- In early days, facts and figures about the financial resources, births and deaths, army strength and income were collected for the purpose of efficient administration which was called statistics i.e, anything pertaining to the state.
- Now a days Statistics is not only the science of state but it plays an important role in all walks of life and in all branch of scientific enquiry. In fact, statistics has become one of the essential tools in modern biology.
- Usually, the word 'statistics' carries different meanings depending on the occasion in which it is used.
    - For e.g., it may mean *statistical data* which refers to quantitative information, *statistical method* which means the methods dealing with quantitative information or statistical measures of a sample. i.e., Arithmetic mean, standard deviation etc. of a sample.
- By statistical data, we mean the aggregate of facts which are affected by multiplicity of causes, numerically expressed, estimated to a reasonable standard of accuracy and collected in a systematic manner for a pre -determined purpose.
- Statistical method includes collection, classification, tabulation, presentation, analysis and interpretation of data.
- Biostatistics is the application of statistical methods to the problems of biology including human biology, medicine and public health.
- Biostatistics is also called Biometry meaning "biological measurement".

Functions of Statistics

- Presents facts in a definite form
- Simplifies mass of figures
- Facilitates comparison
- Helps in formulating hypothesis
- Helps in testing the hypothesis
- Helps in prediction
- Helps in the formulation of suitable policies

Limitations of Statistics

- Statistics does not deal with the individuals.
- It deals only with quantitative characters. However qualitative characters can be numerically expressed and analyzed.
- For eg. Intelligence of students by marks obtained, poverty by income received.
- Statistical results are true only on an average.

- Statistics is only one of the methods of studying a problem.
- Statistics may be sometimes misused, if not properly interpreted.

DEFINITIONS

Population

- A set or collection of objects pertaining to a phenomenon of statistical enquiry is referred to as universe or population or census. (e.g.) animals in a farm.

Sample

- When a few units are selected from a population, it is called as a sample. (e.g.) animals of a particular breed in a farm.

Variable

- The quantitative or numerical characteristic of the data is called as a variable. (e.g.) weight of an animal.

Constant

- It is a numerical value, which is same for all the units in the population. (e.g.) no. of credit hours for B.V.Sc students.

Attribute

- It refers to the qualitative character of the items chosen. (e.g.) breed of an animal.

Parameter

- A statistical measure pertaining to a population is called as a parameter. (e.g.) mean, standard deviation of the population.

Statistic

- A statistical measure pertaining to the sample is called as a statistic. (e.g.) mean, standard deviation of the sample.

Continuous variable

- If a variable takes an intermediate value between any specified interval, it is called as a continuous variable. (e.g.) the weight of animal.

Discrete or discontinuous variable

- If a variable takes only integral values, then it is called as a discrete (or) discontinuous variable. (e.g.) no. of animals in a farm.

MODULE-2: COLLECTION AND CLASSIFICATION OF DATA

- Learning objective
- The learner will get an idea of the ways of collecting and simplifying the data after going through this module.

COLLECTION OF DATA

A statistical investigation always begins with collection of data. One can collect the data either by himself or from available records.

- The data collected by the investigator himself or by his agent from the sample or population are called as the **primary data**.
- The source from which one gathers primary data is called as the primary source.
- The data collected from the available sources is known as **secondary data**.
- The source from which we are getting secondary data is known as secondary source.

PRIMARY DATA

The data collected originally or the first hand information of facts.

Methods of collecting Primary Data

- *Direct personal observation:* The investigator himself goes to the field of enquiry and collects the data.
- *Indirect personal observation:* The investigator collects data from a third person (called as witness), who knows about the data being gathered.
- *Data collection through agents, local reporters etc:* Here the investigator appoints some person called agents or local reporters on his behalf to collect information.
- *Data collection through questionnaires:* The investigator prepares the needed information for the particular study in the form of questions,called questionnaires and sends the same to the respondents to collect data from the respondents.

Merits and Demerits of Primary Data

Methods	Merits	Demerits
Direct personal observation	• It is very accurate. • Intensive details can be collected.	• Expensive in terms of time and money. • Not suitable when the field of enquiry is large.
Indirect personal observation.	• It saves time.	• Witness should possess thorough knowledge of the facts regarding the problem of investigation. • Witness must be willing to give information.

Data collection through agents and local reporters etc.	• It saves time. • Large area can be covered	• The agents will collect information in their own fashion. • Only approximate results can be obtained. • It is expensive.
Data collection through questionnaires	• It saves time. • It is less expensive. • Geographically dispersed area can be covered	• It cannot be used if the informants are illiterate. • Response may be poor. • Possibility of vague/inaccurate answers.

Note

- Of all the types of collection of primary data if the questionnaires are framed skillfully so that it can be answered easily and if there is a compelling force through which we can collect the questionnaires then the data collection through questionnaires will be the best method.

SECONDARY DATA

The data collected from the available sources like published reports, documents, journals etc. are called secondary data.

- The source from which the secondary data are collected is called as secondary source of data.
- While the primary data are collected for a specific purpose, the secondary data are gathered from sources which were done for some other purpose.

Sources of obtaining secondary data

- Published reports/ documents of institutions, NGOs etc.
- Scientific journals
- Government reports
- Books and news papers

Merits and Demerits of Secondary Data

Merits

- It saves time, labour and money.

Demerits

- It may not be very accurate
- All the data needed may not be available
- It might have been collected by some improper methods and in some abnormal condition.

CLASSIFICATION OF DATA

Classification is the process of arranging data into sequences and groups according to their common characteristics or separating them into different but related parts.

Objectives of Classifications

- To remove unnecessary details
- To bring out explicitly the significant features in the data
- To make comparisons and drawing inferences

Methods of Classification

- *Numerical Classification*
  - Classification of data according to quantitative characters. (e.g) classification of animals in a farm according to their weight
- *Descriptive Classification*
  - Classification according to attributes i.e, qualitative characters. (e.g). classification of animals according to breeds
- *Spatial or Geographical Classification*
  - Classification according to geographical area. (e.g) district-wise livestock population in Tamil Nadu
- *Temporal or Chronological Classification*
  - Classification according to time (e.g) livestock population in different years
- *Classification according to class interval or frequency distribution*
  - When the data are grouped into classes of appropriate interval, showing the number in each class, we get frequency distribution.This is called grouped data.The original data is called raw data.
- The following is the frequency table showing the distribution of chicks in different weight classes.

Weight(in gm.)	No. of Chicks
36-40	12
40-44	25
44-48	17
48-52	05
52-56	06
56-60	10
Total	*75*

Terms used in Frequency distribution

Class Interval and Class Limits

- Data are classified or grouped into regular intervals with the range of values of the data (Class Interval) with the lower and upper limits which iare known as Class Limits.

- *True Class Interval*
  - When the Class Intervals are continuous, it is called True or Inclusive Class Interval.
- *Apparent Class Interval*
  - When there is a small gap between the upper boundary of any class and lower boundary of sucessive class, then the Class Interval is called Apparent or Exclusive Class Interval.

Class Frequency or Frequency

- Class frequency or frequency is the number of observations in that class.

Width or Length of the Class Interval

- Width or length of the Class Interval is the difference between the upper boundary and lower boundary of the same class.

Class Mark

- It is the midpoint of the class.
- It is given by half of the sum of the lower limit and upper limit of any class.

TABULATION OF DATA

It is a systematic arrangement of statistical data in columns and rows. It is the next process of condensation of data after classification. Tabulations is a mechanical part of classification. The objects of tabulation are

- Tables are more comprehensive and intelligible and carry a lasting impression on the mind of the reader.
- Tables facilitate quick comparisons.
- Tables facilitate economy of space (while presenting) and time (while reading)
- Relationship and other relevant characteristics of item can be easily marked out in tabulated data.

The following are the points to be considered carefully in preparing a table.

- The title should be short but clear and it should give a full idea of its contents.
- The column and row headings should be self explanatory.
- Footnotes may be given if absolutely necessary.
- Prominence may be given to important facts by different methods of mailing and spacing.
- To have better clarity, space should be left after every five to ten rows.
- It the table is taken from secondary data, it is advisable to give a source note for the table mentioning the source for which the data is collected.

Types of table

- Reference table or General Table
  - These table contain a great deal of summarized information. They appear usually at the end of the report in the form of appendixes.
- Text or summary tables.
  - They are used to analyse or assist in the analysis of classified data. They are included in the discussion of the body of the report.
- Statistical tables

- These are special tables used by statisticians in interpreting the results of statistical analysis. The commonly used tables are 't' tables, z table, F table etc.,

RULES IN FORMING FREQUENCY DISTRIBUTION

- The class interval should be of equal width and of such size that the characteristic features of the distribution are displayed.
- Classes should not be too large (or) too small. If too large, it will involve considerable errors in assuming that the midpoints of the class intervals are the average of that class. If too small, there will be many classes with zero frequency (or) small frequency. There are however certain type of data, which may require the use of unequal or varying class intervals.
- When there is irregular flow of data and wide fluctuating gap among the varieties, varying class intervals are to be taken (or) otherwise there may be a possibility of classes without any frequency or observations falling in that category.
- The range of the classes should cover the entire range of data and the classes must be continuous.
- It is convenient to have the midpoint of the class interval to be an integer. As a general rule, the number of classes should be in the range of 6-16 and never more than 30.

FORMATION OF FREQUENCY DISTRIBUTION

- Method of Tally Marks
- Array Method

Method of Tally Marks

- First we have to form the class interval. The difference between maximum and minimum values in the collected data are noted and it is to be divided by the number of required classes. This value should be rounded off to our convenience.
- The number of required classes can be calculated using the formula suggested either by Sturge's rule or Yule's rule.
- *Sturge's rule*

K = 1 + 3.322 log n (approx.)

where K is the number of required classes and n is the number of observations.

- *Yule's rule*

K = 2.5 x n ¼ (approx.)

where K is the number of required classes and n is the number of observations.

- After forming the class interval each should be written one below the other and for each item in the collected data a stroke is marked against the class interval in which it falls.
- Usually after every four such strokes in the class interval, the fifth item is indicated by striking the previous four strokes, thus, making it easy to count.
- These strokes are counted and this is called formation of frequency distribution by the method of tally marks.

Array Method

- An array is an orderly arrangement of the data by magnitude in the ascending or descending order.
- Form the class interval as in the previous method.
- Then arrange the given data in the ascending order of magnitude.
- From the array, we will count the number of observations belonging to each class and then we will write.
- This method is not easy, when the number of observations is large.
- We can adopt this method in cases, where the number of observations are less than 50.

MODULE-3: PRESENTATION OF DATA

- Learning objective
- This helps the reader to know about the various ways of representing the data by means of diagrams and graphs so that the voluminous numerical data can be exhibited by attractive pictures.

PRESENTATION OF DATA

Introduction

- Classification and tabulation reduce the complexity of vast and complicated statistical data but still it is not easy to interpret the tabulated data. Diagrams and graphs will catch the eye more easily than tables which provide array of figures. A glance over a graph or diagram will enable any layman (without statistical knowledge) to get an idea about the essential characteristics of the tabulated data without much strain or effort.

FUNCTIONS AND LIMITATIONS OF DIAGRAMS AND GRAPHS

Functions

- It will attract the attention of a large number of persons.
- They carry a *"birds – eye view"* impression in the human mind.
- It saves a lot of valuable time if presented in a form of suitable charts & graphs instead of pages of numerical figures.
- To facilitate comparison between two or more sets of data.
- Prediction equations can be represented by graphs and these will be much helpful in forecastings.

Limitations

- They are approximate indicators.
- Exact and accurate informations can be obtained from original tabular information.
- They cannot substitute the tabular information.
- They fail to disclose small difference when large figures are involved.

GRAPHICAL REPRESENTATION OF DATA

- Graphical representation is done when the data are classified in the form of a frequency distribution.The different graphs are
    - Histogram

- Frequency Polygon
- Frequency Curve
- Ogive
- Lorenz Curve

Histogram

- It is a vertical bar diagram without gap between the bars.
- It consists of bars erected over the true class interval, their areas being proportional to the frequencies of the respective classes.
- Since the intervals are of equal width, the height of each bar serves as a measure of the corresponding frequency.
- Draw the two diagonals in the highest modal class rectangles at its top corner to the pre and post modal rectangle corners and the x co-ordinate of the point of intersection is the mode.

Frequency Polygon

- If points are plotted with the x co-ordinate equal to the mid value of the class intervals and the corresponding frequencies as the y co-ordinate and these points are joined by means of a straight line, we obtain frequency polygon.
- These points are the midpoints of the top of the bars in the histogram.

Frequency Curve

- If points are plotted with the x co-ordinate equal to the mid value of the class intervals and the corresponding frequencies as the y co-ordinate and these points are joined by means of a smooth curve then we get frequency curve.

Ogive

- This is cumulative frequency curve.
- This curve is obtained by making use of cumulative frequency instead of the simple frequency.

Cumulative Frequency Distribution

- A frequency distribution gives the number of observations that lie in any class interval whereas the cumulative frequency distribution gives the number of frequencies that lie below any mark or above any given mark.
- When derived from a frequency distribution, the cumulative frequency distribution of one kind gives the number of observations less than the lower boundaries of the successive class and the cumulative frequency distribution of the second kind gives the number of observations that exceed the lower boundaries of the class which are respectively known as the less than and greater than cumulative frequency distribution.
- If we draw frequency polygon to the above two distribution we get cumulative frequency polygon (less than and greater than).
- If we draw a frequency curve to the above two distribution in the same graph, we get cumulative frequency curve or Ogive.
- The x co-ordinate of the point of intersection of less than and greater than cumulative frequency curve is the median.

Lorenz Curve

- This is a modification of the Ogive when the variables and the cumulative frequencies are expressed as percentages.
- It serves to measure the evenness of the distribution and is useful in picturing the distribution and dispersion of wealth, sales and profits etc.,

DIAGRAMMATIC REPRESENTATION OF DATA

Points to be followed in drawing a diagram

- For each diagram, a suitable short heading should be given.
- It should be drawn to exhibit the statistical matter clearly. It should be such as to allow its significant feature to be clearly shown out by adopting suitable scale and will depend upon the space available.
- Diagram should be drawn accurately with the help of drawing instruments.
- Colouring and different markings should be done with pencil or with colours.
- Different colours or marks or dottings are used to show different items. In such cases legend should be given for the column and item it refers. In doing so, we should see that the visual impression conveyed by the diagram is not in any way affected.
- The original data on which the diagram has been based should be given, if necessary facing the diagram as this will help the observer to see the details with clarity.
- Reference to the source of the table should be provided.

Types of a diagram

- One dimensional diagram
  - Line diagram
  - Bar Diagram
- Two dimensional (or) Area Diagram
  - Pie diagram
  - Square diagram and rectangle diagrams
- Three dimensional (or)Volume diagrams
  - Cubes
  - Spheres, Cylinders etc.
- Pictogram
  - Actual pictures

ONE DIMENSIONAL DIAGRAM

Line diagram consisting of curves and lines as well as bars

- *Line diagram*
  - This requires vertical lines to be drawn at equal intervals each of length proportional to the magnitude of the variable for the different items.
  - It has no width and hence of very poor visual effect.
  - It makes comparison easy although it is less attractive.
- *Bar Diagram*
  - It is the simplest of all statistical diagrams.
  - It consists of bars of equal width (all horizontal or vertical) standing on a common base line at equal intervals, the length of the bars being proportional to the magnitude of the variable for different items.
- *Sub-divided bar diagram or component bar diagram*

    - Sometimes the variable is capable of being sub-divided into two or more component parts each representing a sub variable.
    - In this case, all the bars are subdivided by lines in the same order so that each subdivision represents the parts in magnitude in the same scale.
    - They are properly coloured or marked differently for visual guidance.
    - Small squares should be given below the diagram containing the same colour or mark to show their significance.
- *Superimposed or Multiple bar diagram*
    - Bars may sometimes be superimposed for comparative purpose.
- *Percentage bar diagram*
    - When the component parts are expressed in percentages of the whole, the resulting bar diagram is called a percentage bar diagram.
    - In this case all the bars are of equal length.

TWO DIMENSIONAL (OR) AREA DIAGRAM

Pie diagram

- Circles with area proportional to the magnitudes of the data are drawn (i.e.) radii proportional to the square root of the magnitude of the data and the components(sub variables) are drawn with sectors proportional in area to their magnitude.
- A circle subtends an angle of 360^{o} at the centre and this represents the total. The required angle of the sector representing the component is calculated and area distinguished by different colours or markings and key for this should be given.
- It is usual to start from a horizontal radius to the right and proceed in the anti-clock wise direction giving the quantities in descending order of magnitude except the miscellaneous which is shown at the end.
- The lengths have more visual effect than areas and hence it is of less use for comparative purpose. It is commonly used to represent single observation with different components.

Square diagram and Rectangle diagrams

- Their areas should be proportional to the magnitudes of the data.
- For square diagrams, we will have to take the square root of the given figures which will give the measurement of the sides of the square. By adopting suitable scale we can draw squares.
- In the case of rectangle diagrams, if we take equal breath (width) for the rectangles, then the areas will be proportional to the lengths and hence the lengths will be proportional to the magnitude of the given variables.

THREE DIMENSIONAL (OR) VOLUME DIAGRAM

- These comprise of cubes, spheres, prisms, cylinder and blocks.
- Of these cubes are mainly used and their sides are drawn in proportion to the cube roots of the magnitudes of the data.
- They are particularly used when the data has a very wide range. In such a case, it would be difficult to represent the quantities even by squares.

PICTOGRAM

- Tabular data can also be represented by *pictogram, cartogram, maps and pictures* as these device help in attracting the attention to statistical matter which when presented in the ordinary diagrammatic form is very often ignored.
- Pictogram are diagrams of pictorial or semi-pictorial nature and are drawn in different sizes according to scale. Though they are useful in attracting the attention of the people, they very often lean on tables, ignoring the pictorial diagrams.
- They cannot be made use of with certain complicated data.

MODULE-4: MEASURES OF AVERAGES

- Learning objective
- Readers of this module will come to know the methods of condensing the data by means of a single figure and comparing two or more distributions.

MEASURES OF AVERAGE

Need of an average

SUMMATION NOTATIONS

- If there are 'n' observations and their values are given by $x_1, x_2, ..., x_n$ then the sum of the observations, $x_1 + x_2 + ... + x_n$ can be written using the symbol ' ' (read as sigma) as follows:

$$x_1 + x_2 + x_3 + ... + x_n = \sum_{i=1}^{n} x_i$$

which means the sum of x_i's, i taking values from1 to n

- When no ambiguity is likely to arise, then the suffix i can be removed and we can write as $x_1 + x_2 + ... + x_n = \sum x_i$.
- Then using the notation $\sum f_i x_i = f_1x_1 + f_2x_2 + ... + f_nx_n$ which can be written as $\sum fx$.

- A statistical average condenses a frequency distribution or raw data and presents it in one single representative number.
- It is a single value which is considered as the most representative or typical value for a set of values.
- Such a value can neither be the smallest one nor the largest one but is one which usually lies somewhere near the centre of the group.That is why an average is usually referred to as a measure of central tendency.
- It is located at a point around which most of the other values tend to cluster and therefore it is also termed as a measure of location.
- It is considered as a measure of description because it describes the main characteristics of the data.

Objectives of averaging

- To get a single, summary figure describing the prominent characteristics of the entire group of data.
- To facilitate inter-comparison of different phenomena

Different measures of averages

- The different averages are broadly classified into two groups namely mathematical averages or algebraic averages and positional averages or averages of position.
- Mathematical averages are based on all observations and are calculated by algebraic formula. They are
  - Arithmetic Mean (AM) or Mean
  - Geometric Mean (GM)
  - Harmonic Mean (HM)
- Positional averages are based on few observations and occupy certain position among the observations. They are
  - Median
  - Mode

ARITHMETIC MEAN

- It is the value obtained by dividing the sum of the values of the given items (of a variable) by the number of items. Thus,
  - Mean = (sum of the values of the items in the series / Total number of items)
  - It is usually denoted by $\bar{x}$.
- If we denote all the 'n' observations in a series by $x_1, x_2, x_3, \ldots\ldots\ldots x_n$, then arithmetic mean or mean for that series will be given by

$$\text{Arithmetic Mean} = \frac{x_1 + x_2 + x_3 + \ldots\ldots + x_n}{n}$$

- In the case of a frequency distribution if the different class marks of the 'n' classes are denoted by $x_1, x_2, \ldots, x_n$ and the corresponding frequencies by $f_1, f_2, \ldots, f_n$, then the mean of the data is

$$\bar{X} = \frac{f_1x_1 + f_2x_2 + \ldots + f_nx_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum f_i x_i}{\sum f_i} = \frac{\sum f_i x_i}{N}$$

Thus, A.M $= \sum f_i x_i / N$ where x_i is the mid value of the class whose frequency is f_i and N= total frquency $= \sum f_i$

Weighted arithmetic mean

- In computing simple AM, it was assumed that all the items are of equal importance. This may not be always true.
- When items vary in importance they must be assigned weight in proportion to their relative importance. Thus, a weighted mean is the mean of weighted items.
- In calculating weighted A.M. each item is multiplied by its weight and the products so derived are summed up. This total is divided by the total weights (and not by the number of items) to get the weighted mean.
- Symbolically if $x_1, x_2, \ldots, x_n$ are the different items with weights w1, w2, ..., wn respectively then the weighted mean is given by

$$\bar{x}_W = \frac{w_1x_1 + w_2x_2 + \ldots + w_nx_n}{w_1 + w_2 + \ldots + w_n} = \frac{\sum w_ix_i}{\sum w_i}$$

- In fact, AM of a grouped data is a weighted average of the class marks i.e. middle value of the class-interval whose weights are the respective frequency.
- Weighted mean is a much better measure than the simple mean in case when the items in a series are not equally important.

GEOMETRIC MEAN

- The geometric mean is the nth root of product of 'n' items of a series. If $x_1, x_2, \ldots\ldots x_n$ are the 'n' observations in a series then GM is given by:

$$GM = \sqrt[n]{(x_1\, x_2 \ldots x_n)} = (x_1\, x_2 \ldots x_n)^{1/n}$$

- To simplify the above by taking logarithms on both sides,

$$\log GM = \log(x_1\, x_2 \ldots x_n)^{1/n}$$

$$= \frac{1}{n}(\log x_1 + \log x_2 + \ldots + \log x_n)$$

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

$$\therefore\ GM = \text{antilog}\ \frac{\sum \log x_i}{n}$$

- Thus, geometric mean is the antilogarithm of the arithmetic mean of the logarithmic values.
- Logarithm of geometric mean is the arithmetic mean of logarithmic values.
- In the case of frequency distribution (grouped data), GM is given by $GM = (x_1^{f1}\, x_2^{f2} \ldots x_n^{fn})^{1/N}$ where N = total frequency = $\sum f_i$, x_i is the mid point of the class with frequency f_i.
- Simplifying by taking logarithm on both sides,

$$\log GM = \frac{1}{N}\log\left(x_1^{f_1} x_2^{f_2} \ldots x_n^{f_n}\right)$$

$$= \frac{1}{N}\left(f_1 \log x_1 + f_2 \log x_2 + \ldots + f_n \log x_n\right)$$

$$= \frac{\sum f_i \log x_i}{N}$$

$$\therefore GM = \text{antilog}\left[\frac{\sum f_i \log x_i}{N}\right]$$

where x_i is the mid value of the class whose frequency is f_i

HARMONIC MEAN

- Harmonic mean is the total number of items of a variable divided by the sum of the reciprocals of the items. If x_1, x_2, ... x_n are the 'n' observations and HM represents the harmonic mean, then

$$HM = \frac{n}{\left(\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}\right)} = \frac{n}{\sum \frac{1}{x_i}}$$

- Harmonic mean is the reciprocal of arithmetic mean of the reciprocal values.
- In the case of a frequency distribution, HM is obtained by using the formula,

$$HM = \frac{f_1 + f_2 + \ldots + f_n}{\left(\frac{f_1}{x_1} + \frac{f_2}{x_2} + \ldots + \frac{f_n}{x_n}\right)}$$

$$= \frac{N}{\left[\sum \frac{f_i}{x_i}\right]}$$

$$HM = \frac{f_1 + f_2 + \ldots + f_n}{\left(\frac{f_1}{x_1} + \frac{f_2}{x_2} + \ldots + \frac{f_n}{x_n}\right)}$$

$$= \frac{N}{\left[\sum \frac{f_i}{x_i}\right]}$$

where x_i is the mid value of the class whose frequency is f_i, N is the total frequency.

MEDIAN

- It is the value which has got equal number of observations on either side when the items are arranged in the ascending or descending order of magnitude.
- Median divides the series into two equal parts, one part will consist of all variables less than median and the other part greater than median.
- For an ungrouped (or raw) data,
- Median = size of [(n + 1) / 2] th item after arranging the data in the ascending or descending order of magnitude, *when n is odd*.
- Median = mean of size of n/2 th and [(n / 2) + 1] th item after arranging the data in the ascending or descending order of magnitude, *when n is even*.
- For a groupped data(or frequency distribution),

$$\text{Median} = L + \left[\frac{(N/2 - m)}{f}\right] \times c$$

- where L is the lower boundary of the medianal class, N is total frequency, m is cumulative frequency upto the medianal class, c is width of the class interval and f is frequency in the medianal class.

Note

- In the case of frequency distribution, median is the value which has got equal number of frequencies on either side (i.e.) which corresponds to the cumulative frequency of N/2. It is obtained by finding the Medianal class from the less than cumulative frequency distribution (the lower boundary of the medianal class corresponds to the size of the cumulative frequency just less than N/2 and upper boundary of the medianal class corresponding to the size of cumulative frequency just greater than N/2).
- Median can be computed using ogive. It is the x coordinate of the point of intersection of the less than and greater than cumulative frequency curve.

MODE

- It is the size of the most frequent item in a large set of data. Thus mode is the value of that variable which occurs most frequently or repeats itself the greatest number of times.
- In the case of grouped data mode can be calculated by

$$\text{Mode} = L + \frac{(Cf_2)}{f_1 + f_2}$$

where L is the lower boundary of the modal class, f_1, f_2 are the frequencies in the preceding and succeeding modal class and c is the class interval.

Alternative formula

$$\text{Mode} = L + \frac{C(f - f_1)}{2f - f_1 - f_2}$$

where L is the lower limit of the modal class, f_1, f_2 are the frequencies in the preceding and succeeding modal classes respectively and f is the frequency of the modal class.

Note

- Mode can be computed from histogram . It is the x coordinate of the point of intersection of the two diagonals from the top corners of the modal class to the pre and post modal class top corners.
- As a first approximation, mid point of the modal class will be taken as the value of the mode which is called crude mode.
- In a moderately asymmetrical distribution mean-mode = 3 (mean-median), (approximately), Mode = 3 median – 2 mean (approximately). This is empirical mode.
- A distribution can have more than one mode. If it has got one mode, it is called unimodal distribution; if it has got two modes, it is called bi-modal distribution; if it has got three modes, it is called tri-modal distribution; if it has got more than three modes, it is called multi-modal or poly-modal distribution.

PROPERTIES OF AVERAGES

Properties of arithmetic mean

- The sum of the deviations of the items from the mean is equal to zero.
- The sum of the squared deviations from the mean is smaller than the sum of the squared deviations of the items from any other value

$$\text{i.e} \sum (x_i - \bar{x})^2 \text{ is minimum}$$

- The product of the mean with the number of observations gives the total of the original data,

$$\text{i.e.} \sum x_i = n\bar{x}$$

- If $\bar{x}_1, \bar{x}_2$ are the means of the two groups with the number of observations, n_1 and n_2 respectively the mean of the combined group $\bar{x}$ is given by

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

- AM > GM > HM
- When all the values are equal, AM = GM = HM
    - For a symmetrical distribution, AM = median = mode
    - For a positively skewed distribution, AM > median > mode (short tail on the left)

- For a negatively skewed distribution AM < median < mode (short tail on the right)

Properties of geometric mean

- GM will be zero if one or more of the values are zero.
- GM < AM, GM > HM

Properties of harmonic mean

- HM < GM < AM

Properties of Median

- Mean deviation taken about median as the origin is the minimum.
- If the distribution is symmetrical, median = mean = mode,
- Median < mean, and median > mode, if the distribution is positively skewed.
- Median > mean, and median < mode, if the distribution is negatively skewed.

Properties of Mode

- If the distribution is symmetrical, mode = median = mean.
- If the distribution is positively skewed, mode < median < mean,
- If the distribution is negatively skewed, mode > median > mean
- If the distribution is moderately asymmetrical then, Mode = 3 median – 2 mean (approximately).

SITUATIONS WHERE DIFFERENT AVERAGES ARE USED

- AM is generally applicable for all sorts of data. It should be used when the distribution is reasonably symmetrical and further statistical analysis is to be carried out such as the computation of the standard deviation etc. and also algebraic manipulation is to be followed subsequently.
- GM is used when it is desired to give more weights to small items and less weight to large items and in the case of ratios, percentages and microorganisms.
- HM is used in averaging certain types of ratios and rates and problems involving time. It gives more weight to small items.
- The median is to be used when the attribute of the data are not directly measurable. As it can be easily located by mere inspection, it can be calculated when the data are incomplete. Use the median when the distribution is highly skewed and the extreme items may have distorting effects on the mean.
- Mode can be used to know the most typical value or the most common item. It is also used when the quickest estimate of centrality is required.

Click here for overview...

MODULE-5: MEASURES OF DISPERSION

Learning objective

Learner of this module will come to know that averages alone cannot describe the distribution and have to measure how the different observations scatter about an average. One can estimate the consistency of two or more distributions after learning this lesson.

MEANING OF DISPERSION

- The measures of central tendency indicate only the central position. But they have their own limitations and do not throw light on the formation of the series of data.
- Sometimes they may offer misleading results too. *For eg*. Consider the following three series.
    - Series A 50,50,50,50
    - Series B 50,56,48,42,43,59,52
    - Series C 1,29,120
- They have the same mean 50. Hence we may conclude that these series are alike in nature. But a close examination shall reveal that the distributions differ widely from one another.
- In one distribution, the values may be closely packed and in the other, they may be widely scattered. Such a variation is called *scatter, spread or dispersion*.
- Hence an average is more meaningful when it is examined in the light of dispersion.
- When dispersion is not significant then the average appears to be a true representative figure of the series and when dispersion is significant, it implies that the average is far from being a true representative figure.
- The measurement of the scattering of item in a distribution about the average is called a measure of variation or dispersion.
- Measures of dispersion also enable comparison of two or more distributions with regard to their variability or consistency

Objectives of measures of dispersion

- To determine the reliability of an average
- To serve as a basis for control of the variability
- To compare two or more series with regard to their variability

Differrent measures of dispersion

- Range
- Quartile Deviation(Q.D)
- Mean Deviation(M.D)
- Standard Deviation (S.D)

RANGE

- It is the difference between the highest(H) and lowest(L) values in the raw data.
- For the grouped data, the range is the difference between the lower limit(L) of the first class and the upper limit(H) of the last class. It is given by

$$\text{Range}=H-L$$

- It is a very simple measure of dispersion.
- It is useful in the study of variation in money rate and rates of exchange, weather forecast etc.
- Relative measure of dispersion for range is the ratio of range (R.R) which is given by

R.R = (H-L) / (H+L)

QUARTILE DEVIATION

- It is also known as semi-inter quartile range. It is based on quartiles which are points which divide the data into four equal parts.
- The lower or first quartile (Q_1) divides the lower half of the distribution into two equal parts, i.e., it is the value below which 25% of the observation lie and above which 75% of the observations lie.
- Similarly, the upper or third quartile (Q_3) divides the upper half of the distribution into two equal parts, i.e. it is the value below which 75% of the observations lie and above which 25% of the observations lie.
- The difference, $Q_3 - Q_1$ is called inter quartile range and QD is given by $(Q_3 - Q_1) / 2$
- For grouped data, Q_1 is the value which corresponds to the cumulative frequency of N/4 and Q_3 is the value which corresponds to the cumulative frequency of 3N/4.
- Quartile Deviation is used in the case of open end distribution.

Formula to compute QD

- In the case of raw data, after arranging the data in the ascending order

$$Q_1 = \text{Size of the } \left[\frac{(n+1)}{4}\right]^{th} \text{ term}$$

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

then,

$$QD = \frac{(Q_3 - Q_1)}{2}$$

- In the case of frequency distribution or grouped data

$$Q_1 = L_1 + \frac{(N/4 - m_1) \times c}{f_1}$$

where L_1 is the lower boundary of the first quartile class, m_1 is the cumulative frequency upto the first quartile class, f_1 is the frequency in the first quartile class and c is the width of the class interval

$$Q_3 = L_3 + \frac{(3N/4 - m_3) \times c}{f_3}$$

where L_3 is the lower boundary of the third quartile class, m_3 is the cumulative frequency upto third quartile class, f_3 is frequency in the third quartile class and c is the width of the class interval

- Then,

$$QD = \frac{Q_3 - Q_1}{2}$$

- Relative measure of QD is known as the quartile co-efficient of dispersion (QC).

$$QC \text{ of dispersion} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

MEAN DEVIATION

- Mean deviation or average deviation in a series is the AM of the deviations of the various items from an average (mean, median or mode) of the series taking all deviations as positive.

For raw data,

$$\text{MD about mean} = \frac{\sum |x_i - \bar{x}|}{n}$$

$$\text{MD about an average A} = \frac{\sum |x_i - A|}{N}$$

where A is mean or median or mode

For grouped data,

$$\text{MD about mean} = \frac{\sum f_i |x_i - \bar{x}|}{N}$$

$$\text{MD about an average A} = \frac{\sum f_i |x_i - A|}{N}$$

where A is mean or median or mode and x_i is mid point of the i^{th} class with frequency f_i

- The relative measure of Mean Deviation is known as mean coefficient of dispersion or coefficient of mean deviation and is obtained by dividing the MD by the average from which it is computed

$$\text{Coefficient of MD about an average} = \frac{[\text{MD about an average A}]}{A}$$

$$\text{Coefficient of MD about an average A in \%} = \frac{[\text{MD about an average A}]}{A} \times 100$$

STANDARD DEVIATION

- It is the most perfect and widely used measure of dispersion. It is an improved method over MD.
- It is the root mean square of the deviations measured from the mean.
- In other words, SD is the positive square root of the AM of the square of the deviation of items taken from AM of the series.
- It is denoted by `σ' (read as 'sigma').

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}} \quad \text{for raw data}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i\left(x_i - \bar{x}\right)^2}{N}} \quad \text{for grouped data}$$

Where x_i is the mid value of the class interval whose frequency is f_i and N is the total frequency.

- Simplifying the above,

$$\sigma = \sqrt{\frac{\sum x_i^2}{n} - \left\{\frac{\sum x_i}{n}\right\}^2} \quad \text{for raw data}$$

$$\sigma = \sqrt{\frac{\sum f_i x_i^2}{N} - \left\{\frac{\sum f_i x_i}{N}\right\}^2} \quad \text{for grouped data}$$

- Shortcut method,

$$\text{Put } d_i = \frac{x_i - A}{c},$$

$$\text{then, } \sigma = c \times \sqrt{\frac{\sum f_i d_i^2}{N} - \left\{\frac{\sum f_i d_i}{N}\right\}^2}$$

SHEPPARD'S CORRECTION

- In computing the standard deviation, sometimes grouping error may occur on account of grouping of data into different classes. For statistical adjustment of this grouping error, Sheppard has suggested a correction value to be deducted from the variance of the grouped data, which is given by

$$\frac{C^2}{12} = \frac{\text{Square of width of class interval}}{12}$$

$$\text{Hence, Corrected S.D} = \sqrt{\frac{\sum f_i X_i^2}{N} - \left\{\frac{\sum f_i X_i}{N}\right\}^2 - \frac{C^2}{12}}$$

Coefficient of Variation

- Relative measure of standard deviation is known as coefficient of variation (CV or COV) and is defined as SD / Mean.

$$\text{i.e., CV} = \frac{\sigma}{\bar{x}} \times 100$$

- Thus, CV is the percentage variation from the mean, with SD being treated as the total variation.
- Higher CV indicates greater variability and less CV implies better consistency of data.

VARIANCE

- Square of standard deviation is called as variance. It is the mean square deviation.
- It is the sum of the squared deviation of individual observations from the mean divided by the number of observations. It is denoted by σ^2.

Standard Error (SE)

- The mean of random sample may be taken as a representative of the population mean.
- The difference between the sample mean and population mean is due to sampling and it is called sampling error or standard error.
- It is defined as the SD of the mean of different samples, taken from the population.
- If we study only one sample, then

$$SE\left(\bar{X}\right) = \frac{SD}{\sqrt{n}}$$

where ‘n’ is the size of the sample and SD is that of the sample.

Probable Error(PE)

- PE = 2/3 SD (approx.)

Properties of QD

- QD = (2/3) SD = PE (approx.)
- Mean ± QD will cover 50% of the cases.

Properties of MD

- MD = (4/5) SD (approx.)
- MD about median as the origin is the minimum.

PROPERTIES OF SD

- SD is greater than MD, QD and PE
- Mean square deviation will be minimum, if the deviation is taken from AM as the origin.
- Mean ± 1 SD will cover 68.27% of the items
- Mean ± 2 SD will cover 95.45% of the items
- Mean ± 3 SD will cover 99.73% of the items
- By adding or subtracting a constant from all the observations, SD is unaltered.
- If $\bar{x}_1, \bar{x}_2$ are the means of two samples of sizes n_1, n_2 respectively with SD σ_1, σ_2 then the combined SD (σ) is given by

$$(n_1+n_2)\sigma^2 = n_1\sigma_1^2 + n_2\sigma_2^2 + n_1d_1^2 + n_2d_2^2$$

where $\bar{x}$ is the combined mean and

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

MERITS AND DEMERITS OF MEASURES OF DISPERSION

The essential requisites of a good measure of dispersion are the same as that of averages.

easure of ispersion	Merits	Demerits
ge	It is easy to calculate and simple to understand.	It is not rigidly defined. It is not based on all observations. affected much by extreme values and open end classes. It d not possess sampling stability. It is not amenable for furthe mathematical treatment.
rtile iation	It is easy to calculate and simple to understand and is not affected by extreme items and open end classes.	It is not rigidly defined. It is not based on all observations. does not possess sampling stability. It is not amenable for further mathematical treatment.
n Deviation	It is based on all observations, it is rigidly defined and easy to calculate and simple to understand.	It is affected much by extreme values and open end classes. does not possess sampling stability. It is not amenable for further mathematical treatment.
ndard iation	It has a rigid formula and is based on all observations. It is capable of further algebraic treatment. It is less affected by	It is affected much by extreme values and open end classes.

	sampling	

Click here for overview...

MODULE-6: MOMENTS, SKEWNESS AND KURTOSIS

Learning objective

The learner of this module will have an idea about the shape of the curve of the distribution.

MOMENTS

- The r^{th} moment of 'n' observations about their mean $\bar{x}$ (called as central moments), usually denoted by μ_r(read as Mu r)is defined as

$$\mu_r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^r}{N} \text{ for raw data}$$

where $x_1, x_2, \ldots\ldots x_n$ are the 'n' observations and

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ for } r = 0, 1, 2, \ldots.$$

- In the case of frequency of distribution,

$$\mu_r = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^r}{N}$$

where N is the total frequency $x_1, x_2, \ldots\ldots x_n$ are the mid value of the classes whose frequencies are $f_1, f_2 \ldots\ldots f_n$

Note

$$\mu_o = \frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})^o}{N} = \frac{\sum_{i=1}^{n} f_i}{N} = \frac{N}{N} = 1$$

$$\mu_1 = \frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})^1}{N} = 0$$

$$\mu_2 = \frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})^2}{N} = \sigma^2$$

- In case of symmetrical distribution, μ_3, μ_5, μ_7..... will be zero; $\mu_r = 0$ for all odd values of r.

Moments about arbitary point A

- The r^{th} moment of 'n' observation about any point A(called as raw moments) is denoted by μ_r' is defined as

$$\mu_r' = \frac{\sum_{i=1}^{n} (x_i - A)^r}{n} \quad \text{for a raw data}$$

where $x_1, x_2 \ldots x_n$ are 'n' observations

$$\mu_r' = \frac{\sum_{i=1}^{n} f_i(x_i - A)^r}{N} \quad \text{for a frequency distribution}$$

where $x_1, x, \ldots x_n$ are the mid value of the classes whose frequencies are $f_1, f_2 \ldots f_n$

Note

1. when r = 0

$$\mu_o' = \frac{\sum_{i=1}^{n} f_i(x_i - A)^o}{N} = \frac{\sum_{i=1}^{n} f_i}{N} = \frac{N}{N} = 1$$

2. When r = 1

$$\mu_1' = \frac{\sum_{i=1}^{n} f_i(x_i - A)^1}{N} = \frac{\sum_{i=1}^{n} f_i x_i - \sum_{i=1}^{n} A}{N} = \frac{N\bar{x} - NA}{N} = \bar{x} - A$$

$$\bar{x} = A + \mu_1'$$

3. When r=2

$$\mu_2' = \frac{\sum_{i=1}^{n} f_i(x_i - \bar{A})^2}{N} \text{ and so on}$$

Relation between μ_r and μ_r'

$$\mu_r = \frac{1}{N}\sum f\left(x-\bar{x}\right)^r$$
$$= \frac{1}{N}\sum f\left(x-A+A-\bar{x}\right)^r$$
$$= \frac{1}{N}\sum f\left(x-A-\mu_1'\right)^r$$

Using binominal expansion

$$\mu_r = \frac{1}{N}\sum f\left((x-A)^r - rc_1(x-A)^{r-1}\mu_1' + rc_2(x-A)^{r-2}\mu_1'^2 - rc_3(x-A)^{r-3}\mu_1'^3 + \ldots + (-1)^r\mu_1'^r\right)$$
$$= \frac{1}{N}\sum f\left[(x-A)^r - rc_1\mu_1'(x-A)^{r-1} + rc_2(x-A)^{r-2}\mu_1'^2 - rc_3(x-A)^{r-3}\mu_1'^3 + \ldots + (-1)^r\mu_1'^r\right]$$
$$\mu_r' = \frac{\sum f(x-A)^r}{N} - rc_1\frac{\sum f(x-A)^{r-1}}{N}\mu_1' + rc_2\frac{\sum f(x-A)^{r-2}}{N}\mu_1'^2 \ldots\ldots$$
$$\mu_r = \mu_r' - rc_1\mu_{r-1}'\mu_1'^1 + rc_2\mu_{r-2}'\mu_1'^2 - rc_3\mu_{r-3}'\mu_1'^3 \ldots\ldots$$

SKEWNESS

- It is seen that the measures of central tendency indicate the central position or central tendency of the frequency distribution and the measures of dispersion give an indication to the extent to which the items cluster around or scatter away from the central tendency. But none of these measures indicate the form or type of the distribution.

Consider the following two distributions

Class	Frequency
0-5	10
5-10	30
10-15	60
15-20	60

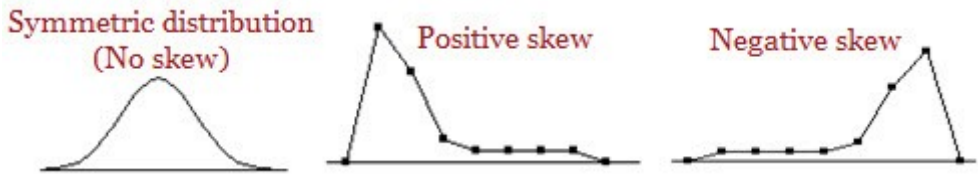
Class	Frequency
0-5	10
5-10	40
10-15	30
15-20	90

20-25	30
25-30	10

20-25	20
25-30	10

The above two distributions have the same mean (=15) and SD (= 6), but yet they are not identical distribution.

- The distribution on the left hand side is symmetrical one, whereas the distribution on the right hand side is asymmetrical or skewed.
- Skewness refers to the symmetry or departure from symmetry.
- Symmetry means that the number of values above the mode and below the mode is same in a data.
- When a distribution departs from the symmetrical form, it is said to be asymmetrical or skewed.
- Evidently, in the case of symmetrical distribution, the two tails of the curve are of equal size and in the case of asymmetrical distribution, one tail of the curve is longer than the other.
- A distribution is said to be skewed in the direction of the excess tail. Thus if the right tail is longer than the left, the distribution is positively skewed; if the left tail is longer than the right, the distribution is negatively skewed.



MEASURES OF SKEWNESS

- Pearsonian measure of skewness
- Bowley's measure of skewness
- Measure of skewness based on moments

Pearsonian measure of skewness

- It is given by, mean – mode or 3 (mean - median) and the coefficient of skewness

$$\frac{\text{mean - mode}}{\text{SD}} \text{ or } \frac{3(\text{mean - median})}{\text{SD}}$$

Bowley's measure of skewness

- It is given by $Q_3 - M - (M-Q_1)$. i.e. $Q_3 + Q_1 - 2M$ and the coefficient of skewness is

$$\frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$$

Measure of skewness based on moments

It is denoted by β_1 and is given by,

$$\beta_1 = \frac{\mu_3^{\,2}}{\mu_2^{\,3}}$$

where μ_r r^{th} moment about $\bar{x}$ is given by,

$$\mu_r = \frac{\sum (x_i - \bar{x})^r}{n} \text{ for raw data}$$

$$\mu_r = \frac{\sum f_i (x_i - \bar{x})^r}{N} \text{ for grouped data}$$

Note: where μ_2 is the variance of the data

KURTOSIS

- It is the measure of the peakedness.
- It indicates the degree of flatness or peakedness in the region or area relating to the mode of frequency curve.
- A normal curve is symmetrical and bell shaped.
- When two distributions are compared, the top of the frequency curve may be flat, very narrow or peaked. This characteristic of frequency curve is known as kurtosis or peakedness.
- The normal curve is known as mesokurtic.
- A curve more peaked than normal curve is leptokurtic and the curve which is flatter than the normal curve is platykurtic.

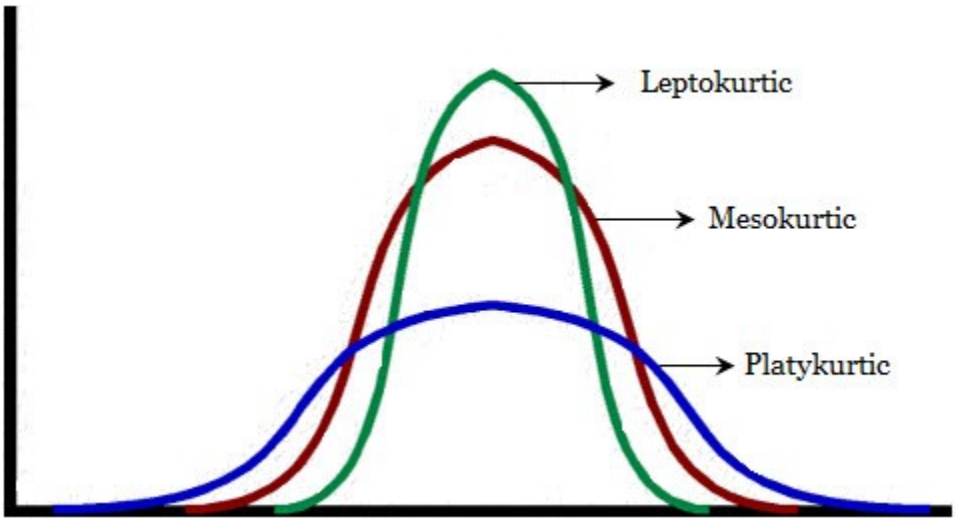
Measure of Kurtosis

- It is measured by $\beta_2 - 3$ where, $\beta_2 = \mu_4 / \mu_2^{\,2}$ where μ_4 is the 4th moment about about $\bar{x}$ is given by,

$$\mu_4 = \frac{\sum (x_i - \bar{x})^4}{n} \text{ for raw data}$$

$$\mu_4 = \frac{\sum f_i (x_i - \bar{x})^4}{N} \text{ for grouped data}$$

- $\beta_2 = 3$ for a normal curve
- $\beta_2 < 3$ for platykurtic curve
- $\beta_2 > 3$ for leptokurtic curve



Click here for Graph...

MODULE-7: PROBABILITY

Learning objective

Reader of this module will know about the concept of Probability and to measure it.

MEASURE OF PROBABILITY

- Probability is a measure of chance.
- If A is a desired event and 'r' is the number of occurrences of A out of the total number of possible occurrences say 'n', then the probability of occurrence of A, denoted by P(A), is given by:

$$P(A) = (\text{No. of occurrence of A}) / (\text{Total no. of occurrences}) = r / n$$

- Probability is a ratio taking values from 0 to1. It can never be negative.
- If an event is an impossible event, then the probability is 0.
- If an event is certain to occur, then its probability is 1.

Definitions

- Event
  - Any phenomena occurring in nature
- Equally likely events
  - All events have equal chance of occurrence
- Mutually exclusive events
  - If the occurrence of an event completely avoids the occurrence of another event, then the events are said to be mutually exclusive.
- Independent events
  - If the occurrence of one event in no way affects the occurrence of another event, then these events are said to be independent .
- Simple and compound event

    - When two or more events occur together, their happening is described as a compound event, while if only one event takes place at a time, it is called as a simple event.
- Exhaustive cases
    - Refer to all possible outcomes without any omissions
- Sample space
    - Represent the set of all possible outcomes of a phenomenon

MATHEMATICAL AND EMPIRICAL PROBABILITY

- Mathematical or "*a priori*" probability is one, which can be determined deductively without any experimentation or trial.
- It is based on the assumption that one has full confidence of an event happening out of several possible alternatives (even before the event happens), which are mutually exclusive and equally likely.
- It assumes that all cases are equally likely, i.e. equally probable all the time.
- However, for many kinds of chances and events in real life especially in social and economic aspects, this definition is unsuitable, as one cannot declare *a priori* that all cases are equally likely.
- Thus, without assuming any of certainty of the event happening, if one tries to base the probability of an event on past experience of certain outcomes based upon a long series of experiments (that is, on the basis of statistical data), then the probability is known as statistical or empirical probability.
- If 'n' is the number of cases observed which is not only a large number but also increases indefinitely up to infinity which sets a limit to the probability of the event happening and when an event A is found to be occurring in 'm' number of cases, the ratio (m / n) is close to P(A), the statistical probability. In other words,

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

THEOREM OF TOTAL PROBABILITY OR ADDITION THEOREM

- The probability of occurrence of one or other of a set of mutually exclusive events is the sum of the probabilities of occurrence of the separate events of the set.

Proof

- Let 'n' be the exhaustive mutually exclusive ways in which all events can occur.
- Let the first event occur in a_1 of these ways, the second in a_2 of these ways ... and k^{th} event occur in a_k of these ways.
- Then, if $p_1, p_2, ..., p_k$ be the probabilities of the occurrence of these events, $p_1 = a_1 / n$; $p_2 = a_2 / n$; $p_3 = a_3 / n$, ..., $p_k = a_k / n$.
- Since these events are mutually exclusive, the number of ways in which one or other of k events will occur is:

$$= \frac{a_1 + a_2 + \ldots + a_k}{n}$$

$$= \frac{a_1}{n} + \frac{a_2}{n} + \ldots + \frac{a_k}{n}$$

$$= p_1 + p_2 + \ldots + p_k$$

= Sum of the independent probability of the events

- Suppose if we take two mutually exclusive events A and B, then the probability of any one of the events A or B which are mutually exclusive shall happen, is the sum of the probability of occurrence of A and the probability of occurrence of B

$$P(A \text{ or } B) = P(A) + P(B)$$

MULTIPLICATION OR COMPOUND PROBABILITY THEOREM

- The probability of simultaneous occurrence of a set of independent events is the product of the separate probabilities of those independent events.
- If A and B are two independent events and their individual probabilities are P(A) and P(B), then the probability of joint occurrences of A and B, i.e the probability of the compound event A and B, is the product of their respective probabilities.
- Suppose if the event A occurs in n_1 ways of which m_1 ways are successful and the event B occurs in n_2 ways of which m_2 ways are successful, then

$$P(A) = \frac{m_1}{n_1} \text{ and } P(B) = \frac{m_2}{n_2}$$

- The total outcome of the events $= n_1 n_2$;
- The total no. of successful outcomes for A and B = $m_1 m_2$. Therefore

$$P(AB) = \frac{m_1 m_2}{n_1 n_2} = \frac{m_1}{n_1} \times \frac{m_2}{n_2} = P(A) \times P(B)$$

PROBABILITY OF COMPOUND EVENTS

- If p is the probability of the event, the probability that it will occur in exactly x out of n cases is $nC_x q^{n-x} p^x$.

Example

- An animal Scientist wishes to choose three animals for his research from 8 sheep, three of which are Suffolk, and the other are Mecheri. What is the probability that a randomly selected group will have two Mecheri and one Suffolk sheep?

The total possible group of 3 $= {}^8C_3$

= 8 x 7 x 6/1 x 2 x 3

= 56

We can select two Mecheri out of five in $^5C_2$ ways and we can select one Suffolk in $^3C_1$ ways.

Therefore, the total possible ways of selecting 2 Mecheri and 1 Suffolk = $^5C_2$ x $^3C_1$ = 10 x 3 = 30. Probability of getting two Mecheri and one Suffolk is = 30 / 56 = 0.54.

- An animal scientist works with three different cattle breeds to improve their efficiency of beef production. Suppose he has 10 Steers, 5 of which are Angus, 3 are Brahman, and 2 are Hereford and he selects three animals at random. What is the probability that all the three selected are Angus?

The total number of ways of getting 3 animals = $^{10}C_3$
= 10 x 9 x 8/1 x 2 x 3

= 120

No. of ways of getting three Angus is = $^5C_3$ = 10

The probability of getting three Angus = 10 / 120 = 0.0833

- If we assume that male and female calves are equally likely, what is the probability of exactly three out of 10 calves born will be male calves?

'p' is probability of success and 'q' is the probability of failure.

Then, probability of getting 'x' success in 'n' trial = $nc_x . q^{n-x} . p^x$

In the given case, n = 10, p = q = ½; x = 3

Let the event of getting a male calf be the success, then we have to find the probability, x = 3

probability = $10C_3 (½)^{10-3} (½)^3$

= $10C_3 (½)^{10}$
= (10 x 9 x 8 /1 x 2 x 3) x $1/2^{10}$

= 120/1024

= 0.1171

MODULE-8: PROBABILITY DISTRIBUTIONS

Learning objective

The learner of this module will know about few important probability distributions and their properties

PROBABILITY DISTRIBUTIONS - DEFINITION

- The probability distribution shows how the set of all possible mutually exclusive events is distributed.
- The probability distribution can be regarded as the theoretical equivalent of an empirical relative frequency distribution, with its own mean and variance.
- A probability distribution comprises all the values that the random variable can take, with their associated probabilities.

BINOMIAL DISTRIBUTION

- It is a probability distribution expressing the probability of one set of alternatives, i.e. success or failure. It is developed under the following assumptions:
  - An experiment is performed under the same condition for a fixed number of trials, say 'n'
  - In each such trial, there are only two possible chances of the experiment, success or failure
  - The probability of success denoted by 'p' which remains constant from trial to trial and the probability of failure denoted by q=1-p
  - The trials are independent
- This distribution was discovered by James Bernoulli and hence it is also called Bernoullian distribution.
- In a series of n independent trails, if p is the constant probability of success at a single trial, then the variate 'x ', i.e., the number of success at these 'n' trails is said to follow binomial distribution. The variate takes values from 0 to n (all integers), the probability of getting 0, 1, 2, ..., n successes at these n trails is q^n,$nC_1q^{n-1}p$,$nC_2q^{n-2}p^2$,..., $nC_xq^{n-x}p^x$, ..., p^n respectively, which are the respective terms of binomial expansion $(q+p)^n$.
- Suppose if we have N sets of 'n' trials, the number of sets in which we will have 0, 1, 2, ..., n success will be given by the successive terms of binomial expansion $N(q+p)^n$. Thus, we classify the sets according to the number of successes which they contain and we get a frequency distribution which is known as the binomial distribution.

No. of success	0	1	2	...	x	...	n
Frequency	Nq^n	$NnC_1q^{n-1}p$	$NnC_2q^{n-2}p^2$		$NnC_xq^{n-x}p^x$		Np^n

Properties of Binomial Distribution

- It is a distribution of discontinuous or discrete variate.
- It has two parameters (constants).They are n and p, where n denotes the number of independent trials and p denotes the constant probability of success at a single trial.
- It takes values from 0 to n (all integers), i.e. 0, 1,2,..., n
- It's mean is np and variance is npq, where q=1-p, SD = $\sqrt{(}$npq) ; Variance is always less than mean .
- The different frequencies are different terms of binomial expansion $N(q+p)^n$
- It is symmetrical, when p=q=½
- When n is large and p is small such that np is constant, the binomial distribution tends to a Poisson distribution .
- When n is large and p=q=½, the binomial tends to become a normal distribution .

POISSON DISTRIBUTION

- It is a discrete probability distribution and is limiting form of binomial distribution, when n is large and p is small such that np is constant.

- It is a distribution of rare events. It is also called as the law of improbable events. The equation is:

$$P(x) = \frac{Ne^{-m}.m^{x}}{x!}$$

where m is the mean, N is the total frequency and x is the no. success

Properties of Poisson distribution

- It has one parameter m
- It is a distribution of discrete variate
- It takes values from 0 to ∞
- Mean is approximately equal to variance
- Skewness is 1/ m
- Kurtosis is 3 + [1/ m]

NORMAL DISTRIBUTION

- Binomial and Poisson distributions are the more useful theoretical distributions for discrete variables. That is, they relate to the occurrence of distinct events.
- In order to have mathematical distribution suitable for dealing with quantities whose magnitude is a continuous variable, a continuous distribution is needed.
- The normal distribution is the most useful theoretical distribution for continuous variable.
- It was discovered by De Movire as the limiting form of Binomial distribution and was also known to Laplace.
- Gauss is the first one who made reference to this and it was erratically named after him as Gaussian distribution.
- The frequency curve corresponding to normal distribution is normal frequency curve or normal curve or Gaussian or Laplacian or probability curve.

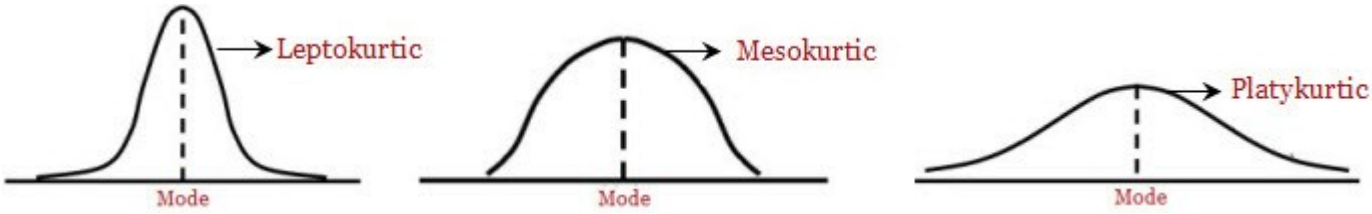
$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{\frac{-(x-m)^2}{2\sigma^2}}$$

Where m = mean, N = total frequency σ^2 = variance

Properties of Normal Distribution

- It is a distribution of continuous variates
- The variate takes values from −∞ to +∞
- It is a symmetrical distribution, mean=median=mode
- The slope of the curve is bell shaped. The ends of the curve tails off asymptotically to the base
- It has two parameters m and σ
- The first and the third quartiles are equi-distant from the median
- MD = 4/5 σ or 0.7979 σ
    - mean ± 2/3 σ covers 50% of the observation
    - mean ± σ covers 68.27 % of the observation
    - mean ± 2σ covers 95.45% of the observation

- mean ± 3σ covers 99.73% of the observation
- All odd moments about mean = 0
- Skewness is zero
- Kurtosis is 3. It is mesokurtic.
- A frequency curve is leptokurtic if kurtosis > 3
- A frequency curve is platykurtic if kurtosis < 3
- If mean =0; SD=1 then the normal distribution is a standard normal distribution



MODULE-9: CORRELATION

Learning objective

A study of this module will make one to know the definition of correlation and the methods of studying it and its uses.

DEFINITION OF CORRELATION

- Correlation is the strength of relationship or the intensity of association between two variables.
- If two variables vary in such a way that as one increases (or decreases), the other also increases (or decreases), then the correlation is said to be positive or direct. (eg.) feed intake and growth rate of animals.
- If two variables vary in such a way that as one increases, the other decreases and vice versa, then the correlation is said to be indirect or negative (eg.) litter size and birth weight of piglets.
- If there is no relationship between the two variables, they are said to be independent or uncorrelated.

Coefficient of correlation

- A measure of correlation free from units of measurements is called coefficient of correlation.
- It is denoted by 'r', r takes values from –1 to +1.
- When r = +1, the correlation is perfect and positive, r = -1, the correlation is perfect and negative, r = 0, there is no correlation

TYPES OF CORRELATION

- Simple, partial and multiple
- Linear and non-linear

Simple, partial and multiple

- It is based on the number of variables studied. When only two variables are involved in the study of correlation, it is called as the *simple correlation*. (eg) (i.) feed in take and growth of animals, (ii). birth weight and number of piglets.
- When more than two variables are involved in the study, it is either multiple or partial correlation. In multiple correlation, we study the correlation of one dependant variable over all the other independent variables, (eg.) milk yield vs. first lactation period, food supplied, age etc.
- In *partial correlation* we study the relationship between two variables, assuming that the other variables are constant. (eg) correlation between the weight of broiler and feed intake assuming the other factors like area provided, labour used, medicinal cost etc. as constant.

Linear and Non-linear

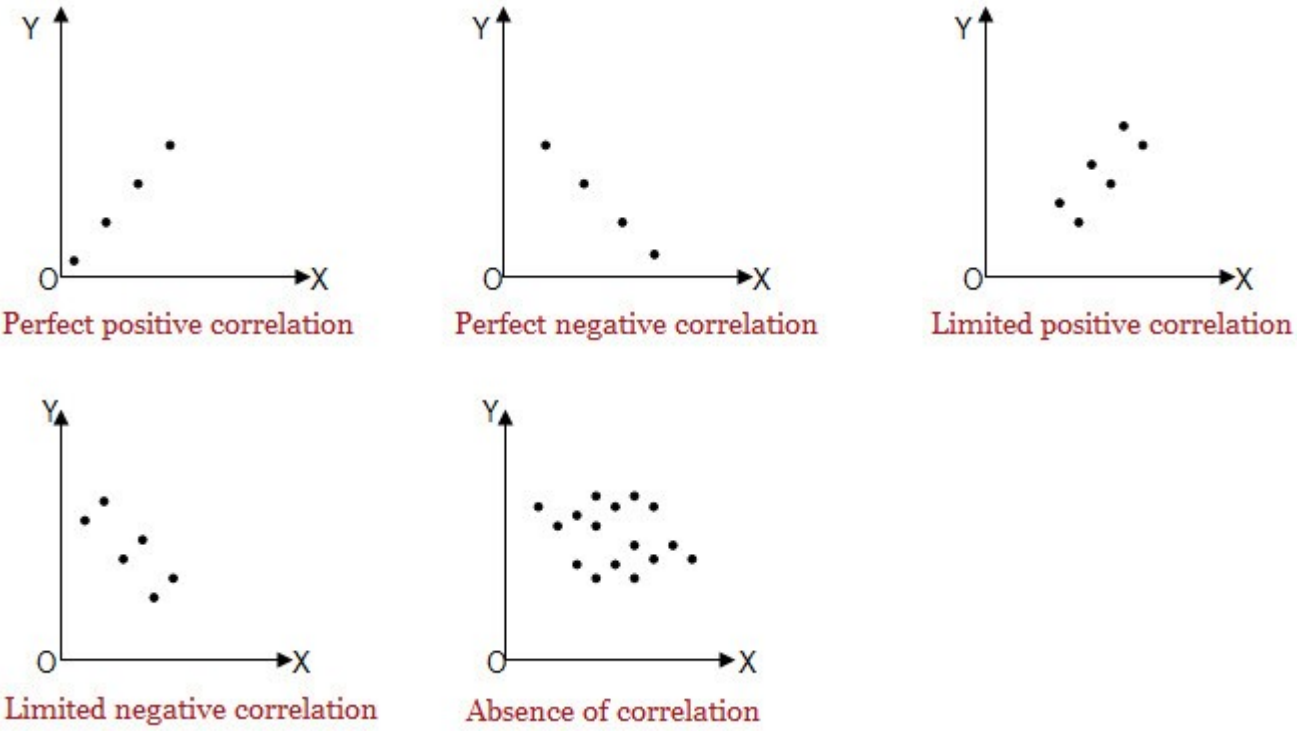
- On the basis of degree of covariation, correlation may be termed as linear or non-linear.
- When the correlation between two sets of variables is perfect of degree one or unity which means that the two variables have an exact functional relationship, the correlation is said to be *linear*.
- The graphical depiction of a linear correlation presents a straight line and its functional relationship is represented by the relation, y = a + bx, where 'a' and 'b' are constants.
- A perfect linear correlation may be positive or negative. Thus, its numerical coefficient will be either +1 or −1. These are the limits of correlation.
- Thus, coefficient of correlation cannot be greater than +1 or less than −1. If the correlation is imperfect, its graphic exposition will be *non-linear*.
- It will not form a straight line. Non-linear correlation will always be less than unity and it will lie between -1 and +1.

METHODS OF STUDYING CORRELATION

- Scatter diagram
- Correlation graph
- Karl pearson's coefficient of correlation
- Concurrent deviation method.
- Rank method

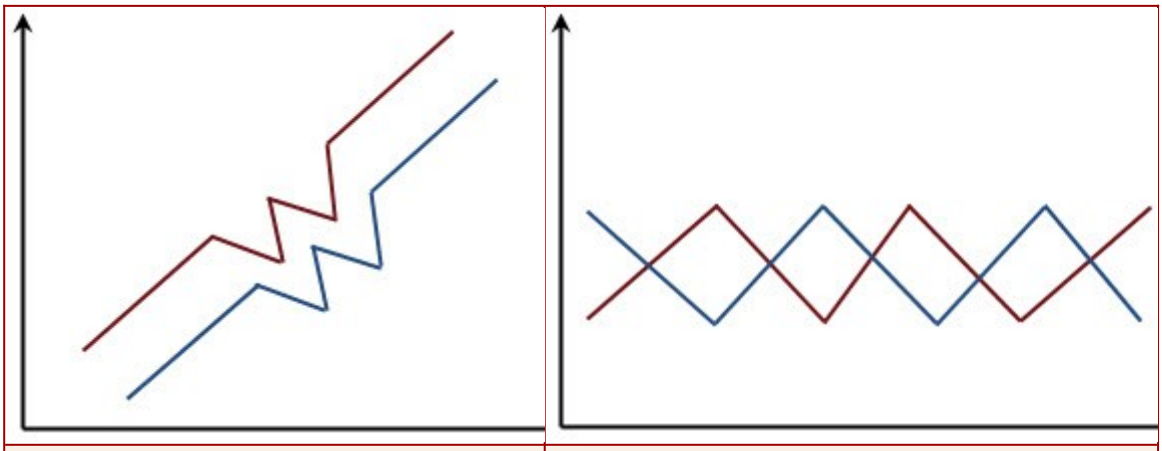
SCATTER DIAGRAM

- A scatter diagram or scattergram or scatterplot or dot diagram is a chart prepared to represent graphically the relationship between two variables.
- Take one variable on the horizontal and another on the vertical axis and mark points corresponding to each pair of the given observations after taking suitable scale.Then, the figure which contains the collection of dots or points is called a scatter diagram.
- The way in which the dot lies on the scatter diagram shows the type of correlation.
- If these dots show some trend either upward or downward,then the two variables are correlated. If the dots do not show any trend, there is absence of correlation between the two variables.



CORRELATION GRAPH

- In this method, curves are plotted for the data on two variables. By examining the direction and closeness of the two curves so drawn, we can infer whether or not the variables are related.



If both the curves drawn on the graph are moving in the same direction (either upward or downward), correlation is said to be positive	*On the other hand, if the curves are moving in the opposite direction, correlation is said to be negative*

- This method is normally used for time series data. However, like scatter diagram, this method also does not offer any numerical value for coefficient of correlation.

KARL PEARSON'S COEFFICIENT OF CORRELATION

- Of the several mathematical methods of measuring correlation, the Karl Pearson's method, popularly known as Pearsonian coefficient of correlation is often used.
- It is denoted by r. It is also called product moment formula. It is given by

$$r = \frac{\frac{\sum xy}{n} - \bar{x}\bar{y}}{\sigma_x \sigma_y}$$

Last modified: Monday, 30 April 2012, 02:33 PM

CONCURRENT DEVIATION METHOD

This method of studying correlation is the simplest of all the methods. What is to be found in this method is the direction of change of x and y variables.

- The stepwise procedure is:

Step i:

- Find out the direction of change of x variable, i.e as compared with the first value, whether the second value is increasing or decreasing or constant. If it is increasing, put a + sign, if it is decreasing, put a – sign and if it is constant, put zero. Similarly, as compared to second value, find out whether the third value is increasing, decreasing or constant. Repeat the same process for the other values also. Denote the column as Dx.

Step ii:

- In the same way, find out the direction of change of y variable and denote this column as Dy.

Step iii:

- Multiply Dx with Dy and determine the value of c, the number of concurrent deviations or the number of positive signs obtained after multiplying Dx with Dy.

Step iv:

- Then apply the formula

$$r = \pm\sqrt{\pm\left[\frac{2c - n}{n}\right]}$$ sign is taken as that of $(2c - n)$

RANK METHOD

- Sometimes we may not know the actual values, but their ranking may be known. In such occasions, this method would be of use. Even when the actual values are available, we can rank them and measure correlation using the formula:

$$r_s = 1 - (6 \sum d^2) / n(n^2 - 1)$$

where 'd' is the difference in the ranking of the two series x and y, and n is the number of paired observations.

- When some of the values are equal, give the average rank and then add 1/12 ($m^3 - m$) to $\sum d^2$, where m stands for the number of items whose values are equal.
- When the number of paired observations exceed 30, it is very difficult to rank them and hence, unless rank is given, it is better to avoid this method. This method is also called as *Spearman's rank correlation coefficient* as it was due to the statistician Spearman.

PROPERTIES OF CORRELATION COEFFICIENT

- Correlation coefficient lies between –1 and +1
- Correlation coefficient is independent of change of scale and origin of the variable x and y
- Correlation coefficient is the geometric mean of the two regression coefficients † (in magnitude).

Note: † discussed in the next chapter

Coefficient of determination

- Square of correlation coefficient (r^2) is defined as coefficient of determination. If $r^2 = 0.85$, it implies that 85% of the variation in the dependent variable is due to the independent variable studied.

Coefficient of non-determination

- $(1 - r^2)$ is defined as the coefficient of non-determination. If $1-r^2 = 0.24$, we infer that 24% of the variation in the dependent variable is due to the other variables which are not studied. This is called as unexplained variations.

MODULE-10: REGRESSION

- Learning objective
- Reader of this module will understand the meaning of regression and its uses.

REGRESSION

- Regression is the amount of dependence of one variable on the other. This gives the rate of change of one variable with respect to another.
- The meaning of regression is the act of returning or going back.
- This term was introduced by Francis Galton, when he studied the relationship between the heights of father and sons.
- We may be interested in estimating the value of one variable given the value of another. This is done with the help of regression.
- The problem of regression is to find out equations of the lines (or curves) with functional relationship between two variables, independent - x and dependent - y.
- The simple linear regression is of the form $y = a + bx$, where 'b' represents the slope of the line (also called as regression coefficient) and 'a' the intercept of the line.

- In a study where data on age and weight of animals are available, age could be considered as the independent variable, while weight as the dependent variable.
- It means that weight regresses on age.
- In the regression line y = a + bx, a and b are calculated by

$$b = \frac{\frac{\sum xy}{n} - \bar{x}\bar{y}}{\sigma x^2}$$

$$a = \bar{y} - b\bar{x}$$

- b is called the regression coefficient of the line y on x denoted by ' b_{yx}'
- The above is the regression line of y and x. Similarly we have the regression equation of x on y which can be given by:

$$x = c + dy,$$

- where d is the regression coefficient of the regression line x on y ,denoted by ' b_{xy}' which is given by

$$d = \frac{\frac{\sum xy}{n} - \bar{x}\bar{y}}{\sigma_y^2} \quad \text{and} \quad c = \bar{x} - d\bar{y}$$

PROPERTIES OF REGRESSION COEFFICIENT

- Regression coefficient gives the rate of change in the value of dependent variable (y) for a unit change in the independent variable (x).
- Regression equation is useful in estimating the values of y by knowing the values of x.
- Both the regression coefficients will be of the same sign.
- If one of the regression coefficients is greater than one, the other will be less than one.
- Product of the regression coefficients is the square of correlation coefficient.

$$r = +\sqrt{bd} \text{ if b and d are positive.}$$

$$r = -\sqrt{bd} \text{ if b and d are negative.}$$

Exercise

- Following are the data on milk yield (in kg/day) (y) in a lactation and number of days (x) after calving:

Day (x)	10	14	18	22	26	30	34
Milk yield (y)	1.78	1.66	1.62	1.59	1.55	1.60	1.58

- Fit a regression equation of the form y = a + bx and estimate the milk yield on 15th day and 23rd day.

MODULE-11: THEORY OF SAMPLING

- Learning objective
- After going through this module, one will realise the need of sampling. The reader will know to estimate sampling error and various sampling procedures.

INTRODUCTION

In practice, any required information may be obtained by following either census method or sampling method of data collection.

- Census method
  - It refers to the complete enumeration of the data from each and every unit of population or universe, i.e. total set of items in a particular investigation.
- Sampling method
  - It is the process of learning about the population on the basis of the samples drawn from the population. i.e., a part of the universe or population selected from the population for the purpose of investigation.
- Parameter
  - It is the statistical measure pertaining to the population, e.g., mean of the population, SD of the population, etc, while statistic is the statistical measure pertaining to the sample. e.g., mean of the sample, SD of the sample, etc.

NEED, ADVANTAGE AND DISADVANTAGE OF SAMPLING

Need

- In many situations, census method, in spite of its advantages, is not possible especially when population is large or infinite. As census method proves to be costly and time consuming, sampling method of data collection appears appropriate.

Advantage

- As only a part of the population is studied, this method saves time.
- It saves cost as the amount of labour and expenses involved will be less for the sample.
- Since only a few and standard number of observations are involved, it provides results with greater accuracy.
- It provides more detailed information as it deals with only few observations.
- Sometimes sampling is the only method available. When the population is infinite or if articles are to be destroyed for testing the quality, census method can not be carried out.

Disadvantages

- What we get from that sample is only the estimate of population parameter which may differ from the actual value of the parameter, yet it is possible to calculate the sampling error or standard error which is given along with the estimate.

STANDARD ERROR / SAMPLING ERROR

- It is the difference between parameter and statistic

- It is the standard deviation of the sample means. If only one sample is studied, then standard error is SD/$\sqrt{\mathbf{n}}$ where n = size of sample.
- *Estimate:* is the value of population parameter obtained from the sample.
- *Estimator:* is the one which is used to estimate the value of the population parameter.

There are two types of estimate:

- *Point estimate:* is a single value which is used to estimate the population parameter
- *Interval estimate:* is an interval in which population parameter lies between. It is also called fiducial limit or confidence interval.

Notations

Particulars	Sample	Population
Observations	$y_1, y_2, y_3, \ldots$	$Y_1, Y_2, Y_3 \ldots$
Size	n	N
Mean	$\bar{y}$	$\bar{Y}$
Standard deviation	s	S or σ

USES OF STANDARD ERROR

- Standard error of the mean indicates the average variations in sample means from the population mean. Standard error is used as an instrument or basis for testing the hypothesis.
- The magnitude of standard error gives us an idea about the reliability of the sample. The greater the standard error is greater the departure of actual value from the expected value and hence greater the unreliability (lesser the reliability) of sample. The reciprocal of standard error is taken as a measure of reliability or precision of sample. That is,

$$\text{Precision} = 1/\text{S.E and S.E} = \frac{SD}{\sqrt{n}}$$

- As n increases, SE decreases and hence precision increases. In large samples, sampling distribution of statistics approximates normal distribution
- Standard error helps us to determine the limits, within which the population parameters are expected to lie.

LIVESTOCK CENSUS

- The livestock sector is important to national economy and considering the short span of re-productivity and life of domestic animal, the first Census was organized during 1919-1920 and since then it is being conducted quinquennially(once in five years) by all the States/UTs in India. So far, 1 8 such Censuses have been conducted and the latest one is the 18th Livestock Census in the series with 15/10/2007 as the date of reference. 17th Livestock Census was due to be conducted in 2002 but got delayed and was conducted with reference date 15th October, . As per 17th

livestock census (2003), livestock population in our country is 485 millions and that of poultry is 489 millions. India ranks first in respect of buffalo, second in cattle and goats, third in sheep, fourth in ducks, fifth in chickens and sixth in camel population in the world. The country has 57% of the world's buffalo population The 18th livestock census has been conducted throughout the country and final draft is awaited. 18th Livestock census for India and Tamil Nadu is given below.

17th Livestock census(2003)

Livestock	World*	India
Cattle	1371.1	185.2
Indigenous		160.5
Crossbred		24.7
Buffalo	170.7	97.9
Sheep	1024.0	61.5
Goat	767.9	124.4
Pigs	956.0	13.5
Horses and Ponies	55.1	0.75
Donkeys		0.65
Camels	19.1	0.632
Yaks		0.065
Mithun		0.278
Poultry	49877	489.0

Source: Dahd.nic.in

SAMPLING PROCEDURES

Random sampling or probability sampling

- Unrestricted random sampling or simple random sampling
- Restricted random sampling
  - Stratified sampling
  - Systematic random sampling
  - Cluster sampling
  - Multistage sampling.

Non random sampling or non probability sampling

- Judgment sampling
- Convenient sampling
- Quota sampling

SIMPLE RANDOM SAMPLING (SRS)

- Simple randon sample(SRS) is the technique of drawing sample from the population, in such a way that each unit of population has an equal and independent chance of being included in the sample. If N is the population size and n is sample size, the probability that any one of N units in population to be considered as a unit in the sample is 1/N.

SRS can be done by any one of the following two methods

- Lottery method
- Method using random number table

Lottery Method

- This is a very popular method of selecting random samples.
- In this method, all the items of universe are numbered or named on separate slips of papers of identical shape and size.
- These slips are then folded in the same manner and are put in a container. After shuffling, one slip after another is taken till the required sample size is obtained.
- The number in the slips would constitute the sample and thus the selection of items depends entirely on chance.

Method Using Random Number Table

The lottery method is quite cumbersome when the size of the population is large. The alternative method of random selection is using table of random numbers. The following are some of the tables of random numbers available:

- Tippets random numbers table consisting of 41,600 random units grouped into 10,400 sets of 4 digits random numbers
- Fisher and Yates random number table consisting of 15,000 random units assigned into 1,500 sets of 2 digits random numbers
- Kendall and Smith random numbers table having 10 lakh random digits grouped into 25,000 sets of 4 digits random numbers
- Rand Corporation table of random numbers having of 1 lakh random number digits grouped into 20,000 sets of 5 digited random numbers
- Rao, Mitra and Matthai table of random numbers consisting of 20,000 random digits grouped into 5,000 sets of 5 digits random numbers

Example of selecting samples through random numbers

- If the population size is 30 (N = 30) and sample size is 15 (n = 15), then, number the population from 1 to 30. Since 30 is a two digit number, choose a two digit random number table.
- In the two digit random number table, choose 01 to 90 rejecting 91 - 99 and 00 so as to give equal chance to all units.
- Have a random start in random number table.
- Start any where, select that number if the number is less than 30; if the number is greater than 30, divide it by 30 and take the reminder; if the reminder is 0, that corresponds to 30. The process has to be repeated till we get 15 different numbers (i.e. required sample size).

ESTIMATION OF POPULATION MEAN AND CONFIDENCE INTERVALS

- If $y_1, y_2, \ldots, y_n$ are the sample values chosen from the population of size N, then the mean of the sample

$$\bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n}$$

$$\text{i.e. } \bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$$

- Sample mean will be taken as estimate of population mean

$$\hat{\bar{Y}} = \bar{y}$$

Confidence interval

- For this, we have to find out standard error of population mean. Estimation of standard error of population mean,

$$\text{i.e., S.E.}\left(\hat{\bar{y}}\right)$$

$$\text{S.E.}\left(\hat{\bar{Y}}\right) = \sqrt{\left(\frac{N-n}{N}\right)\frac{s^2}{n}} \text{ where } s^2 \text{ is variance of sample}$$

- When n > 30, 95% confidence interval for population mean $\bar{y}$ is given by

$$\bar{y} \pm 1.96 \text{ S.E.}\left(\hat{\bar{Y}}\right)$$

$$\bar{y} \pm 1.96 \sqrt{\left(\frac{N-n}{n}\right)\frac{s^2}{n}}$$

- 99% confidence interval is

$$\bar{y} \pm 2.58 \sqrt{\left(\frac{N-n}{n}\right)\frac{s^2}{n}}$$

- For small samples, if n<30, then 95% and 99% confidence interval for $\bar{Y}$

$$\bar{y} \pm t_{(n-1)df}(5\%)S.E\left(\hat{\bar{Y}}\right) \text{ and}$$

$$\bar{y} \pm t_{(n-1)df}(1\%)S.E\left(\hat{\bar{Y}}\right) \text{ respectively}$$

STRATIFIED RANDOM SAMPLING (StRS)

- In this method, the population of size 'N' is subdivided into a definite number of non overlapping and distinct sub population of sizes $N_1, N_2, \ldots, N_k$ such that $N_1 + N_2 + N_3 + \ldots + N_k = N$. This procedure of dividing the population into distinct sub populations is called stratification and each sub population is called as a stratum.
- While forming a stratum, we see that the units within each stratum are more homogenous with respect to the character under study. Between strata, there should be greater diversity or variability. This is done with the idea of improving the precision of estimate.
- Within each stratum of size 'Ni', a simple random sample of n_i will be drawn such that $n_1 + n_2 + \ldots\ldots + n_k = n$, while 'n' is the size of the sample. Such a sampling method is called as the stratified random sampling. The 'n' sampling units chosen ar the strtified random sample(St. RS)

Situations where St.RS could be used

- When the character under investigation is highly variable
- When estimates are required for subpopulations also

Notations

Particulars	Sample	Population
Characteristic value	$y_{11}, y_{12} \ldots y_{ij}$, j^{th} unit in i^{th} stratum	$Y_{11}, Y_{12} \ldots Y_{ij}$, j^{th} unit in i^{th} stratum
Mean	$y_1, y_2 \ldots$	$Y_1, Y_2 \ldots$

Advantages of Stratified Random Sampling

- It is more representative than SRS, as there is a possibility for all sample units drawn from a single corner of the population. However in St.RS, all stratums are equally represented and thereby avoid the possibility of any essential group of population being completely eliminated from sample.
- St.RS provides more precise population estimate, besides estimates for each of the stratum.
- As compared with SRS, the St.RS would be concentrated geographically, accordingly, time and money involved in collecting the data and interviewing the individuals will be considerably reduced and supervision of field work could be done with greater efficacy and convenience.

SYSTEMATIC SAMPLING

- It is a commonly employed technique, if the complete and upto date list of sampling units are available.
- This consists of selecting only the 1^{st} unit at random, the rest being selected according to some predetermined pattern involving regular spacing of units.

- Let us suppose that 'N' population units are serially numbered from 1 to N in some order and a sample of size 'n' is to be chosen, then let k = N/n and this k is called sampling interval or sampling ratio.
- Choose a number (p) at random, which is less than k, then choose every k^{th} item, i.e., p, p+k, p+ 2k,... This random number p is called random start.

Merits and Demerits of Systematic Sampling

Merits

- It is more convenient than simple random sampling and stratified random sampling as time and work involved is relatively less.
- Systematic sampling will be more efficient than simple random sampling, provided the list from which sample units are drawn is known.

Demerits

- Actual sample size will be different from that of required size.
- Sample mean is not an unbiased estimate of population mean.
- It may lead highly biased estimates, if there are periodic features associated with sampling interval. i.e., if the list has a periodic features and sampling interval is equal to or multiple of that period.

CLUSTER AND MULTISTAGE SAMPLING

Cluster sample

- In this case, total population is divided depending on the problem under study, into some recognisable subdivisions named as clusters and simple random sample of these clusters is drawn.
- Each and every unit in the selected clusters is observed as sample.
- For example, If we are interested in obtaining income of a city, whole city is divided into N different blocks or localities and simple random sample of 'n' blocks is drawn.
- The individuals in selected blocks form the cluster sample.

Multistage sample

- Instead of enumerating all sample units in the selected clusters, one can obtain better and more efficient estimators by resorting to sub sampling at different stages within the clusters. This technique is called second stage sampling, third stage sampling and so on ,clusters being termed as first stage sampling unit.
- The entire above procedure is called multistage sampling.
- For example, if we want to study consumption pattern of households in Tamil Nadu.
- In the first stage, a simple random sample of few districts will be selected.
- Selected districts then divided into villages and from which a simple random sample of villages is selected, which is called second stage sampling units.
- In the third stage of sampling, the selected villages will be further divided into households and a simple random sample of households will be selected as final sample units.

NON-RANDOM SAMPLING METHOD

Purposive or deliberate or subjective or judgment sampling

- It is the one in which the investigator takes the samples exclusively at his discretion.

Convenient sampling

- If the investigator chooses the samples at his convenience, it is called convenient sampling.

Quota sampling

- It is a type of judgment sampling wherein quotas are setup according to some specified characteristics such as 'this much in this group', 'this much in other group' and so on.

MODULE-12: TESTS OF SIGNIFICANCE : INTRODUCTION AND 'Z' TEST

- Learning objective
- After going through this module,the reader will understand the purpose of test of significance and various tests of significance and will know the all large sample tests.

INTRODUCTION

- Test of significance is a statistical procedure followed to test the significance of the difference between statistics and the parameter or between any two statistics. i.e. between sample mean and population mean or between two sample means.

Hypothesis

- Any statement made about the population

Null hypothesis

- It is the hypothesis under test or initial hypothesis proposed .
- It may or may not be true.
- It is usually denoted by H_o. Null hypothesis is never proved.
- It is either accepted or rejected at some level of significance.

Levels of significance

- Usually we will have two levels of significance. 5% and1% level of significance.
- 5% level of significance means that, if this experment is repeated under identical conditions 100 times,then the chance for this conclusion to go wrong is five out of 100 .
- 1% level of significance means that, if this experment is repeated under identical conditions 100 times,then the chance for this conclusion to go wrong is one out of 100 .

Alternate hypothesis

- Statement contrary to null hypothesis is alternate hypothesis and is denoted by H_1.

Degrees of freedom(d.f)

- The number of observations which are free to move or free to vary.

DIFFERENT TESTS OF SIGNIFICANCE

Parametric test

- A distribution will be attached to the test. The various parametric tests are
  - Normal deviate test or large sample test or Z test
  - Students t test or small sample test
  - Chi-square test
  - Variance – ratio test or F-test

Non- parametric test

- It will be free from distribution and it is called distribution free method.
  In any test, we take any four type of decisions:
  - Null hypothesis may be true but we reject it by our test which is Type I error.
  - Null hypothesis may be false but we accept it which is Type II error.
  - Null hypothesis is true and we accept it which is correct decision.
  - Null hypothesis is false and we reject it which is correct decision.

STATISTICAL PROCEDURE FOLLOWED IN ANY TEST FOR SIGNIFICANCE

- *Step 1:* Formation of null hypothesis
- *Step 2:* Calculation of test statistics

Test statistics = (Statistics – Parameter) / SE of difference
= (Difference in the value of two statistics) /SE of difference

- *Step 3:* Depending on the value of test statistics, we take decisions different from test to test.

LARGE SAMPLE TESTS OR Z TESTS

- Z test is carried out, when sample size is > 30. When n > 30, it will follow normal distribution whose equation is

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{\frac{-(x-m)^2}{2\sigma^2}}$$

where, m = mean; σ = standard deviation

To test the significance of the difference between sample mean and population mean

- *Step 1: H_o:* No significant difference between sample mean and population mean
- *Step 2:* Test statistics or Z statistics is given by,

$$Z = \frac{\text{Sample mean - Population mean}}{\text{SE of difference}}$$

$$= \frac{\bar{x} - m}{SE(\bar{x} - m)}$$

Where x = mean of sample; m = mean of population

$$Z = \frac{\bar{x} - m}{\frac{\sigma}{\sqrt{n}}}$$

Where σ = SD of population; n = size of sample

$$Z = \frac{\bar{x} - m}{\frac{s}{\sqrt{n}}}$$

When 's' is not known replace it by which 's' is SD of sample.

Then,

- *Step 3 :* Conclusion
  - | Z | < 1.96, we say that Z is not significant and Ho is accepted. We denote this by Z = ( )^{NS}.
  - | Z | > 1.96, Z is significant and Ho is rejected. We denote this by Z = ( )*.
  - | Z | > 2.58,Z is highly significant and Ho is rejected. We denote this as Z = ( )**.

Exercise

- A sample of 300 broilers is taken from a farm and their mean weight was found to be 1.9 kgs with Standard Deviation 0.24 kgs. Verify whether the sample could have been taken from a population of mean weight 2.23 kgs.

TEST THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN TWO SAMPLE MEANS

Case -1 : When they are taken from different populations

- Step-1 Ho: There is no significant difference between two sample means.
- Step-2: The test Statistics is given by

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{SE(\bar{x}_1 - \bar{x}_2)}$$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Where $\bar{x}_1, \bar{x}_2$ are sample means with sizes n_1, n_2 respectively taken from population with standard deviations σ_1, σ_2

- If SD of the populations are not known, replace it by the sample SD

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where $\bar{x}_1, \bar{x}_2$ are mean of sample of sizes n_1, n_2 with SD of s_1, s_2 respectively

- Step-3: Conclusion - As in the previous test

Exercise

- The mean milk yield during a lactation in two farms are estimated as 2176 kgs and 2,425 kgs with SD 56.2 kgs and 43.2 kgs by taking a sample of sizes 50 and 60 respectively. Test whether the two farms differ significantly in their lactation yield.

TEST THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN TWO SAMPLE MEAN

Case-2: When they are taken from same population

- Step-1: Ho: There is no significant difference between the two sample means
- Step-2: The test statistics is given by

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_2} + \frac{s_2^2}{n_1}}}$$

Where σ (SD of Population) is not known,

$$\text{replace } \sigma^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$$

- Step-3: Conclusion: as in the previous test

Exercise

- Two samples of sizes 35 and 48 are taken from a broiler farm and their mean weight (in kgs) were worked out as 1.8 and 2.0 respectively with SD 0.72 and 0.56. Test the significance of the two samples.

DIFFERENCE BETWEEN SAMPLE PROPORTION WITH POPULATION PROPORTION

To test significance of the difference between sample proportion with population proportion

It 'n' is the number of trials, and 'p' is the proportion of success out of 'n' trials, and P is the expected proportion of success, then p follows normal distribution when 'n' is large.

- Step-1:Ho: There is no significant difference between p and P.
- Step-2:

$$\text{Test Statistic } Z = \frac{p - P}{SE(p - P)} = \frac{p - P}{\sqrt{\frac{pq}{n}}} \text{ where } q = 1 - p$$

- Step-3: Conclusion: as in the previous test

Exercise

- In a farm 120 calves were born in a year out of which 73 are female. Test the hypothesis that sexes are born in equal proportion.

TO TEST THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN TWO PROPORTIONS

To test the significance of the difference between two proportions

- Ho: There is no significant difference between the two proportions
- Let p_1, p_2 are observed proportions of success out of n_1, n_2 trails respectively,then the test statitics is

$$z = \frac{p_1 - p_2}{SE(p_1 - p_2)} \quad \text{or}$$

$$= \frac{p_1 - p_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \quad \text{where } q_1 = 1 - p_1, q_2 = 1 - p_2$$

$$= \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{where } p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \text{ and } Q = 1 - P$$

TO TEST THE SIGNIFICANCE OF AN OBSERVED CORRELATION COEFFICIENT

To test the significance of an observed correlation coefficient

- Let 'r' be the sample correlation coefficient and ' ρ ' be the population correlation coefficient.
- Ho: There is no significant difference between correlation coefficients of the sample and population.

$$z = \frac{r - \rho}{SE(r - \rho)} = \frac{r - \rho}{\frac{1 - \rho^2}{\sqrt{n}}}$$

- Conclusion: as in the previous test.

Note

- In this case, we are not interested about r being significantly different from ρ , but we are interested whether there is a significant correlation between the variables studied. i.e. whether r is significantly different from o (zero).
- Hence, in the above formula, put ρ = o

$$= \frac{r - o}{(1 - o)/\sqrt{n}} = r\sqrt{n}$$

- If z is not significant, then we conclude that r is not significant i.e. there is no correlation between the variables studied.
- We denote this as r=( )$^{N.S}$.
- It z is significant, then there is significant correlation.
- Then, we denote this as r = ( ).*
- If Z is highly significant, then we say that r is highly significant and r = ( )**.
- In both cases, both variables studied are related.

Exercise

- The correlation coefficient between first lactation yield and second lactation yield of 75 Tharparkar cows was found to be 0.5819. Test the significance of correlation in the sample.

MODULE-13: TESTS OF SIGNIFICANCE : 'T' TEST

Learning objective

Through this module, the reader will know all the small sample tests and whether a sample chosen by him is a good representative of the population and whether two samples are homogeneous.

SMALL SAMPLE TESTS

- When the sample size is large and variates are normally distributed, normal test is employed to test the significance of differences.
- With small samples and when the degrees of freedom less than 30, the variates are not normally distributed.
- Hence we make use of the 't' distribution and 't' test of significance.
- Tables have been constructed relating to 't' at different probability levels for various degrees of freedom.

TO TEST THE SIGNIFICANCE OF THE DIFFERENCE OF THE SAMPLE MEAN FROM THE POPULATION MEAN

To test the significance of the difference of the sample mean from the population mean

- Step 1: Ho: There is no significant difference between the sample mean and population mean
- Step 2: Test statistic is given by

$$t = \frac{\text{Difference in the mean of the sample and population}}{\text{S.E of difference}}$$

$$t = \frac{\bar{x} - m}{S.E(\bar{x} - m)}$$

$$= \frac{\bar{x} - m}{\frac{s}{\sqrt{n}}} \text{ with d.f} = (n-1)$$

$\bar{x}$ = mean of the sample of size 'n' with SD 's' with d. f= (n-1)

m = population mean

- Step 3: Conclusion
  - If | t | < table value of t for (n-1) d.f. at 5% level t is non-significant and H_0 is accepted. We denote this as t = ( ) N.S.
  - If | t | > table value of t' for (n-1) d.f. at 5% level 't' is significant. Ho is rejected, we denote this as t = ( )*.

- If | t | > table value of 't' for (n-1) d.f. at 1 % level t is highly significant. Ho is rejected. We denote this as t = ( ) **

Exercise

- The average fleece weight of a breed of sheep was given as 1700 gm per sheep. From this breed 20 sheep were selected at random and the average weight of fleece per sheep of this sample was found to be 1625 gm with SD 38.75 gms. Find whether the sample mean agrees with the population mean.

TEST OF SIGNIFICANCE OF THE DIFFERENCE BETWEEN TWO SAMPLE MEANS

- To test the significance of the difference between two sample means when the sizes are less than 30 and they are dependent (Paired 't' test)
- Two samples are dependent when they have some common factor linking the observations in the two samples.
  - Step 1 : Ho: There is no significant difference between the two sample means
  - Step 2 : Test statistic is given by

$$t = \frac{\bar{d} - o}{SE \text{ of } \bar{d}} \text{ with d.f} = (n-1)$$

where d = Difference in the observation of the two samples. s = SD of d

$$= \frac{\bar{d}}{\frac{s}{\sqrt{n}}} \text{ with d.f} = (n-1)$$

  - Conclusion: as in the previous test

Exercise

- The first clip wool yield of daughters of the ram and their dams (in 100 gm) are

Daughter (x)	19	14	21	18	21	17	17	15	8	20
Dam (y)	9	17	14	16	15	18	9	8	10	11

- Test the significant difference between the daughter and the dam.

NON PAIRED OR UNPAIRED 't' TEST

- To test the significance of the difference between the means of two samples, when the samples are independent (Non-Paired or Un-paired 't' test)
- By independence of the two samples, we mean that there is no relationship between individuals contributing the sample. Thus, the sample drawn from different populations or different parts of the same population will be independent.

- Step 1: Ho: There is no significant difference between the sample means.
- Step 2: Test statistics 't' is given by

$$\text{i.e., } t = \frac{\text{Difference in the mean of the sample}}{\text{S.E of difference}}$$

$$= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\text{with d.f} = n_1 + n_2 - 2 \text{ where } s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$\bar{x}_1, \bar{x}_2$ are the means of the two samples with SD s_1, s_2 of sizes n_1, n_2 respectively

- Conclusion: as in the previous test (with respect to d.f.)

Exercise

- Following are the data on first lactation yield of Sindhi and Sahiwal. Test whether there is any significant difference between these two breeds.

Sindhi	66	47	54	64	56	66	51	28	55	56	29
Sahiwal	68	53	53	72	54	63	41	33			

TO TEST THE SIGNIFICANCE OF AN OBSERVED CORRELATION COEFFICIENT

- To test the significance of an observed correlation coefficient
  - Step 1: Ho: There is no significant correlation
  - Step 2: Test statistic is given by

$$t = \frac{r - o}{\text{S.E. of } (r - o)}$$

$$= \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

$$= r\frac{\sqrt{n - 2}}{\sqrt{1 - r^2}} \text{ with d.f.} = n - 2$$

- The coefficient of correlation between the first lactation yield and second lactation yield of 25 Tharparkar cows from a large population is 0.48. Is this significant of correlation in the population.

N.B

- Table A11 page 557 of statistical methods by Snedecor and Cochran gives the critical values of correlation coefficient that are significant at 5% and 1% level. If 'n' is the no. of observations, we have to see the table for degrees of freedom at (n-2).

MODULE-14: TESTS OF SIGNIFICANCE : CHI-SQUARE TEST

- Learning objective
- The reader will know through this module, Chi-square tests for significance and whether the observed values are according to theory and whether any two factors of classification are independent or not.

CHI-SQUARE TEST FOR SIGNIFICANCE

- This test is used in the case when observed frequencies are to be tested for their fit with expected or theoretical frequencies or to test whether two factors of classification of a set of individuals presented in the form of two- way table are independent or not.

Chi-square test for goodness of fit

- This test is performed to test whether the deviation of observed frequencies in a given data from the expected frequencies are due to real causes or due to chance.
- In other words, this is a test to decide whether the observed frequencies are in accordance with the frequencies within statistical limits. This test is used to decide whether the given data has a good fit with one of the known forms of distributions, viz., Normal, Binomial or Poisson.
- This test is also used to test the observed number of progenies in a genetic experiment to fit in Mendalian laws of heredity.
- The test statistic for goodness of fit is given by

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

O - observed frequency, E - Expected frequency, df = No. of classes -1

In this case,

- Ho : The fit is good or there is no significant difference between observed and expected frequency.

Conclusion

- If the calculated χ^2 is less than table χ^2 for the respective degrees of freedom at 5% level. χ^2 is not significant which is denoted by $\chi^2 = (\ )^{N.S}$.i.e., Ho is accepted. The fit is good.
- If calculated χ^2 is > table χ^2 for the respective d.f. at 5% level χ^2 is significant and denoted by $\chi^2 = (\ )^{*}$. H_o is rejected. The fit is not good or theoretical frequencies are not according to theory.
- If calculated χ^2 is > table χ^2 for the respective df at 1% level χ^2 is highly significant and denoted by $\chi^2 = (\ )^{**}$. H_o is rejected. The fit is not good or theoretical frequencies are not according to theory.

Exercise

- In a farm 250 calves were born during a particular time. The number of male calves is 135. Test whether the sexes are equally born.
- The following table gives the classification of 400 plants according to the nature of leaves and flower colour.

Blue flower and flat leaves	234
Blue flower and crimpled leaf	63
White flower and flat leaf	76
White flower and crimpled leaf	27

- Test whether the frequencies are in the ratio 9:3:3:1.

CHI-SQUARE TEST FOR INDEPENDENCE

- This is performed when the data is presented in the form of a contingency table.
- A table giving the simultaneous classification of the body of data in two different ways is called a “contingency table” If there are ‘r’ rows and ‘ c’ columns, the table is said to be an "r xc contingency table”. χ^2 test is applied to test whether the factors classified are independent or not. i.e. the two factors are associated or not.
- In the contingency table both factors may be qualitative or one qualitative and the other quantitative or both quantitative.
- The degrees of freedom for r x c contingency table is (r-1) (c-1).
- Application of χ^2 statistics in a 2 x 2, χ^2 contingency table

Factor A

Factor B		Level – I	Level – II	Total
	Level I	a	b	a + b
	Level II	c	d	c + d
	Total	a+c	b+d	a+b+c+d=n

- Step 1: H_0: The factors are independent
- Step 2 : χ^2 is given by

$$\chi^2 = \frac{(ab - bc)^2 \times n}{(a+b)(c+d)(a+c)(b+d)}$$

with df =(2-1) (2-1) = 1

- Conclusion: as in the case of χ^2 test for goodness of fit (with the respective d.f).
- The above formula is applicable when all a, b, c, d are greater than 5. If one or more is less than 5, Yate's correction of continuity is to be applied which is as follows:

$$\chi^2 = \frac{(|ad - bc| - n/2)^2 \times n}{(a+b)(c+d)(a+c)(b+d)}$$

- In a r x c contingency table, the expected value (E) in the ith row and jth column is calculated by

$$\frac{R_i \times C_j}{N}$$

where

  - Ri = Sum of all the values in the i^{th} row
  - Cj = Sum of all the values in the j^{th} column
  - N = Grand total i.e., the sum of all the values in the given contingency table. Then,

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- Conclusion: as in the previous test (with the respective df).

EXERCISE

- 1. The following table relating to the survival to one breeding farm during 5 years. Test the association of the mortality rate and seasons.

Season of birth	*Survived*	*Died*	*Total*
Summer	378	82	460
Rainy	215	16	231
Winter	204	87	291
Total	*797*	*185*	*982*

- 2. Following is the vaccine test results. Find whether the inoculation of vaccine can be regarded as effective against the disease.

	Survived	*Dead*	*Total*
Inoculated	10	12	22
Non-inoculated	14	13	27

- 3. In an experiment on the immunisation of goats from Anthrax, the following results were obtained. Derive your inference on the effectiveness of vaccine.

	Survived	*Died*	*Total*
Inoculated	15	2	17
Noninoculated	4	10	14
Total	19	12	31

MODULE-15: TESTS OF SIGNIFICANCE : 'F' TEST

Learning objective

After reading this module, the reader will know about the 'F' test and it's uses.

'F' TEST OR VARIANCE-RATIO TEST

This is the test to compare the variance of two samples

- H_0 : There is no significant difference between the variances of two samples
- The 'F' Statistics is given by

$$F = \frac{\text{Estimated Variance of first sample}}{\text{Estimated Variance of second sample}} = \frac{s_1^2}{s_2^2} \text{ with d.f} = (n_1 - 1), (n_2 - 1)$$

S_1^2 and S_2^2 are the variance of the sample of sizes n_1, n_2 respectively.

F has 2 degrees of freedom, one for higher variance and another for smaller variance.

N.B

- We have to put the greatest of variance (S_1^2 , S_2^2)in the numerator

Conclusion

- If cal | F | < tab F for d.f. = (n_1-1), (n_2-1) at 5% level, F is not significant ,denoted by F=( ) $^{N.S}$. H_o is accepted.
- If cal | F | > tab F for d.f. = (n_1-1), (n_2-1) at 5% level, F is significant,denoted by F = ( ) *. Ho is rejected
- If cal | F | > tab F for d.f. = (n_1-1), (n_2-1) at 1% level, F is highly significant,denoted by F = ( ) **. Ho is rejected.

Exercise

- Test whether the variances are significantly different in the following sample.

Sample I (X1)	45	46	49	25	17	18	13	56	58
Sample II (X2)	47	49	43	27	29	38	37		

MODULE-16: DESIGN OF EXPERIMENTS

Learning objective

Through this module, the reader will know what is design of experiment and the principles used in it and different designs of experiment. One can compare two or more than two groups.

DESIGN OF EXPERIMENTS

Introduction

- Designing an experiment means planning an experiment so that information collected will be relevant to the problem under investigation.
- The Design of an experiment is the complete sequence of steps taken before experimenting to ensure that the appropriate data will be obtained in a way that furnish an objective analysis leading to valid inferences with respect to the state and problem.
- The purpose of any experimental design is to provide maximum information relevant to the problem under investigation

DEFINITION

Treatment

- What we apply on the subject of investigation is called treatment. e.g. Application of feed to animals ,application of fertilizer to agricultural plot, etc.

Experimental material and experimental unit

- The individual or group of individuals that will be subjected to a treatment is called the experimental unit and the collection of such units will be experimental material.

Response

- Outcome of an experiment. i.e. the treatment effect available from the experimental units.

Experimental error

- It is the unit-to-unit variation within the same treatment group. This is a measure of variation due to uncontrollable causes. It describes the failure of two identically treated experimental units to yield identical results.

BASIC PRINCIPLES OF EXPERIMENTAL DESIGNS

- Randomization
- Replication
- Local Control

Randomization

- It is a device for eliminating bias. i.e. by randomly assigning treatments to the experimental units we avoid personal bias. It involves giving equal chances to all the experimental units to be subjected to different treatments. It can be done by the use of random number table or by drawing lots. The randomization procedure is different for different designs.

Purposes served by randomization

- It avoids personal bias
- It makes the test valid

Replication

- It is the repetition of the same thing,that is by replication we mean the number of experimental units receiving a particular treatment. If an experiment have equal replication for all the treatments studied, then the design is called “Equi-replicated design” If a design has got an unequal replication for different treatments then it is known as a design with unequal replication.

Purposes served by replication

- It provides an estimate of experimental error
- It enables us to obtain a more precise estimate of the mean effect of any factor, since precision is 1/S.E. and S.E. = S.D. / $\sqrt{}$ n. As ‘n’ increases precision increases. The more the replication, the more the precision will be.

Local control

- It refers to the balancing, blocking and grouping of the experimental units that is employed in the experiment. It refers to the skillful logical way of grouping the experimental units in such a manner that there is more uniformity within the same group and there is greater variability between different groups.

Purposes served by local control

- To make the experimental design more efficient
- To make the test procedure more powerful

CRITERIA FOR MAKING BLOCKS

- Characters that have influence on the response value before the commencement of the experiment are considered as the criteria for making blocks.
- For e.g. in milk yield studies, the stage of lactation can be considered as criteria for making blocks and in weight gain study, we can consider the initial weight or initial age or breed or sex as criteria for making blocks.
- In completely randomised design (CRD), there is no local control applied, while in randomised block design (RBD) local control is applied in one direction with one criterion and in Latin square design (LSD) in two directions with two criteria.
- Besides these three principles of experimental designs, there are auxiliary variable and control that should be considered while experimenting.

Last modified: Monday, 30 April 2012, 04:40 PM

OTHER PRINCIPLES OF EXPERIMENTAL DESIGNS

Auxiliary Variable

- Some characters are not being altered by the treatments applied but may have influence on the characters under study.
- Such character can be utilised to improve the precision of the estimate and the efficiency of designs.
- They are so chosen that the collection of information about this do not involve cost and labour.

Control

- When no treatment is applied over a group of experimental units we consider these units to constitute a control group.
- The purpose of this control is to make effective comparison.
- Whenever an experiment is conducted to make recommendation of a new treatment, it is better to include a control group as one of the treatment group.

MODULE-17: COMPLETELY RANDOMIZED DESIGN

- Learning objective
- Through this module, one will know one of the common experimental design-Completely Randomized Design(CRD) and will know about how to test the significance of different groups.

COMPLETELY RANDOMIZED DESIGN (CRD)

- This is simplest of all experimental designs.
- This is the design in which the treatments are assigned completely at random to the experimental units or vice versa. i.e. it imposes no restrictions on the allocation of treatments to the experimental units.
- CRD is preferred when all the experimental units considered for the experiment are known to be homogeneous.

- Any number of experimental units and treatments can be utilized in this design.
- As the design is highly flexible and simple, the CRD is widely used.
- Analysis is simple, even if certain values are missing.
- The experimental units will be allotted to the different treatments by using random number table or by lottery method.

COLLECTION AND ANALYSIS OF DATA

- After having randomised the experimental units over different treatments, initial recordings (if any) of the experimental units are noted against each experimental unit.
- For example, in case of weight gain study, initial weights are to be recorded.
- Then the experimental units are subjected to respective treatments and after the experimental period, the response values will be observed. The data will be tabulated as follows.

Data of response values

T_{r1}	T_{r2}	...	T_{rt}
y_{11}	y_{12}	...	y_{1t}
y_{2t}	y_{2t}	...	y_{2t}
.	.	...	.
.	.	...	.
.	.	...	.
.	.	...	.
y_{n1t}	y_{n2t}	...	y_{ntt}
T_1	T_2	...	T_t

- Let us have' t 'treatments each having replications n_1, n_2 ... n_t, then, $n_1 + n_2 + ... + n_t = N$

STEPWISE PROCEDURE

Ho : There is no significant difference between the treatments

Step 1

- Calculation of treatment total T_1, T_2 ..... T_t
- Calculation of grand total (GT) = $T_1 + T_2 + ... + T_t$
- Calculation of correction factor (C.F)

$$(CF) = \frac{GT^2}{N}$$

Step 2

- Calculation of sum of squares

$$\text{Total sum of squares (TSS)} = Y_{11}^2 + Y_{12}^2 + \ldots + Y_{nt}^2 - CF$$

$$\text{Treatment sum of squares (TrSS)} = \frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \ldots + \frac{T_t^2}{n_t} - CF$$

- Error sum of squares (ESS) = TSS – T_rSS

Step 3

- Formation of Analysis of Variance table

Source of variation (S.V)	Degrees of freedom (d.f)	Sum of squares (S.S)	Mean squares (M.S.) = SS/d.f.	F
Treatments	(t-1)	T_r SS	T_rSS = T_rMS/(t-1)	T_r MS/EMS
Error	(N-t)	ESS	EMS = ESS/(N-t)	
Total	(N-1)	TSS		

Step 4

Interpretation

- If calculated F< table value of F for (t-1), (N-1)d.f. at 5% level, F is not significant. H_o is accepted. All the treatments are alike.
- If calculated F> tab F for (t-1), (N-1)d.f. at 5% level F is significant.F=( )*. Ho is rejected.
- Iif calculated F> tab F for (t-1), (N-1) d.f at 1% level F is highly significant. F=( )**. Ho is rejected.
- If F is significant or highly significant, critical difference between treatment means is to be worked out.

Critical difference between any two treatment means is defined as the least significant difference between any two treatment means, to be exceeded by the difference between two treatment means to declare them as significantly different.

Critical difference = Standard error of the

between any two treatment means at 5% level		difference between the treatment means x table value of 't' for error d.f. at 5% level.
Critical difference between any two treatment means at 1% level	=	Standard error of the difference between the treatment means x table value of 't' for error d.f. at 1% level.

- Critical difference between Tr_1 and Tr_2 at 5% (1%)

$$\sqrt{EMS\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \times \text{'t' error d.f at 5\%(1\%)}$$

- Critical difference betweenTr_1, Tr_3 at 5%(1%)

$$\sqrt{EMS\left(\frac{1}{n_1}+\frac{1}{n_3}\right)} \times \text{'t' error d.f at 5\%(1\%) and so on..}$$

- After this, write treatment means $Tr_1, Tr_2 \ldots Tr_n$ in the ascending order of magnitude. Bar chart representation will be done to show the significant differences among the treatments.

Note

- If we have equal number of replication, i.e, $n_1 = n_2 = \ldots. n_t = n$, then,we have in step 2
- Treatment sum of squares

$$= \frac{T_1^2 + T_2^2 + \ldots + T_r^2}{n} - CF$$

- Critical difference between any two x table 't' for error d.f. at 5% or 1% level treatment means at 5% or 1% level

$$\sqrt{\frac{2EMS}{n}}$$

Exercise

- A set of data involving 4 feedstuffs A, B, C, D tried on 20 chicks is given below. All the 20 chicks are treated alike in all respects except the feeding treatments is given to 5 chicks. Analyse the data.
- Weight gain (in gms) of baby chicks

A	B	C	D
55	61	42	169
49	112	97	131
42	30	81	169
21	89	95	85
52	63	92	154

MODULE-18: RANDOMIZED BLOCK DESIGN

Learning objective

Reader of this module will understand another experimental design, Randomized Block Design (RBD) where in experimental error is minimised compared to Completely Randomised Design.

RANDAMIZED BLOCK DESIGN (RBD)

- In this design, we make use of the principle of local control with the idea of reducing the variability due to experimental error.
- First we arrange the experimental units into homogeneous blocks such that within the blocks, the experimental units are as far as possible homogeneous and between the blocks there is variability.
- This design can be made use of when the experimenter finds that the available experimental units can be grouped into different homogeneous blocks each containing units as many as the number of treatments or as a multiple of the number of treatments.
- The blocking has to be done on the basis of any observable character, which is likely to have influence over the factor under study.
- For e.g. in the study of comparative effects on weight gain in chicks, the chicks may be grouped according to strains. If we are to conduct a Randomised Block Design to try 't' treatments, we need `bt' number of experimental units, where b = number of blocks and t = number of treatments.
- Each block is a replication by itself and number of replication = number of blocks = b. This is the restriction in the case of RBD. RBD is an equi-replicated design.

RBD is a better design than CRD, as this will have greater precision of the estimates and greater efficiency of the designs.

Randomization

- Having formed the number of blocks, experimental units to the different treatments will be allotted independently for each block randomly and the response values will be tabulated as follows:

Treatment	Tr_1	Tr_2	...	Tr_t	Total
1	y_{11}	y_{12}	...	y_{1t}	B_1

2	y_{21}	y_{22}	...	y_{2t}	B_2
...	...	...	...	...	...
...	...	...	...	...	...
b	y_{b1}	y_{b2}	...	y_{bt}	B_t
Total	**T_1**	**T_2**	**...**	**T_t**	**GT**

STEPWISE PROCEDURE

- Ho : Treatment means do not different significantly

Step 1

- Calculation of treatment total, i.e. T_1, T_2 ..... T_t
- Calculation of block total, i.e. B_1, B_2,....B_t
- Calculation of grand total (GT) = $T_1 + T_2 + ... + T_t = B_1 + B_2 + ... + B_t$ = Sum of the treatment totals = Sum of the block totals

$$\text{Calculation of correction factor (CF)} = \frac{(GT)^2}{b \times t} N = bt$$

Step 2

- Calculation of sum of squares

$$\text{Total sum of squares (TSS)} = \sum\sum Y_{ij}^2 - CF$$

$$\text{ie., sum of square of all the bt values} - CF$$

$$\text{Treatment sum of squares (TrSS)} = \frac{T_1^2 + T_2^2 + ... + T_r^2}{b} - CF$$

$$\text{Block sum of squares (BSS)} = \frac{B_1^2 + B_2^2 + ... + B_b^2}{t} - CF$$

$$\text{Error sum of squares (ESS)} = TSS - (TrSS + BSS)$$

Step 3

- Formation of Analysis of Variance table

ANOVA Table

Source of	d.f.	Sum of	Mean squares	F

variation		squares		
Treatments	(t -1)	T_r SS	T_rMS = T_rSS/(t-1)	T_rMS/EMS
Block	(b -1)	BSS	BMS = BSS /(b-1)	
Error	(b-1) (t-1)	ESS	EMS = ESS/(b-1) (t-1)	
Total	**N-1**	**TSS**		

Step 4

Interpretation

- If calculated F< table value of F for (t-1),(b-1)(t-1) d.f. at 5% level, F is not significant. H_o is accepted. All the treatments are alike.
- If calculated F> tab F for (t-1),(b-1)(t-1) d.f. at 5% level F is significant.F=( )*. Ho is rejected.
- If calculated F> tab F for (t-1),(b-1)(t-1) d.f at 1% level F is highly significant.F=( )**. Ho is rejected.
- In the last two cases we have to calculate critical difference between any two treatment means at 5% (1%) level.

$= \sqrt{\frac{2EMS}{b}}$ x table value of `t' for error d.f. at 5% (1%) level.

- Bar chart representation will be done to show the significance between treatments.

ADVANTAGES AND DISADVANTAGES OF RBD

Advantages

- This is a simple design with one local control for more efficient utilisation of the available experimental units. RBD takes into account and eliminates the assignable source of variation among the experimental units by means of grouping the more homogeneous units together.
- This reduces the experimental error and the test of significance become more efficient.
- Any number of treatments and any number of replications may be included but each treatment should have same number of replications.

Disadvantages

- When the data from some experimental units are missing the “Missing plot technique” has to be used.
- If the missing observations are more, this design is less convenient than CRD.

Exercise

- Consider the results given in following table for an experiment involving 6 treatments in 4 randomized blocks.

Response values for a RBD

Blocks	1	2	3	4	5	6
1	24.7	20.6	27.7	16.2	16.2	24.5
2	27.3	28.8	22.7	15.0	17.0	22.5
3	38.5	39.5	36.8	19.6	15.4	26.3
4	28.5	31.0	34.9	14.1	17.7	22.6

MODULE-19: BIOLOGICAL ASSAYS

Learning objective

This module deals with, biological assays and its types.

BIOLOGICAL ASSAYS

Meaning of bioassy dosage response curve

- Biological assay or bioassay is a form of experiment for the estimation of the potency of a substance or comparing the efficacy of two or more substances by means of the reaction that follows their application to living matter. Bio–assay is different from purely comparative experiment. In the usual experiments, the magnitude of effects of different treatments are compared whereas in bio-assays, the potencies of treatments are compared.
- Bio-assays are thus a type of experiments with the object of comparing the efficacy of two or more substances, or preparations, like drugs, by using responses produced by them on suitable living organisms. This technique is used more in pharmacological investigations for comparing the potency of two or more preparations of individual drugs.
- The bio-assay involves a stimulus applied to a subject and the response of the subject to the stimulus.
- The stimulus may be a pesticide, a fungicide, a vitamin, and so on. The intensity of the stimulus may be varied so as to vary the dose given to the subject. The dose can be measured as a weight, a volume or a concentration. The subject may be an insect, a plant, a bacterial culture, etc.
- When a stimulus is applied to a subject there may be a change in some characteristics of the subject. For example, weight of the whole subject or of some particular organ may change, an analytical value may change, recovery from symptoms of a disease may appear, or the subject may die. Such changes in the subject are known as responses. The response may be quantitative as in the case of weight or qualitative as in the case of mortality. The magnitude of response depends upon the dose.
- Normally, two preparations having a common effect are taken for assaying. One of the preparations is of known strength and is called the standard preparation and the other is of unknown strength and is called test preparation. The objective of the assay is to estimate the potency of the test preparation relative to that of the standard preparation.
- Let z_s and z_t denote the doses of the standard and the test preparations respectively such that each of them produces a pre-assigned response in some living organism.

- Then the ratio $R=Z_s/Z_t$ is called the relative potency of the test preparation. If R is greater than unity, it shows that a smaller dose of the test preparation produces as much response as a relatively larger dose of the standard preparation and hence the potency of the test preparation is greater than that of the standard preparation. Similarly when R is less than 1 the potency of the test preparation is smaller than that of the standard preparation. If R = 1,the two preparations are equi-potent.
- An assay with two preparations containing the same effective ingredient, which is responsible for the response, is called *analytical dillusion assay*. An assay with two preparations which have a common effect but do not contain the same effective ingredient, is called a *comparative dillusion assay*.

TYPES OF ASSAY

- There are three main types of bio-assays. They are
  - Direct assays
  - Indirect assays based upon quantitative responses and
  - Indirect assays based upon quantal responses

DIRECT ASSAY

- In direct assays doses of standard and test preparations are administered to randomly selected identical subjects. The administration of the stimuli is stopped as soon as the pre-assigned response has occurred. In direct assays the tolerance doses, the doses below which no response occurs for the standard and test preparations, are measured directly as soon as the response has occurred.
- The tolerance will generally vary considerably from subject to subject. Hence, a number of trails are required to estimate the average tolerance. For obtaining the needed data, the common designs of experiments are used. Once data are obtained, the average of the tolerances, z_s and z_t for the standard and test preparations respectively are calculated. The estimate of the relative potency is then obtained as $R = z_s/z_t$ R is estimated through Feiller's theorem.

INDIRECT ASSAY

Indirect assays based on quantitative responses

- In this type of assays, specified doses of stimuli and their responses are recorded.
- The responses may be a change in weight, a change in analytical value, time of survival of the subject and the like.
- The relationship between the dose and response drawn as a frequency curve is known as the dose response curve.
- This curve is then ascertained which is usually a sigmoidal curve.
- Next, the dose corresponding to a given response is obtained from the relation.
- Such results are obtained for both standard and test preparations. Finally, the relative potency is estimated.

Under indirect assays based on quantitative responses there are two types assays. They are,

  - Parallel line assays,
  - Slope – ratio assays.

Parallel line assays

- The parallel line assays are those in which the relationship between the quantitative response and log dose is linear. The lines for the standard and test preparations shall be parallel.
- A parallel line assay in which the standard and test preparations have an equal number of doses and an equal number of subjects for each dose is called a *symmetrical parallel line assay*. Otherwise, it is called an *asymmetrical parallel line assay*.
- In a symmetrical parallel line assay there are k doses of each of standard and test preparations. In all there are 2k doses in this assay. To each of the doses 'n' subjects are allotted at random. Hence, it is called a 2k – point symmetrical parallel line assay. The constant k may be 2, 3, 4, etc. When, k = 2, we have 4 point parallel line assay. Similarly, we have 6 point parallel line assay, 8 – point parallel line assay, etc. Among these, 4 – point and 6 point assays are used very commonly. The most popular design is the 4-point assay.

Slope ratio assays

- When the response is linearly related to the dose, x raised to power i, the assay is known as slope ratio assay. The value of i is taken as 1 which is adequate and very commonly used constant. The equations for the lines of the standard and test preparations will be,

$$y_s = a + b_s x_s \quad y_t = a + b_t x_t$$

- The two lines converge at zero dose. The relative potency is, therefore, estimated from the ratio of the slopes (regression coefficients) of the fitted lines. Hence the name slope ratio assay.
- The dose – response relationships for the two preparations may be expressed as a multiple linear equations,

$$y_s = a + b_s x_s + a + b_t x_t$$

- The estimates of b_s and b_t are b_s and b_t, respectively. The estimate of the relative potency p is then defined as,

$$\text{Log LD50} = \log A + \left(\frac{50 - a}{b - a}\right)\log^2$$

- As in the case of parallel line assays, the symmetrical designs are ideal for the slope ratio assays. Hence, we shall consider here only assays with (2k + 1) doses. We may have 3 – point, 5 – point, 7 – point, etc., assays. The 3-point assays is the most efficient design.

QUANTAL RESPONSE ASSAY

- In many biological assays the responses are qualitative in nature, for example, in the assay of insecticides the response is mortality of insects. Such qualitative responses are also known as *all – or – none responses*. The assays in which the responses are qualitative are known as *quantal response assays*.
- In the quantal response assay, the subject is given a predetermined dose of the preparation under test and is observed to see whether or not a specified response occurs.
- Thus, quantal response assays are closely related to direct assays. In this type of assay, the strength of a preparation is charactersied by the median tolerance or the dose that induces 50% responses.
- If the response is mortality it is called *median lethal dose* and is denoted by LD_{50}.

- If the response is not mortality, it may be called *median effective dose (ED_{50}), median knock down dose (KD_{50}), median antifeeding dose (AD_{50})* and the like.
- The most commonly used measurement is LD_{50}. The ratio LD_{50} / ED_{50} is called *therapeutic index.*

Last modified: Monday, 30 April 2012, 04:58 PM

METHODS OF ESTIMATING LD50

When the number of subjects used is relatively small and the doses are fairly close, LD50 may be estimated using Dragstedt – Behren's method. The estimate is given by

$$\text{Log LD50} = \log A + \left(\frac{50-a}{b-a}\right)\log^2$$

where,

- A = dose corresponding to a percentage of mortality immediately below 50% mortality,
- a = observed mortality (%) immediately below 50% mortality, and
- b = observed mortality (%) immediately above 50% mortality

SPEARMAN - KARBER METHOD

- Another simple method of estimation of LD50 is Spearman – Karber method. It is used when the log doses are equally spaced. Suppose that the log doses are denoted by $x_1, x_2,.....x_k$. If the log doses are equispaced, then $x_{i+1} - x_i = d$ for all i.
- The LD_{50} is estimated as

$$\text{Log LD50} = \xi = X_k + \left(\frac{d}{2}\right) - d\sum p_i$$

Where,

x_k = the highest log dose,
p_i = proportion of response for i^{th} dose.

MODULE-20: COMPUTER - AN INTRODUCTION

Learning objective

This module gives an outline on the basics of computers.

COMPUTER - AN INTRODUCTION

What is a computer?

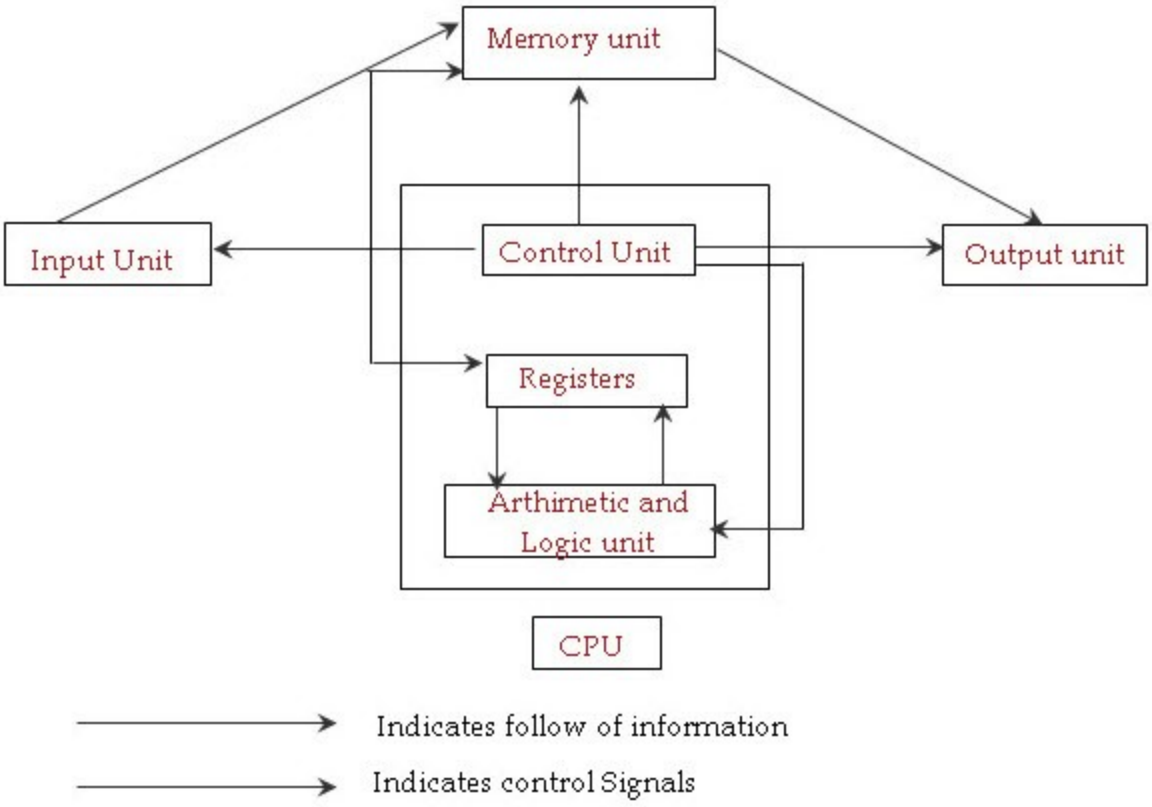
- A computer is a fast, electronic data processing machine/device for receiving, storing, processing, analyzing and retrieving any amount of data or information with 100% accuracy following a set of

instruction given to it by human being. It does all the work assigned to it perfectly without committing any mistake.

Components of computer (anatomy of a P.C)

- The components of computers are
  - Input unit
  - Central Processing Unit (CPU)
    - Control unit
    - Arithmetic and Logic unit
    - Register.
  - Memory unit
  - Output Unit

Block diagram of a computer



COMPONENTS OF COMPUTER

Input unit

- It is a device through which we enter the program and data into a computer. It performs two important functions.
- We feed information or data into the computer for the purpose of processing. Secondly, we instruct the computer to perform various arithmetic operations and the logical sequence in which they are to be computed.

- Punched cards, Punched paper tapes (used in olden days), magnetic tapes, magnetic diskettes, magnetic drums, keyboards, compact disc, etc., are some of the input devices.

Central Processing Unit (CPU)

- It is the main part of a computer system like the heart of a human being. It interprets the instruction in the program and executes one by one.
- It consists of three major units.
- Control Unit
    - It controls and directs the transfer of program instructions and data between various units. The other important functions of control unit are
        - opening and closing of proper logic circuits
        - receiving data from input devices
        - storing them in memory
        - getting instructions from the memory
        - executing the instructions
        - sending the information to the output unit and so on.
- Arithmetic and Logic unit (ALU)
    - Arithmetic operations like additions (+), subtraction (-), multiplication (*), division, exponentiation, etc., and logical operations like comparisons using the operators <, >, <=, >= etc., are being carried out in this unit.
- *Registers:* They are used to store instruction and data for further use.

Memory units

- It is used to store the programs and data. Computers have two types of memories like human being, main memory and auxiliary memory. For a human being brain acts as main memory and notes, books and diaries act as auxiliary memory.
- The main memory is used to store only vital information and auxiliary memory is used to store a lot of information, which is not frequently used.
- The main memory is of the type of Random Access Memory (RAM) and the auxiliary memory is of magnetic memory.
- RAM can retain the information as long as there is electric power. The auxiliary memories are very slow and are very voluminous compared to the main memories. The access time to auxiliary memory is slow compared to that of main memory.

Output units

- Output devices are used to print/display the useful results or processed data that are stored in the memory unit. Paper card punched, paper tape punched (Olden days), dot matrix printer, line printer, plotter, video display unit (VDU), graphic printer, laser printer etc., are some of the output devices.

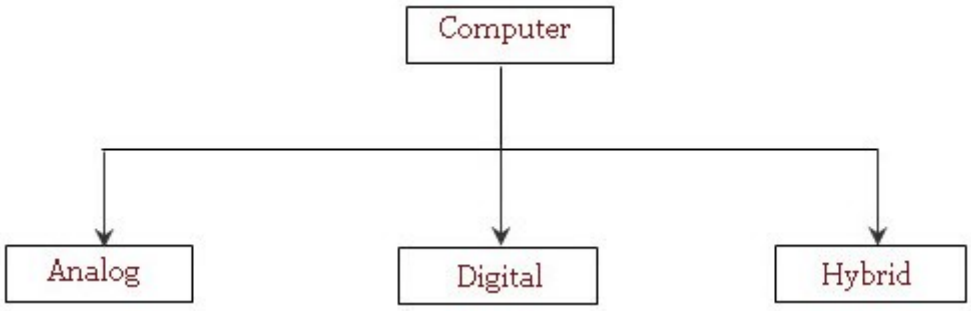
MODULE-21: TYPES OF COMPUTER

- Learning objective
- This module deals with, classification of computer.

TYPES OF COMPUTERS

- The computers are classified based on the type of data they are designed to process. Data may be obtained either as result of counting or through use of some measurement.

- Data obtained by counting are called discrete data. For example, total number of students in a classroom.
- Continuous data is obtained by measurement. For example, measurement of temperature or voltage.



ANALOG COMPUTERS

- These computers use varying physical quantities like voltage, current, temperature, etc as their data values.
- They do not directly count numbers. They deal with variables that are measured along a continuous scale and are recorded to some predetermined degree of accuracy.
- Voltage may be measured to the nearest hundredth of volt and temperature may be measured to the nearest tenth of a degree on Celsius scale.
- Analog computers are frequently used to control process such as those found in an oil refinery where flow and temperature measurements are important. Analog computers may be accurate to within one percent of correct value.

DIGITAL COMPUTERS

- Digital computer is a counting device that operates on discrete data. It operates by directly counting numbers that represent numeral, letters or other special symbols, just as digital watches directly count off the seconds, minutes and hours.
- Digital processors also count discrete values to achieve desired output result. This computer handles binary numbers (1s and 0s).
- A word like READ or any number like 7865 is represented by strings of 1s and 0s.
- They can obtain whatever the degree of accuracy is required simply by calculating additional places to the right of decimal point. In fact, a digital computer once worked out the value of P upto 50,000 decimal places.
- Digital computers can be classified further based on their usage into
    - Special purpose digital computer
    - General purpose digital computer

Special purpose digital computer

- It is one that is designed to perform only one specific task.
- The program of instruction is wired into or permanently stored in such machine. Although it lacks versatility, it does its single task quickly and efficiently. Ex. Atomic submarines to solve complex navigational problems and Black box in the airplane.

General-purpose digital computer

- It is one that can store different programs and thus can be used in various application is known as general-purpose computer.
- By using different instruction such a machine can solve linear equations in one minute. General-purpose computers are very versatile because old programs can be changed and new programs can be written wherever required.
- Based on their size, memory capacity and work length, digital computers are further classified into
    - Microcomputers
    - Minicomputers
    - Mainframe computers
    - Super computers
- General thinking is that the larger the system the greater is its processing speed, storage capacity and cost. The larger systems are also better equipped to handle a greater number of more powerful input and output devices.

MICROCOMPUTERS OR PERSONAL COMPUTERS (PCs)

- Invention of microprocessor in 1972 has changed the computing scene dramatically.
- A microprocessor is a silicon chip that can do the same jobs as the main parts inside a computer.
- It is normally used to mean the single chip containing central processing unit (CPU), but it can also be used to mean the complete microcomputer systems.
- A microprocessor when interfaced with memory and input/output units becomes microcomputers. They are the smallest general-purpose systems.
- They are cheap and limited in capabilities. They are used for data processing by small business organisation, in homes and for hobbies.
- They may perform same operations and use same program instructions as much as the larger computers.
- A PC has all the functional elements found in any larger systems. It has input, storage, ALU, control and output function. Most PCs are light enough to be moved easily from place to place. They are single user oriented i.e. one person at a time.

MINI COMPUTERS

- They are general-purpose computers. They are more powerful and expensive than microcomputers.
- They have more storage capacity and variety of input and output devices. IBM, DEC, NCR, HCL, Univac are some of mini computers.
- In physical size, minis can vary from a small computer sitting on a desktop to one that is the size of a small almyrah.
- The mini computer is usually designed to serve multiple users simultaneously.
- A system that supports multiple users is called multiterminal, time sharing systems. They are popular data processing system in business organisation.

MAINFRAME COMPUTERS

- Mainframe computers are large computers that may offer faster processing speeds and greater storage than minicomputers.
- They may have other special input-output devices. They are used for information retrieval in military and other application. However today, it is becoming increasingly difficult to distinguish between larger minicomputers and smaller mainframe computers in terms of cost, speed and storage capacity.
- They support a large number of terminals for use simultaneously.

SUPER COMPUTERS

- These systems are the largest, fastest and most expensive computers in the world. While the speed of traditional computers is measured in terms of million instruction per second (MIPS), a super computer is rated in terms of million operations per second (MOPS) with an operation consisting of numerous instructions.
- Typically, the super computer is used for large-scale numerical problems in scientific and engineering disciplines. These include applications in electronics, space, petroleum engineering, defense mechanisms, weather forecasting, structural analysis, chemistry, medicine and physics.
- Cyber and Cray-I are some of the super computers.
- Param-10,000 is the super computer produced in India.

HYBRID COMPUTERS

- This type is suitable combination of both analog and digital computer. Nowadays they are becoming obsolete. In a hospital intensive care unit, analog devices could be used to measure a patient's cardiac rate, pulmonary rate, temperature and other vital functional rate.
- These measurements could then be converted into numbers and supplied to digital component in the system.
- This component is used to monitor the patient's vital symptoms and to send immediate signals to nursing terminal if any abnormality in readings is detected.
- Therefore a digital computer can do all the operations that a hybrid computer used to do.
- Analog and hybrid computers obviously perform important specialised task. But overwhelming majority of all computers used for business, offices and scientific application are digital computers.

HUMAN BEING AND COMPUTER

Comparison between HUMAN BEING and COMPUTER

- A computer can be compared to a human being. A human being reads data-using eyes, hears the data using ears, gathers data using feelings, tastes some data using tongue and smells some data using nose.
- So these parts like eye, ear, nose are used to get input and hence these parts are comparable to the input units (or) input devices of a computer.
- The output of information can be given orally using mouth and in writing using hands and sometimes using actions, facial expressions and body postures.
- These parts such as mouth, hands, etc., are used for giving out the information and are comparable to the output devices of a computer. The logical and arithmetic operations are done using the brain.
- The data are also kept in the memory. So, the brain is considered as the memory as well as the processing unit.

Human being	Computer
• Have common sense • Perform arithmetic calculations in minutes or	• Have no common sense • Perform arithmetic calculations in milli-seconds or microseconds.

hours • Having a poor memory power. • Reliability is not bad. • Most of the time gives approximate solutions for a problem. • Have natural intelligence. • Have very good thinking power. • Learn things quickly and act accordingly. • Due to tiredness, cannot do repetitive jobs for a long time.	• Having a very good memory power. • Very good reliability. • At all times gives accurate solutions for a problem. • Have artificial intelligence. • Have no thinking power. • No learning power. • No tiredness and so, do repetitive jobs any number of times with same accuracy and efficiency.

MODULE-22: HARDWARE, SOFTWARE, HUMANWARE AND FIRMWARE

Learning objective

This module gives an overall view of Hardware, Humanware, Firmware and Software.

HARDWARE AND SOFTWARE

Hard ware

- The hardware components of a typical computer are registers, memory cells, adders (Logic gates) and so on.
- Computer hardware consists of five units viz., input unit, output unit, memory, arithmetic logic unit and control unit.

Software

- The computer software is a computer program written in computer languages following certain rules called the syntax of the language, so as to enable the system to obey the instructions carefully.
- In general, the computer software can be categorized as systems software and application software.
- Application software is a set of program necessary to carry out operation for a specified application. Eg: Programs
    - to solve a set of equations
    - to process examination results
    - to prepare a pay-bill for an organization
    - to prepare electricity –bill for each month.
- System software is a general program written for the system, which provides the environment to facilitate writing of application software. Some of the system programs are given below:

- *Complier:* It is a translator system program used to translate a High- level language program into a machine language program.
- *Assembler:* It is another translator system program used to translate an assembly language program into a machine language program.
- *Interpreter*: It is also a translator system program used to translate a High- level language program into a machine language program, but it translates and executes the program line by line.
- Loader: It is a system program used to store the machine language program into memory of the computer.

HUMAN WARE

- Computer Information System (CIS) managers have primary responsibilities for measuring and evaluating of CIS productivity - which is affected by the performance of both computer system and CIS staff members.
- To help maintain productivity, methods are needed to evaluate performance. If performance is below standards, CIS managers may need to provide additional or refresher training for staff members.
- Also, CIS managers must provide training for new employees or employees in new positions.
- A CIS staff involves many categories of employee. CIS training activities cover multiple job functions, including:
  - Data entry operators
  - Operators
  - Programmers
  - System analysts
  - Supervisors and lead people
  - Users.

Data Entry Operators

- Training activities for data entry operators generally are a response to two related sets of requirements. The first requirement involves newly acquired equipment.
- Extensive training may not be needed every time new equipment is acquired. At a minimum, however, orientation sessions usually are held to familiarize operators with new equipment.
- Often, equipment manufacturers supply instructors and materials for these sessions. The second training requirement arises when new or revised application systems are implemented.

Operators

- Training needs for computer operators are similar to the needs for data entry personnel. Purchase of new hardware or installations of new and/or upgraded applications all present needs for training.
- Required training may be in such areas as security and safety procedures.
- A training requirement that is unique to operators involves operating systems. Operating systems may be updated several times a year. Some updates necessitate modifications to application programs. Changes in operating modes lead to additional training for operators.

- Modifications or additions to application or system software also present needs for additional training of programmers. That is, all programmers should be given at least an informal walk-through of all new or modified programs with which they are involved.
  - Programmers should also be given enough information to support their participation in further modifications.
  - This information often takes the form of written memos or updates of software manuals.

System analysts

- A typical CIS practice is to promote systems analysts from the ranks of operators and programmers.
- Systems analysts who follow this route usually have/had no formal training in the theories of analysis and design.
- Of course, the hands-on experience that such professionals do have may be just as valuable as formal training.

Supervisors and lead people

- The major content of training programs for supervisory personnel is aimed at employee and human relations. Supervisors should maintain knowledge of current union contracts, grievance procedures, employment legislation, and the hiring policies of the organization.
- Many organizations require new supervisors to undergo fundamental training in the behavioural sciences.
- Training for supervisors and lead people parallels that of programmers and analysts.
- Training activities pertain to hardware, programming conventions and application or system software modifications and upgrades.

Users

- The proliferation of microcomputers and high-level programming languages has brought about increased user participation in development activities.
- Due to these technological developments, computing and programming are readily accessible to users. As a result, the roles of CIS functions and professionals are changing.
- Under this approach, the CIS function is responsible for the guidance and assistance of users involved in development projects and for the coordination of diverse development activities into an integrated system.
- A user requested an application; analysts and programmers developed the application; then users accepted the system.
- Under current concepts, the user is involved in development activities and may in effect hire the CIS professional as consultant.

FIRM WARE

- Firm ware is the use of microprogrammed read only memory circuits in place of "hardware" logic circuitry. The use of special software (microprograms) to perform the functions of special hard wares (electronic control circuitry).
- Microprograms stored in a read-only storage module of the control unit interpret the machine language instructions of a computer program and decode them into elementary microinstructions, which are then executed.

Engineering firm

- Computers are widely used for designing, drafting and analysing purposes. The newly introduced facility called CAD/CAM (Computer Aided Design/Drafting and Computer Assisted Manufacturing) acts, as an aid for engineers to design a machine part required.
- Among all areas of drafting, CAD is faster and accurate than manual drawing. With a CAD system a drawing can be stored and can be recalled at any time to make changes easily.
    - Alternative designs can be produced. It makes the designer more creative. CAD can be used in almost every drafting discipline such as mechanical, architecture, electronic and piping.

Space world

- Computer plays an important role in space research. Computer is used to design space vehicle such as rockets and satellites. Computers monitor the projectile motion of a launched rocket.
- Special purpose computers are used to receive and process data from satellite. The processed data could help us in geological research, weather forecasting and defence research.

Business firm

- A computer takes relatively very lesser time to generate a payroll report of organisations. Another application is to maintain accounts receivable/operation in big business organisations.
- It is also used for inventory control to give information to the top officials to get profitable balance and to keep stock levels reasonably low while rendering good service to customers. From a sales analysis report generated by a computer, a manager of an organisation can get information regarding the best customer, the best product that brings maximum profit and the best and dull areas in sales.

Hospital

- A hospital data base management system can generate a list of blood donors, disease incidences, etc., Computers can be used to monitor physical parameters like temperature, pressure size.
- Personal and household computers are used to teach lesson for children. A multiple choice question and answer pattern is made to appear on the screen of a computer. The computer compares the answer with already stored correct answer and evaluates them. These systems are known as CAI (Computer Assisted Instruction/ Information) systems.

MODULE-23: TYPES OF MEMORIES

- Learning objective
- This module deals with, memories used in computers.

TYPES OF MEMORIES

- The memory unit holds (stores) all data, instructions, and find results temporarily in input storage area, programme storage area, working storage area and output storage area respectively. The memory consists of hundreds of thousand of cells called "storage locations" each capable of storing one word of information. The memory unit is called by different names, such as storage, internal storage, main memory or simple memory. Main memory is meant to store very critical information and secondary memory for storing less critical and less frequently used information.
- A byte consists of 8 bits (of 1 and 0) and represents one character. In computer parlance, memories are expressed in kilobytes (KB) where a kilo (K) stands for 1024. The storage capacity of a memory is the number of bytes it can store. Memory sizes vary from 64 KB for a home computer to 20GB for a super computer. The address of the location from where a word is to be retrieved or to be stored is entered in a **Memory Address Register (MAR).** The data retrieved from memory or to be stored in a memory are placed in a **Memory Data Register (MDR).** The time taken to write a word is known as the **Write time**. The time taken to retrieve information is called the **Access time** of the memory.
- Based on the mode of access, or the method of access, memory devices may be classified into two categories.
- Memory built into the computer is called main memory and other external devices that have to augment the storage capacity of the main memory are termed auxiliary or secondary memory. Execution of the current program is carried out by the computer in the working area of the main

memory only. Access time for getting information from main memory is less than that of auxiliary memory. Main memory is random (or direct) access storage whereas auxiliary devices may be random or sequential (serial) access storage.

RANDOM ACCESS MEMORY(RAM)

- This is that part of primary storage where data and program instructions are held temporarily while being manipulated or executed. It is called Random Access Memory because any of the locations on a chip can be randomly selected and used to directly store and retrieve data and instructions.
- It takes approximately equal time for READ or WRITE operation within the memory. Data can be accessed randomly i.e. independent of the address of the word.
- Thus the user memory is generally RAM. RAM is volatile in nature - that is the contents are lost when power is switched off.
- RAM chips may be classified as dynamic or static:

Dynamic RAM Chip

- The storage cell circuits in dynamic RAM chips contain
    - A transistor that acts like a mechanical on-off light switch.
    - A capacitor that is capable of storing an electric charge.
- Depending on the switching action of the transistor, capacitor either contains no charge (0 bit) or does hold a charge (1 bit). The charge on the capacitor tends to leak "off"; therefore, provision is made to "regenerate" the storage charge. Thus a dynamic RAM chip provides volatile storage.

Static RAM Chip

- It also provides volatile storage, but as long as it is supplied with power, it needs no special regenerator circuits to retain the stored data.
- Static RAM chips are more complicated because they require more transistors and other devices to store a bit of data.
- Therefore, static RAM's are used in special applications while dynamic RAM's are used in primary storage section.
- Based on mechanical storage devices, memories are classified as Magnetic core memories and semiconductor memories

Magnetic core memory

- For many years, magnetic cores were the principal elements used for internal memory. A magnetic core is a tiny hollow ferrite ring, which can be magnetised by means of an electric current sent through it.
- The direction of the current determines the polarity of the core. Thus a single core can be made to represent a binary 1 or 0. An array of such cores makes up the internal storage for information storage.
- Magnetic core memory is a non-volatile system, because information is retained even if power is switched off.

Semiconductor memory

- Recently semiconductor devices have largely replaced magnetic cores. Semi conductor memory consists of a large number of transistors etched on to silicon chips in high densities acting as a group of memory cells, by retaining or not retaining electric charges.

- A number of such chips may be grouped together to get the desired memory capacity. A transistor is a two-state device that can be made to conduct or not to conduct, thus representing a binary 1 or 0.
- Transfer of information in semi-conductor memory is much faster than magnetic core memory. Magnetic core memories are non-volatile and retain data even if power is lost, whereas semiconductor memories are volatile and lose data when power is turned off.
- Hence they require power back up which is the main draw back. Otherwise, semiconductor units are tiny, cheap, precise and mass produced. Power consumption is less and as a result, loss of energy is minimised.

SEQUENTIAL ACCESS MEMORIES (SAM)

- The access time (the time elapsed between the call and delivery of data) varies depending on its address, information is read one after another in linear sequence.
  - Example: Magnetic type memories,magnetic drum, charge coupled devices and magnetic bubble memories.
- Every processor has a primary storage section (general-purpose storage section) that holds the active programs, and data being processed. In addition to this many processors also have built-in specialised storage elements that are used for specific processing and control purposes.
- One element used during processing operations is a high-speed buffer memory **(cache memory),** which is both fast and expensive per character stored as compared to primary storage.
- This high speed circuitry serves as a "scratch pad" to temporarily store data and instructions that are likely to be retrieved many times during processing, thereby improving the processing speed. Data may be transferred automatically between the buffer and primary storage. Cache memory is used in large as well as small computers.
- In general the term "RAM" is used to mean the semiconductor READ/WRITE memories. There is another term used in computer industry to denote the read-only action of a memory called "ROM". A user can store special functions or programs in a ROM, which is a non-volatile storage.
- ROM chip may contain microprogram control instructions that cause the machine to perform certain operations such as starting the computer or indicating instructions to the entire operating system.
- ROM chip can only be read; it does not accept any input data instructions from the users. The manufacturers set the actual contents of ROM and they are unchangeable and permanent. ROM is mainly of three types:
- PROM
  - These ROM chips are different from the volatile RAM chips because these can retain stored data even when the power goes off. Such memory units are available in 32K and 64 K capacities.
  - PROM allows a chip to be programmed by the user for converting critical and lengthy operations into microprograms that are fused into a chip. Once they are in hardware form, they can be executed at a very high speed. Usually control instructions that cause the machine to perform certain operations can be repeatedly read from a ROM chips as required.
  - However no data or alteration can be written in to a ROM chip by

computer users. The computer manufacturer as part of computer system supplies the most basic type of ROM chip.

- EPROM
  - Erasable PROM. This memory is used to store programs, erase them subsequently and then reprogram.
  - The user can program EPROM chip. Once programmed it behaves like a ROM chip.
- Electrically Erasable and Programmable Read Only Memory (EEPROM)
  - It can be reprogrammed with special electrical pulses.

Last modified: Monday, 30 April 2012, 05:12 PM

MODULE-24: INPUTS AND OUTPUTS

Learning objectives

- This module gives an overall view of
  - Input Units
  - Output Units
  - Secondary Storage Devices

INPUT DEVICES

- Input is usually through a keyboard (like a typewriter) and output may be obtained either on display screen or on a printer.
  - While the printer produces typed copy or on paper (usually known as hard copy), the screen display (soft copy) allows the user to verify the output before it is printed.

Input devices

- An input device presents data to the processing unit in machine - readable form. Although the keyboard is a common input device for a small computer, a system may also support one or more of the input devices.

Keyboard

- The keyboard is very much like a standard typewriter keyboard with a few additional keys.
- The additional keys are included to perform certain special functions such as loading a program or editing a text.
- These are known as function keys. The function keys help user to perform automatically many tasks that would be tedious and time consuming on an ordinary typewriter.

Mouse

- The mouse is a small hand held input device and is used to position the cursor on the screen.
- It is a small plain sized box. Its manipulation on a flat surface moves the cursor in the same direction as the movement of the mouse. The box contains a ball underneath, which senses the movement and transmits it to the cursor electronically.

Optical character Reader (OCR)

- Optical character readers are input devices used to read any printed text. They can interpret handmade marks and characters and special symbols and codes.
- OCR scans the texts optically character by character convert them into a machine-readable code and store the text on the systems memory. They can read at a rate of upto 2,400 characters per second.

Optical Mark Reading and Recognition (OMR)

- Special pre-printed forms are designed with boxes, which can be marked with a dark pencil or ink. Such a document is read by OMR.
- These are applicable in the areas where responses are one out of a small number of alternatives and the volume of data to be processed is large.
- For example
    - Objective type answer in examination in which large number of candidates appear
    - Market surveys, population surveys etc.,
    - Order form containing a small choice of items
    - Time sheets of factory employees in which start and stop times may be marked.

Magnetic Ink Character Recogniser (MICR)

- A MICR can identify characters printed with special ink that contain particulars of magnetic material.
- MICR is used mainly in the banking industry to read cheque/drafts. This eliminates the need to manually enter data from cheques/drafts into a floppy/hard disk. Since the MICR system can recognise only certain character styles, the characters have to be accurately formed.
- Besides saving time, this method ensures accuracy of data entry and improves security.
- Small bars of varying thickness and spacing are printed on packages, books and tags etc., which are read by optical readers and converted to electrical pulses.
- The patterns of bars are unique and standardised. For example, each grocery product has been given a unique 10-digit code and this is represented in a bar code form on every container of this product.

Speech Input Unit

- A unit which takes spoken words as its input, and converts them to a form that can be understood by a computer is called a speech input unit.
- By understanding we mean that the unit can uniquely code (as a sequence of bits) each spoken word, interpret the word and initiate action based on the word.
- Scanners, Floppy diskette, CD-Rom, Hard disk, mark sense reader, graphics tablet, joystick, light pen etc., are other input devices.

PROCESSING UNIT

- The processing unit receives data and instructions, stores them temporarily and then processes the data as per the instructions.
- This part contains the following units.
    - Memory unit (Main memory, secondary memory)
    - Arithmetic logic unit (ALU)
    - Registers
    - Control unit.
- ALU, Registers and control unit are together called as the Central Processing Unit (CPU).

Memory unit

- The memory unit holds (stores) all data, instructions and find results temporarily in input storage area, programme storage area, working storage area and output storage area respectively.
- The memory consists of hundreds of thousand of cells called "storage locations" each capable of storing one word of information. The memory unit is called by different names, such as storage, internal storage, main memory or simple memory.
- Main memory is meant to store very critical information and secondary memory for storing less critical and less frequently used information.

Arithmetic Logic unit (ALU)

- This unit is used to perform all the arithmetic and logic operations, such as addition, multiplication, comparison, etc. For example, consider the addition of two numbers A and B.
- The control unit will select the number A from its location in the memory and load it into the arithmetic logic unit. Then it will select the number B and add it to A in the arithmetic logic unit.
- The result will then be stored in the memory or retained in the arithmetic logic unit for further calculations.

Control unit

- This unit co-ordinates the activities of all the other units in the system. Its main functions are
    - To control the transfer of data and information between various units and
    - To initiate appropriate functions by the arithmetic unit.
- Conceptually, the control unit fetches instructions from the memory, decodes them and directs various units to perform the specified functions.

More on the CPU

- There are two cycles or phases that occur as the CPU considers each program instruction.
- One is the execution cycle and the other the instruction cycle. Although the operations of the ALU and the control unit may seem cumbersome, they are performed with incredible speed.
- Most computers are synchronous machines. That is, the two cycles mentioned above are synchronised by an electronic clock that emits millions of regularly spaced electrical pulses each second.
- Commands are interpreted and then executed at proper intervals, and the intervals are timed by a specific number of these pulses.
- Thus the speed with which an instruction is executed is directly related to the computers built in clock speed - that is the number of pulses produced each second.
- This clock speed is measured in megahertz (or MHz), where Mega means million and Hertz mean times per second. Most of today's popular personal computers have clock speeds in the 273 to 373 MHz ranges. But microprocessors are being designed with a rating of upto 1000 MHz.

OUTPUT DEVICES

- Output devices receive information from the CPU and present it to the user in the desired form. Some of the common output devices are

Visual Display Unit (Screen/monitor)

- When a program is keyed in, the screen (similar to TV Screen) displays the characters. The user can read program line by line and make correction before it is stored or printed on a printer.

- It is also possible to bring to the screen a portion of the program stored in the external storage for editing. The use of cursor facilitates quick and easy editing.
- Screen sizes differ from system to system. The standard size is 24 lines by 80 characters. Most systems have provision for scrolling.
- This facilitates the user to move the text vertically or horizontally on the screen thus bringing to the screen the hidden text.
- Thus the user can scan through the entire file either to review or to select a particular portion. The cursor keys on the keyboard control the cursor on the screen.
- Initially there were only monochrome monitors. Gradually, we began having monitors that display color. Monitors are of different types and have different display capabilities. A special circuit called the Adapter card determines the capabilities.
- Some popular adapter cards are
    - Colour Graphics Adapter (CGA)
    - Extended Graphics Adapter (EGA)
    - Vector Graphics Adapter (VGA)
    - Super Vector Graphics Adapter (SVGA)
- The smallest dot that can be displayed is called a pixel. The number of pixels that can be displayed vertically or horizontally give the maximum resolution of the monitor. The resolution of the monitor determines the quality of the display. The higher the resolution the better is the quality of display. Some popular resolutions are 800X640 pixels, 1024X768 pixels, and 1280X1024 pixels.

Graphic plotters

- Plotters are used to produce output containing graph or diagrams. They use either pen or inkjet approach.
        - Pen plotters are available in two forms, drum type or flatbed type. In the drum plotters, both pen and paper move, while in the flatbed plotter, the paper is fixed and the pen moves.
        - The injection plotters use jets of ink with different colours and are able to produce large drawings containing many colours.
    - Apart from the above printers, other output devices are as follows:
        - Flat Bed Plotter
        - Microfilm and microfiche
        - Drum Plotter
        - Graphic display device (Digitising tablet)
        - Speech output unit.

PRINTERS

- The final output can be obtained from printers. The paper copy obtained from a printer is often referred to as the print out or hard copy.
- There are many types of printers, but they fall into one of the two categories - impact and non-impact printers.
    - In the case of an impact printer, an inked ribbon exists between the print head and paper and the head striking the ribbon prints characters.
    - Non-impact printers use techniques other than mechanical method of head striking the ribbon. Thermal printers, electrostatic printers and laser printers are examples of non-impact printers.
- Printers are available with a variety of printing mechanism, speeds and varying quality. Printers that can print only one character at a time are called the character printers as against the line printers, which print an entire line at a time. Provision is made to obtain a carbon copy if necessary.
- Line printers, Dot matrix printers and Daisy wheel printers are examples of the impact printers.

Line Printer

- Printing speed varies from 150 lines to 2500 lines per minute with 96 to 160 characters on a 15-inch line.
- Six to eight lines per vertical inch are printed. Usually 64 and 96 character sets are used with English letters.
- Two types of line printers are available i.e. Drum Printers, which consist of cylindrical drum. The characters to be printed are embossed on its surface.

Chain printers

- It has a steel band on which the character sets are embossed.

Dot Matrix Printer

- It prints one character at a time, with the print head moving across a line. They are normally slow (30 to 300 Characters per second).
- A character to be printed is made up of a finite number of dots and so, the print head consists of an array of pins.
- Character to be printed is sent one character at a time from the memory to the printer. The character code is decoded by the printer electronics and activates the appropriate pin in the print head.
- They print from left to right as well as from right to left on return. This bi-directional movement enhances the speed of printing.

Letter quality printer (Inkjet Printer)

- The characters are represented by sharp continuous lines and so the output is good looking.
- This printer consists of print head with a number of small holes or nozzles.
- An integrated circuit resistor can heat individual holes very rapidly. When the resistor heats up, the ink near it vaporises and is ejected through the nozzle and makes a dot on a paper placed near the head.
- A high-resolution inkjet printer has around 50 nozzles within a height of 7 mm and can print with a resolution of 300 dots per inch.
- Latest inkjet printers have multiple heads, one per colour, which allows colour printing. The printing speed is around 120 characters per second.

Laser Printer

- An electronically controlled laser beam traces out the desired character to be printed on a photoconductive drum.
- The drum attracts an ink toner on to the exposed areas. This image is transferred to the paper, which comes in contact with the drum.
- Low speed laser printer can print 4 to 16 pages per minute. Very fast printers can print 10,000 lines per minute.

SECONDARY OR AUXILIARY STORAGE DEVICES

- Magnetic surface recording devices commonly used in computers are Hard disks, Floppy disks, CD-ROMs and magnetic tapes.
- These devices are known as secondary or auxiliary storage devices.

Floppy Disk Device (FDD)

- In this device, the medium used to record the data is called as floppy disk. It is a flexible circular disk of diameter 3.5 inches made of plastic, coated with a magnetic material.
- This is covered in a square plastic jacket. Each floppy disk can store approximately one million characters.
- Data recorded on a floppy disk is read and stored in a computer's memory by a device called floppy disk drive.
- This disk is normally rotated at 300 revolution per minute. 5 1/4 " Floppy, 3 1/2 " Floppy.

Compact Disk (CD)

- CD - ROM (Compact Disk Read Only Memory) uses a laser beam to record and read data along spiral tracks on CD.
- A disk can store around 650 MB of information. CD-ROM's are normally used to store massive text data (such as encyclopedias), which is permanently recorded and read many times.
- Recently CD writers have come in the market. Using a CD writer, lot of information can be written on CD-ROM and stored for future reference.

Hard disks (HD)

- Unlike a floppy disk that is flexible and removable, the hard disk used in the PC is permanently fixed.
- The hard disk used in a higher end PC can have a maximum storage capacity of 40-80GB (1Giga byte = 1024 MB or 2^{30} bytes). The data transfer rate between the CPU and hard disk is much quicker as compared to that of between CPU and the floppy disk.
- The CPU can use the hard disk to load programs and data as well as to store data. The hard disk is very important input/output device.
- In summary, a computer system is organised with a balanced configuration of different types of memories. The main memory (RAM) is used to store program being currently executed by the computer. Disks are used to store large data files and program files. Tapes are serial access memories and used to back up the files from disk. CD-ROMs are used to store user manuals, large text, audio and video data.

MODULE-25: EXECUTION OF A PROGRAM

- Learning objective
- This module deals with an overall view of execution of a program.

EXECUTION OF A PROGRAM

- Execution of computer programming or program development requires two specifications. First, the details of the problem to be solved and the type of information required.
    - Second the type of computer available in which the program can be run. Once these specifications are known, we are ready to start from the problem specification, pass through seven stages of development to the successful completion of the program.
    - The seven stages of program development are given below

Algorithm

- Once the problem is known the first thing to be done is the construction of the Algorithm. Algorithm is nothing but the sequence of finite number of steps or computer operations, which if followed will give the solution of the problem.
- This stage requires a complete knowledge of the problem with some creative thinking and problem solving ability.
- The best way of constructing the algorithm is putting them in broad terms so that the programmer can visualise the various possible alternatives.

Flowchart

- The next step is to record the Algorithm into graphical representation known as Flow charts.
- By using suitable data as input, the validity of the Algorithm may be checked.

High Level Language Program (Program Coding)

- Now that, the validity of the logic is checked, the sequence of operations outlined by the flow chart is converted into the format allowed in the programming language.

Input Preparation

- The instruction must be prepared in some form suitable for the computer to receive them.
- This depends upon the type of the input device available in the particular computer.

Compilation

- The source program, which is in a high level language, is fitted into the computer and the compiler converts it into the machine version. At the same time, the compiler also scans.
- Any mistake in the source program for the use of the language is known as "syntactical errors".
- These syntactical errors if any are pinpointed and printed out as "error message" for our attention.

Corrections

- The syntax errors if any are corrected and the source program goes through the process of compilation again.
- Only when the program is syntactically correct, the object version is passed on for the execution by the computers.

Testing Process

- The compiler of a language can detect the syntactical errors in a program. But if there is an error in the logic of the program, it is left undetected by examining the output of the program, which will definitely be wrong.
- So, when testing, specimen data, which the program will be able to handle, is fed in, including deliberate errors to make certain the program is capable of identifying them.
- If the program is not capable of identifying between the valid and invalid input or gives erroneous output go to step 1 to redesign the Algorithm

MODULE-26: DATA TYPES

Learning objectives

- This module deals with various data types such as,
  - Constants
  - Variables
  - Expressions
  - Operators

DATA TYPES

- Data names are user-defined names. Corresponding to every data name, there will be a memory location.
- Whenever we refer a data name the corresponding memory location will be referenced.
- Data items are of two types. They are
  - Elementary data item
  - Grouped data item.
- First let us see an example of an elementary data item and grouped data item.
  - For example the data name HOUR, MINUTE, SECONDS are elementary data items.
  - TIME is a group data item.
  - The memory space referenced by data name TIME is the combined memory space for HOUR, MINUTE, SECOND.
  - So, grouped data item is formed using elementary data items.
  - Elementary data item is the one, which does not contain any data item within it.
  - In the above example TIME is a grouped data item; HOUR, MINUTE and SECONDS are elementary data item.
- Character set consists of letter, digits and special symbols.
  - Letter - A to Z, a to z
  - Digits - 0,1,2,3,4,5,6,7,8,9
  - Special symbols - +, -, #, /, =, :, :, ", ', <, >, (), [], { }, @, $, ^, &, *,
- Identifiers are names denoting constants, types, bounds, variables, procedures and functions. An identifier must begin with a letter, which may be followed by any combination of number of letters and digits. No distinction is made between the upper and the lower case letters.

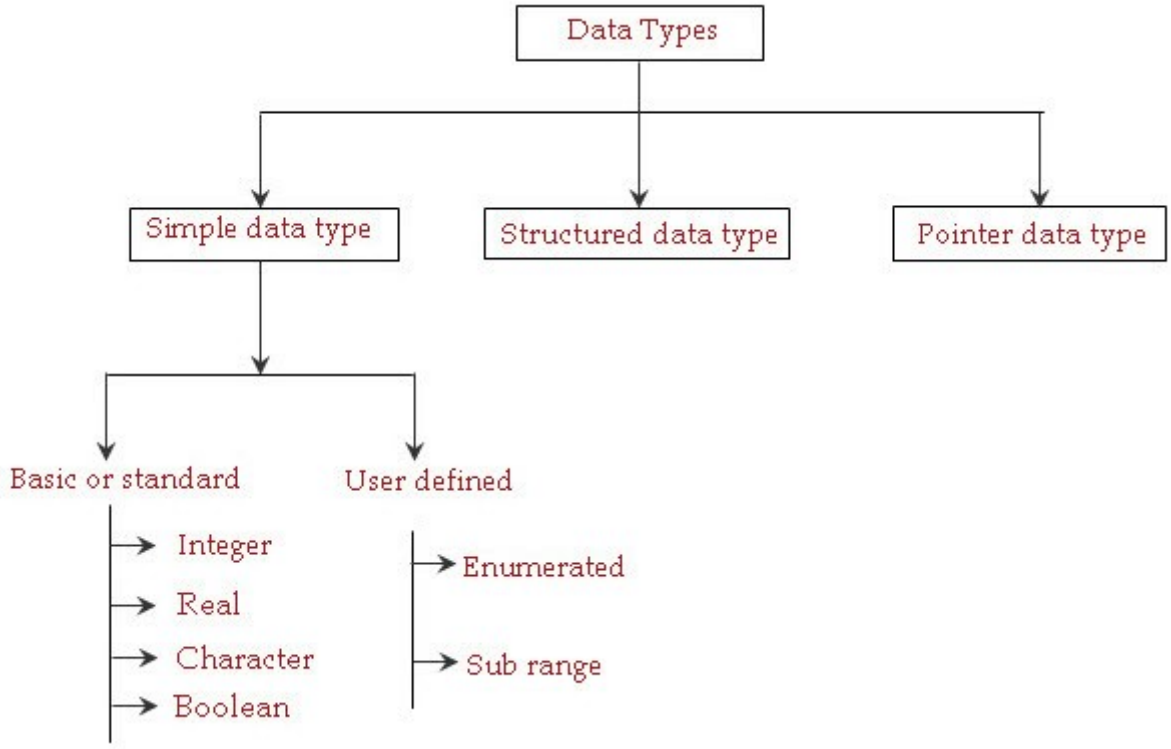
Numbers

- Decimal notation is used for numbers, which denote either integer or real values.
- Integer Number
  - It contains neither a decimal point nor an exponent. Thus a number is a sequence of digits or a digit and may be preceded by a sign. Example: 3, 66, 273243, -34
- Real Number
  - A real number must contain either a decimal point or an exponent (or both). If a decimal point is included at least one digit must precede and follow the point.
  - In the exponent notation, a number 1.32 X10^2 is written as 1.32 E+2. The notation E+2 stands for 10^2. Similarly 0.000012 is written as 1.2 E -5. Example: 2.10059 4.3 E+2 -3E+2 5E-3

Strings

- A string is a sequence of characters enclosed by apostrophes (single quote marks). Example: 'a' '+' 'beginning'.

CLASSIFICATION OF DATA TYPES



- Simple data type
- Structured data type
- Pointer data type

SIMPLE DATA TYPES

- Represents single items (numbers, characters, etc.,) that are associated with single identifiers on a one to one basis.
    - There are four basic data types (integer, real, character, boolean) and two user defined data types (enumerated and sub range types) with the simple data type.
- Basic data type
    - The basic data type includes integer, real, character and boolean data.
- User defined type
    - The user defined data type includes enumerated and subrange.

INTEGER TYPE

- Integer type data are integer numbers, which is an implementation of defined subset of whole number.
- The operators that are used to carry out numerical type operations are called arithmetic operators.
- The following arithmetic operators can be used with integer type operands

Arithmetic operator	Meaning	Type of the operands	Type of the result	Example /a,b,c are integers

-	Unary minus	Integer	Integer	-a
+	Addition	Integer	Integer	a+b
-	Subtraction	Integer	Integer	a-b
*	Multiplication	Integer	Integer	a*b
/	Division	Integer	Integer	a/b
DIV	Truncated division	Integer	Integer	aDIVb
MOD	Remainder after division	Integer	Integer	aMODb

- It should be noted that the integer division (/) gives a real result. The symbol (-) is used for both subtraction and unary minus (-A+B here the (-) before A is an unary minus). The DIV operation for integer gives the quotient after division. The MOD operation gives the remainder obtained in integer division; thus 9DIV4 gives 2 and 9MOD4 gives 1.

REAL TYPE

- Real type refers to data items that represent real numbers. The operators, which can be used with real type operands, are given below.

-	Unary minus	Real	Real	a-b
+	Addition	Real	Real	a+b
-	Subtraction	Real	Real	a-b
*	Multiplication	Real	Real	a*b
/	Division	Real	Real	a/b

CHARACTER TYPE

- Character type data are single character strings i.e. single characters enclosed within apostrophes.
- The character type data can never be subjected to arithmetic operations. But just like numbers, characters also form an ordered set. We mean the digits are ordered set.
- That is the digits are ordered consecutively in their numerical sequence and the letters are ordered in their alphabetical sequence.
- Thus '0' < '1' < '2' < '3' .....< '9' and 'A' < 'B' < 'C' < 'D' .....< 'Z'. Even though there are numerals, they are within quotes and there can be no arithmetic operations.
- The following set of operators, which are relational operators, can be used with all the three data types namely integer, real and character.

=	Equal to

<>	Not equal to
<	Less than
<=	Less than or equal to
>	Greater than
>=	Greater than or equal to

BOOLEAN TYPE

- Boolean type data are truth-values that are either "true" or "false". Logically a statement can be either "true" or "false". We say that the values true and false are logical data or boolean data. These are also ordered false preceding true.

Boolean operators	Meaning	Example /a, b are boolean type data
OR	The result will be true if their operand is true (or if both operands are true) else false	a OR b
AND	The result will be true only if both operands are true else false	a AND b
NOT	The result is the opposite boolean value of the operand (here only one operand is allowed)	NOT a

ENUMERATED TYPE

- An enumerated type specifies an ordered set of values by enumerating the constant identifiers, which denote the values.
- For example, if we wish to work with values that ranged over the months of an year it would be convenient if we have a data type whose constants are Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec and that too ordered in the given sequence.
- The ordinal number of the first constant is 0, second constant is 1, etc. This is done by the following enumerated type declaration.

Type

- Months = (Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec);
- In general, the type definition is written as
- Type name = (data item 1, data item 2, data item3.... data item n);
- Here 'name' is the name of the enumerated data type (ex. Months). The "data item 1", "data item 2".... "data item n" are the ordered constant just as Jan, Feb, ...Dec.
- Thus the association between the "name" of the data type (Months) and the individual data items (Jan, Feb, ...Dec) is established by the type definition.
- The '=', "(",")" are a must and ";" is a delimiter.

Example - Type

- Continents = {Africa, Antarctica, Asia, Australia, Europe, North America, South America}
- Days = (Monday, Tuesday, Wednesday, Thursday, Friday,Saturday, Sunday)
- Colours = (White, Red, Blue, Yellow, Purple, Green, Orange, Black)

Subrange type

- The subrange type is a portion or subrange of any other previously defined ordinal type called its host type.
- The subrange type data are data items that fall within this subrange thus forming a subset of continuous, ordered data.
- The definition of a subrange simply indicates the least and the largest constant value in the subrange, where the lower bound must not be greater than the upper bound.
- The host of the subrange may be previously defined enumerated data as well as the standard data type; integer, char, boolean . Real is not allowed because real is not an ordinal type.
- The general form of a subrange type definition is written as
- Type name = first data item ..... Last data item;
- Here "name" is the name of the subrange data type. "first data item" is the first of the ordered data items within the subrange and the "last data item" is the last of the ordered data items.
- The complete subrange consists of all the data items contained within these two bounds, including the bounds themselves.

Example -Type

- Months = (Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec);
- Summer = Apr... Jul;
- Letter = 'A' ...'Z'

STRUCTURED DATA TYPES

- Structured data types are complex higher-level data types that are built from the collection of simple data types and that contain some additional relationships between the various elements of that simple data type. Structured data types are frequently required because we may need data values, which are a part of an unified collection of related data rather than independent related data.
- Examples: Arrays (the collection of data items of a same type known by a single name), Records, Sets, Files.

CONSTANTS

- It is often convenient to associate a quantity such as a numerical value- a string or a boolean value, with an identifier, thus providing a name for the quantity.
- The identifier is called a constant if the quantity is assigned permanently to it. In other words, the value of the constant identifier remains unchanged throughout the program.
- A constant should always be defined before it is used. Constant refers to the quantity, which does not change during the entire execution of a program. The declaration establishes that identifier to be a constant and associate a value to it.
- The general format or syntax of constant declaration is
- CONST
    - name 1 = value 1; name 2 = value 2; .....name n = value n;

- Here CONST word indicates that it is a constant declaration (as we have already said lower case or upper case makes no difference) name 1, name 2, ... name n are identifiers which represent the constant name and value 1, value 2, value n, are the actual data assigned to the respective constant name. The (;) semicolon indicates the end of one declaration.
- Example:
    - Const Rate = 0.50
    - Bonus = 500; Years = 7;
    - Company = ' The super stockist'
    - Sex = 'm'

VARIABLES

- By a variable, we mean a quantity or data items whose value may change during execution of the program. A variable name is the programmer -supplied name or identifier of a variable.
- Normally, one chooses names so that they indicate the kind of the quantity they represent: e.g. CELSIUS might be chosen to denote the Celsius temperature. Observe that CELSIUS may change during the execution of the program; it may be 20^{o} or 30^{o} depending upon the input.
- So a variable-name can be thought of as a "name tag" for the storage location. The value of the variable at any instant during the execution of the program is equal to the number stored in the corresponding storage location. But the address of memory location to store it remains the same. CELSIUS is the value of variable name CELSIUS at this instance.
- *Single Precision Variable:* This type of variable name ends with the special character "!". Single precision variable has a maximum of 7 digits including the decimal points.
- Examples: Pass marks !, Payment!
- *Double Precision Variable:* Variable name of this type ends with # symbol. This variable can have eight or more digits including decimal point. Example: Pure #, Trace #.
- *String variable:* These are used to hold string constant(s) and the name ends with dollar ($) sign. No arithmetic operations are allowed on these variables. Example: Name$, House $,
- *Integer variable:* Integer variable names end with % sign. These are used to store full numbers. Example: Total num%, Numpass%
- When a variable-name can contain only one element, then it is called a **"scalar variable"**. Variables, which can store an integer number or real number or a character, are example of scalar variables.

Variable Declaration

- It is important thing to have in mind that a variable should be declared to which data type it belongs in the beginning of the program or sub-program itself. The format or syntax, which is to be followed while a variable is declared, is given by VAR
- Variable list: data type
- The declaration statement starts with the VAR (upper case or lower case doesn't matter), then the variable names are given one by one separating each other by comma, then after putting a colon mark, the data type of the variable is given.
- The type can be any one of the four data types integer, real, char or boolean. The word VAR must appear only once even though any number of variable lists is declared
- Examples:
    - VAR
    - Cost, Rate: Real
    - Sex, Grade: Char
    - Teenager, Indian: Boolean
    - No of pupils, count A: Integer

EXPRESSIONS

- It is a combination of constants, variables and arithmetic operators, combined according to the syntax of the language used.
- Even a single constant or variable may form valid expression.
    - Example : P*(R/100), 2.45A

Arithmetic expression

- An arithmetic expression is a series of variable names and constants connected by arithmetic operations. Forming arithmetic expression with arithmetic operators is very simple.
- But two things come as at most consideration; first the data types we use in the expression and second the order in which the arithmetic operations are performed within the expression.
- Depending on the data types used in the arithmetic expression, we shall broadly categorize it, as integer expression, real expression and mixed mode expression.

Integer expression

- An arithmetic expression is an integer expression if the operations and operands in it are of integer type. As we know the operations can be +, -, *, /, DIV, MOD.
- The operands can be integer constants and variable names declared as integers. All valid integer expression produces integer result.
- Example: (a, b, c are all integer)
- -a+b
- a Div b +c
- -a +b*c

Real Expression

- An arithmetic expression is a real expression if the operators and the operands in it are of real type.
- The operators are +, -, *, / and the operands are real constants or real identifiers. All real expressions produce a real result.
- Example: (x,d, c are all reals)
    - -d+x
    - x*d
    - x - 5.09+e
    - x/6.75

Mixed mode expression

- In each of the integer and real expressions the operands are of a same type, both are integer or both are real. Performing operation on operands of different types is called mixing data types.
- Integers are converted into real in expression where both reals and integers appear. Such expression are known as Mixed mode expression. In all other cases other than integer and real mixing of data type is illegal and will cause fatal error.
- Example : (a is integer and b, c are reals)
    - a + b
    - a - b
    - a * b
    - a / c
    - a R b where R can be (< , >, =, < >, < =, > =)

Boolean expression

- A boolean expression is any expression that has the value true or false. For example the expression (average >= 85) would have an answer true if average is greater than or equal to eighty five, and false if average is less than eighty five.
- Here we use relational and boolean operators to construct boolean expression of varied complexity. The relational operators (<, < =, = , < >, >, >=) can be used to compare either real, integer, char, boolean or user defined quantities, but in any given comparison, both operands should be of the same type. The result of a comparison is a boolean value true or false. The logical operators (AND , OR, NOT) operate on boolean values to produce a boolean result.
- Consider the boolean expression (count >= 100) and (not found) or (Ch = '.') or (ch = '!'), In which order it will be evaluated ?
- The expression within parentheses will be evaluated first, then NOT is evaluated, next come AND and OR. So the boolean operators also have a place in the hierarchy of operators. The following table shows the precedence levels for all of the operators
- Highest precedence - Parenthesized expressions
    - NOT, Unary minus
    - *,/, DIV, MOD,
    - +, _ , OR
- Lowest precedence < > =, < =, >=, < >
- Examples:
    - Not found
    - Not ((X>25) or (X< 0))
    - (X> = 0) and (X>= 25)

HIERARCHY OF OPERATORS

- Consider the expression
    - a - -8.201*b - c + d
    - Here in what order the operations are performed? Whether a -8.201 is performed or 8.201*b is performed first? Clearly the result of the computation will be different for each case.
    - But we can use parentheses to obtain the effect we want, as shown below.
    - (a – -8.201)*b - (c+d)
    - (a - (-8.201*b) - c)+d)
    - In each case the computed value is different because of the use of the parentheses. If the parentheses are absent, program uses the hierarchy of operators.
    - Some operators are said to have higher precedence than others. In the absence of the parentheses, which explicitly specifies the ordering, the program will evaluate the arithmetic expressions in such a way that the operators of higher precedence are evaluated before operators of lower precedence.
- The precedence rules are
    - Sub-expressions inside parentheses are evaluated first.
    - Exponentiation (or **)
    - *, /, DIV, MOD operations are evaluated next in their order of occurrence.
    - +, - are evaluated last in their order of occurrence.
- Having seen the rules of precedence we shall determine the order of evaluation of arithmetic expressions.

Expression	Order of evaluation
a*10 + bDiv20 - K MOD	((a*10) + (bDiv20)) - (K MOD 13)

13	
a + b - a * b/c	(a + b) - ((a * b)/c)
a * (b-c)/b+c	(((a*(b-c))/b) +c)

- In all these expressions we can see if a sub expression within a parentheses arises it is taken first, or *, / , DIV, MOD are taken in their order of occurrence next and +, - are taken last in their order of occurrence.

MODULE-27: FUNCTION COMMAND

- Learning objective
- This module deals with function command.

FUNCTION COMMAND

- A standard function is one that performs a number of useful operations. Functions are used when a set of instructions has to be executed repeatedly.
- This set of instructions is kept as a block or function. Whenever we need the instructions we can call the function from the main program. This function block can also be called a module.
- An error in the program is called a 'Bug'. The activity to remove the bug is called 'Debugging'. It is easy to debug if the lengthy programs are split up into function modules. The technique is called modular programming.
- To use a function, we write function name followed by an argument with parentheses.
- Function - name (argument)
    - The argument is the value, which we want the function to use during computation. If we want to compute square root of 5.76, we can simply write Sqrt(5.76). Here Sqrt is the built in standard function name and 5.76 is the argument.

Example

Function	Computes
Sqrt(x)	√x
Sqrt (x+5.76)	√(x+5.76)
Sqrt((x+y) - (x*y)/4.0)	√ (x+y) - (x*y)/4.0)
Sqrt (sin (x))	√ sin (x)

- When we use function two things are of importance. The data type of the function's argument and the data type of the function itself (the data type of the output).
- These two data types need not be the same. For example the function ROUND (x) will take real type data and give integer type data. Some functions accept as input data of more than one type.
- We must always provide an argument of proper type. The following table shows some functions together with their arguments

Function	Name	Definition of argument	Type of argument	Result of argument
Exponential	exp	e^x	Real or Integer	Real
Natural logarithm	ln	Ln(x)	Real or Integer	Real
Sine	Sin	Sin(x) x in radians	Real or Integer	Real
Cosine	Cos	Cos(x)x in radians	Real or Integer	Real
Arctangent	arctan	$\tan^{-1} x$	Real or Integer	Real
Squaring	Sqr	X^2	Real or Integer	Real or Integer
Square root	Sqrt	$\sqrt{x}$	Real or Integer	Real
Rounding	round	x rounded	Real	Integer
Truncation	trunc	x truncated	Real	Integer
Absolute value	abs	{x}	Real or Integer	Real or Integer
Ordinal	Ord	Integer representation of X	Character	Integer
Odd value	Odd	Is x odd	Integer	Boolean
Character	Chr	Character that corresponds to the given integer x	Integer	Character
Successor	Succ	Successor of x in the ordinal set	Any scalar type except real	Same type argument
Predecessor	Pred	Predecessor of x in the ordinal set	Any scalar type except real	Same type argument

MODULE-28: FLOWCHART

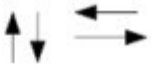
Learning objective

This module deals with flowcharts and its advantages.

FLOW CHARTS

- To illustrate the algorithm in a diagrammatic form, flow charts are used. Flow charts are nothing but a method to lay out in a visual format, the sequence of steps or events necessary to solve a problem by a computer.
- A flow chart is composed of symbols that represent specific activities. Some of the important symbols and their meaning are given in the following table

- Eventhough flow chart is a widely used diagrammatic representation of algorithm, it is not the only way. There are other forms of program description aids such as pie charts and decision tables.

SYMBOL	MEANING
 Rounded Rectangle	TERMINAL Start or Stop
 Parallelogram	INPUT/OUTPUT Get or Give
 Small Circle	CONNECTOR To be continued or continued from
 Rectangle	PROCESS Do something
 Diamond	DECISION Yes or No True or False Reply
	PROCEED THIS WAY

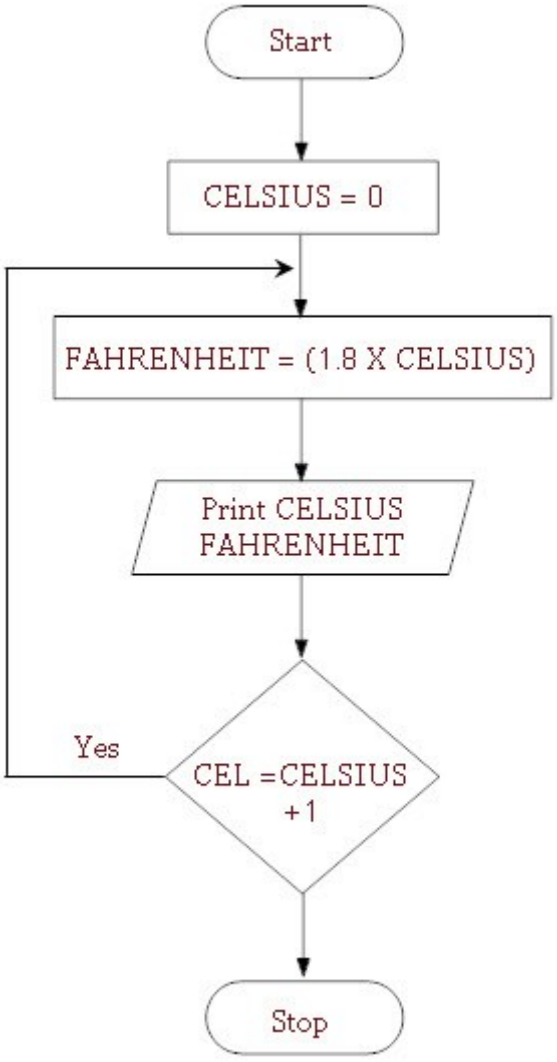
ADVANTAGES OF THE FLOW CHARTS

- They are precise. They represent our thoughts exactly.
- It is easy to understand small flow charts.
- The disadvantage is that in real life flow charts can occupy many pages and hence very difficult to understand. So no one uses flow charts in such situation.

- Example 2: The initial value of CELSIUS is one. The equivalent FAHRENHEIT is calculated and printed out. Now CELSIUS incremented by one. So, CELSIUS becomes two. The equivalent FAHRENHEIT is calculated and printed out. Like wise the same steps are carried out till CELSIUS is less than or equal to hundred.

ALGORITHM

- Step 1 Initialize CELSIUS = 1
- Step 2 FAHRENHEIT = (1.8 x CELSIUS) + 32
- Step 3 Print CELSIUS, FAHRENHEIT
- Step 4 CELSIUS = CELSIUS + 1
- Step 5 If CELSIUS < 101, then GOTO Step 2
- Step 6 Stop



- Human beings often find the repetition of same steps boring, but computers do not. In fact they excel us in many ways. In the above example also the same steps are carried out as long as a condition is true else it comes out. It is obvious that some steps within the group of steps must affect the condition. The path or the group of steps, which are repeated, is called **a loop**. This type

of flow where it involves loop is known as **repetitive flow**. In contrast, the type of flow where the instructions are carried out in the explicitly prescribed sequence is **sequential flow**

MODULE-29: STORAGE OF DATA

Learning objectives

- This module deals with,
  - Storage of data
  - Filing
  - Retrieving and Reprodution

STORAGE OF DATA - INTRODUCTION

- Storing data is to enter the data from hard copies to a magnetic medium such as floppy disk. The data are then stored or classified. Such a step is called "file creation". For example, the financial manager of a firm may wish to group the employees by departments for a payroll program; or the sales manager may wish to sort out the sales completed in terms of salesman.
- These operations produce results, which can be presented in the form of tables or summaries. An example is the total sales of different months in a year by a particular salesman or total commission paid to different salesmen during a particular period.
- It is always better to check the data entered into the computer before these data are being processed. Normally a program called "edit program" can be used to check the correctness of the data entered into the file. Such a step enables the operator to detect the wrong entries, if any, and to correct them.
- This ensures the validity of data. The installed computer facility should possess sufficient secondary storage space in order to store the different sets of data, which can be retrieved whenever they are required.
- Storing of data is arranging data or results of data in some order. The data or results of data can be arranged in ascending or descending order. By means of storing, conclusions can be arrived at easily.

STORAGE OF DATA - FILING

File

- A collection of related data records treated as a unit. Sometimes called a data set.
- All information in a computer is stored in files. Every file has a unique name that helps you to identify it.
- A file name is made up of two components:
  - Main component
  - Extension.

Main component

- The first part of the file name is the main component.
- This is the name given to the file by the user. It can contain alphabets, numbers, spaces and other characters like @, $, #, & {,}, however, there are a few characters that a file name cannot contain. They are: \/: *?"<>,! etc.,

Extension

- This is the second part of the file name. The extension is used to identify the type of file. It is separated from the main component by a period (.) or full stop and is usually three characters long. Usually, when a file is created using an application, the extension is automatically added by the application itself. Some examples of file extensions are .DOC, .BAS, .TXT and .XLS.
- The file name including extension can be a maximum of 255 characters long (windows98) though you can give any name you prefer to your file, it is always better to use a name that tells you some thing about the contents of the file.
- It should also be noted that in MS DOS - an operating system, follows a different set of rules for giving file names. The main components of files created or used on DOS based computers can have a maximum of eight characters and cannot contain spaces. A file created and stored by one program may also be accessed by another program.
- Finding one particular file is very difficult, if you go through all the file names one by one till you locate the files that you need.
- This problem can be overcome by using folders in Windows 98 (and Subdirectories in DOS). A folder is nothing but a collection of related files or subfolder. For example, consider an organisation, its office will have hundreds of papers relating to products, customers, suppliers, personnel, finance and accounts.
- Normally these papers are filed into different folders and stored in a filing cabinet.
- Labels on the folders and the cabinets make it easier to find what you need. So when a person wants some information about a supplier named Hoe&Co., he has to look only in the cabinet marked suppliers and search for the folder marked Hoe&Co. In the same way Windows allows us to organise the files on the disk by grouping them into folders.
- Files, which allow the random selection of file elements, are known as Random Access files. Records in random files can be accessed directly by specifying its record key. Normally these files can be stored in floppy disk and hard disk. File components, which can be accessed sequentially as in, cassette tape is known as sequential, file access. Sequential files may be stored in ascending or descending order according to a record key. File processing starts with the first record and proceeds with the next record sequentially.

File Management System

- A software package that allows users to define data items, place these items into specified records, combine these records into designated files, and then manipulate and retrieve stored data in various ways to achieve user goals. A FMS can typically access records from only one field at a time.

File maintenance

- The activity of keeping a file up to date by adding, changing or deleting data.

File processing

- Utilizing a file for data processing activities such as file maintenance, information retrieval, or report generation.
- When we retrieve/open a file, job of setting apart the working memory area for a file is done automatically. This is important operation for any file retrieval. Using a sub routine subprogram by specifying the file name and its number can retrieve files.

- Once a file is retrieved, it can readily be accessed for reading or writing purposes. While closing a file, the primary working memory area reserved for it is released and the remnants of information in it are written into the floppy or hard disk.

STORAGE OF DATA - RETRIEVING AND REPRODUCTION

Retrieval

- It is bringing back the stored file ready for further work.

Reproduction

- All storage mediums have the potential of corrupting the stored data. The reproduction process fortunately takes this into account and is prepared to fix any errors caused by the storage medium. The stored data is reproduced as soft copy (VDU)/ hard copy by using a printer.

MODULE-30: COMPUTER LANGUAGES, THEIR SCOPE AND LIMITATIONS

- Learning objective
- This module deals with scope and limitations of various computer languages.

COMPUTER LANGUAGES THEIR SCOPE LIMITATIONS

- There are two types of computer programming languages. They are
  - Low level languages
  - High level languages

Low level languages

- The earliest computer language developed is called "Machine Language". A machine language program consists of *a set of machine codes*. It is very tedious to write program using machine codes.
- In the early 1950's some programmers tried to simplify the task of writing programmes in machine codes. With a suitable combination of binary codes (machine codes), they formed a set of alphabetic codes called "*assembly codes (mnemonic codes)* ".
- These codes are short words or abbreviations such as READ, ADD, SUB, MOVE, etc. These assembly codes or memories constituted a language called "*Assembly Language*".
  Later, *assembler programs* were written to do translation between assembly codes and machine codes. These languages are called *low-level* because they remain closer to computer's language (Machine language) than to computer user.
- The program given below shows an assembly language program to add two numbers A & B.
- Program code description
  - READ A It reads the value of A.
  - ADD B The value of B is added with A.
  - STORE C The result is stored in C.
  - PRINT C The result in 'C' is printed
  - HALT Stop execution.

High level languages

- Though the discovery of assembly languages reduced the hardship of writing programs to some extend, it is not fully eliminated still they were not closer to the human understandable languages like English.
- The next step was the development " High Level Languages" like FORTRAN, COBOL, PASCAL, C+++, JAVA, ORACLE, etc.,. Since these languages are easy for a beginner to understand and remember the command verb, they are called "High level languages".
- The following program written in BASIC language is to add two given numbers.
- Program code description
    - 10 INPUT A, B To read the values of A &B
    - 20 LET C= A+B A&B are added and result is stored in C
    - 30 PRINT C Print the value of C
    - 40 END Stop execution.

Advantages of High Level Languages

- Since high-level languages command replaces many machine instructions, a program written in high-level languages seems to be shorter.
- Short programs reduce labour and improve clarity.
- High-level language program is machine independent, it can be run in any computer with negligible number of modification unlike machine language, which is machine dependent.
- Since commands in high-level languages are like English, it would be easier for a beginner to learn.

PROGRAM

- A computer program is a set of instructions that converts the high-level language commands into sequences of machine language instructions (Assembly Codes).
- This process is called an implementation of a language/program development. There are two ways of implementation
- Compiler
    - A compiler is a program that translates the entire program text written in high level language (source program) that can later be executed independently of the compiler.
    - The source program is retained for possible modification and corrections and the object program is loaded into the computers for executions.
- Interpreters
    - It is a program that translates a program written in high-level languages, one statement at a time, converting each high-level language command into machine instructions.
    - During the execution of a program, if it finds any error in a particular statement, it halts temporarily there and displays an error message on the screen. It will do further execution only after that error is getting corrected.
    - The important advantage of compiler is speed. On the other side, it is very easy to debug (error finding) a program using a interpreter.
- Some of the High Level Languages:
    - **FORTRAN,** which stands for **FOR**mula **TRAN**slation, was developed at IBM between 1954 and 1957 for scientific and engineering applications.
    - **COBOL** which stands for **CO**mmon **B**usiness **O**riented **L**anguage was developed in 1959 by a committee called **CODASYLC** (**C**onference **O**n **Da**ta **Sy**stem **L**anguage **C**ommittee) and was approved by **ANSI** (**A**merican **N**ational **S**tandard **I**nstitute) in 1968 and another version in 1974 for business application.
    - **Pascal:** It is a structured programming language developed by Nikalaus Wirth in 1968 for scientific application.
    - **C language**

        - It was developed in 1972 by Dennis Ritchie and Ken Thomson of Bell Laboratories.
        - The famous UNIX Operating System is written in 'C'. 'C' has been the primary development languages for personal computers and workstations.
        - The interesting application of 'C' language is common scientific applications such as aircraft testing, oil exploration, modeling fusion reactors, economic planning, cryptoanalysis, astronomy, biomedical analysis, real time speech recognition and robotics. Now the latest version of Visual C+++ is used by the software developers in the industry.
    - **BASIC** which stands for **B**eginners **A**ll **P**urpose **S**ymbolic **I**nstruction **C**ode was developed in Dartmouth College, USA under the direction of J.G. Kemeny and T.E.Kurtz in the mid 1960s.
        - It is easy to learn and it has powerful additional facilities for advanced users. It is designed for interactive use.
        - Now the programmers use the latest visual BASIC version extensively.
    - **d BASE:** It is a product of Aston-Tate of USA, is one of the most popular data base management system (DBMS) package for PCs. All the IBM compatible PCs support this package. It is possible for a manager to learn it in a couple of hours for retrieval of data quickly for decision-making. Even though it is the basis, nowadays, visual Foxpro is more popular among the programmers to solve the modern problems.

MODULE-31: DATABASE MANAGEMENT SYSTEM

- Learning objective
- This module deals with database management system.

DATABASE MANAGEMENT SYSTEM

- *Database:* Data is essential for every activity in an organisation (farm/firm).
- Some of these activities for instance are :
    - Farm Employee Pay roll accounting - (which uses data on employees)
    - Farm crop and livestock accounting - (which uses data on products)
    - Sales analysis - (which uses data on invoices)
- Each entity (e.g. Employee, products, customer) is described by a set of *attributes*. For example, the entity "employee' in pay roll accounting has attributes such as name, code, designation and basic pay; similarly entity "invoice" in sales analysis has attributes like name, consumer number, sex, address, date of purchase, quantity purchased, value of purchase, etc.,
- Each attribute is also called a *data item* or a *field*. All the related fields of data for a particular entity are grouped together to constitute *a record* for the entity. A collection of records for an entity constitutes a *data files* - employee files, inventory file, sales file, consumer file, student file etc., A collection of data files constitute a *data base* and can be defined as an organised collection of operational data used by the application system in an organisation.

NEED FOR DATABASE

- Before knowing the need for modern computerised database, one should understand the drawbacks of conventional system

Data redundancy

- It involves unnecessary duplication or storage of common data items in several registers.
- Example:

- Students Mark statement,
- Attendance Registers,
- Employees pay Register, UPF Register

Data inconsistency

- It is only due to the delayed or irregular updation of records in every register.
- Example:
  - Designation on promotion changed in Pay Register but remains unchanged in UPF Register.

Data dependence

- The data organisation and retrieval from secondary devices are dictated by the requirements of the specific application.
- Example:
  - Retrieval of employee's record in pay register has to be necessarily through his /her last name but not through other attribute like employee code, designation, etc.,

Program dependence

- All the reports generated from conventional systems are program dependent. In other words, if any report format or its contents need to be altered, it calls for reprogramming.

DATABASE MANAGEENT ON PC

- A data base management system (DBMS) overcomes the most of the above drawbacks of conventional system.
- Data redundancy and inconsistency are minimised by maintaining an integrated data base management system and providing access to all application programs.
- The most important advantage of data base management system is the data program independence it offers to application programs. The user is completely relieved from all details about the physical organisation of data on secondary storage devices.
- Adhoc queries can be responded and standard report easily reformatted or changed to suit the individual need. PC based data base management system involves 3 major functions.
  - Creating the database
  - Querying the database
  - Updating the database

Creation of the database

- It consists of two steps
  - Defining a data base
  - Populating a data base
- Defining a database
  - A database is defined by describing the characteristics of the data items in each file.
  - A data item (field) is characterised by its name, type and width. For example, a mark statement can be defined as follows.
  - File name: Mark - (Statement)
  - Description: Field Name Type Width
  - Name Character 15
  - ID Number Alphanumeric 8

    - Sub.Code Alphanumeric 8
    - Mark Numerical 3
    - Ave.Mark Numerical 6.2
    - A width of 15 characters for NAME of student reserves 15-character position to carry name. A width of 6.2 for Ave. Mark reserves 6 position for the numerical filed by using three places for the integer component, two places for decimal components and one place for the decimal point itself.
- Populating a database
    - Once a database has been defined, it must be created on the system. This step is referred to as *populating database*.
    - Care has to be taken to ensure that the data at the time of creation is consistent with its definition given earlier.

Querying the database

- Queries are input to the database using simple commands as in English. Query languages are superior to standard programming languages and are command oriented.
- The system interprets the query commands and provides responses to queries by retrieving the necessary data from the database.
- Retrieval is the most attractive and visible feature of any data base management system.

Updating the database

- This step involves adding, deleting, editing (changing) or updating a given set of data items.
- While updating, great care has to exercise to ensure data integrity, which plays an important role in the design of any data management software.

Data Base Management System

- A few popular data base management software packages are d Base III, R Base, REFLEX, MS Access, FoxPro (visual), PARADOX, ORACLE.

MODULE-32: COMPUTER PROGRAMMES, THEIR SCOPE AND LIMITATION

Learning objective

- This module deals with,
    - Operating systems
    - Utility programmes
    - Language processors
    - Application programmes
    - Statistical packages

COMPUTER PROGRAMMES-THEIR SCOPE AND LIMITATION

- In order to use the computers in problem solving, it is essential to establish a communication between them and their users.
- Computers speak and understand electrical language says "ON" or "OFF" symbolised by the binary numbers 1 or 0. The users, of course, speak their own native language.

- This communication gap between the computer and the users has to be closed if co-operation is to be possible. This is done through a process called programming.
- Computer programming is writing programs or detailed instructions in a computer language to solve a given problem.
- These detailed instructions are to be executed by the computer to have the problem solved.
- Translation of these instruction into machine language is then done by the machine (computer) itself using translation programs such as compilers or interpreters depending upon the way in which these functions.
- Generally the computer programs for farm employee pay roll preparation, farm inventory control, farm production control, customer accounts etc., may be provided as software packages by the computer manufacturers.
- There are four types of computer programmes.
    - Operating System
    - Utility Program
    - Language Processor
    - Application Programs - a) Standard and b) Unique.

OPERATING SYSTEM

- The software that manages the resources of a computer system and schedules its operation is called the operating system. The operating system acts as an interface between hardware and the user and facilitates the execution of the programs.
- The operating system limits the variety and nature of devices, which can be attached to the computer and kind of software, which can be supported.
- The principal functions of operating system include
    - To control and co-ordinate peripheral devices such as printers, display screen and disk devices.
    - To monitor the use of the machine resources.
    - To help the application programs execute its instructions.
    - To help the user develop program.
    - To deal with many faults that may occur in the computer and inform the operator.
- The operating system is usually available with hardware manufacturers and is rarely developed in-house owing to its technical complexity. Small computers use different operating systems. Hence an operating system that runs on one computer may not run on another.
- Example:
    - MS DOS
    - UNIX
    - MACINTOSH
    - WINDOWS 2000
    - CP/M-86

UTILITY PROGRAMS

- There are many tasks common to a variety of applications. For example
    - Sorting a list in a desired sequence
    - Merging of two programs
    - Copying a program from one place to another
    - Report writing
- One need not write programs for these tasks. They are standard and normally handled by Utility Programs, which are pre-written by the manufacturers and supplied with the hardware. They may also be obtained from standard software vendors. A good range of utility programs can make life much easier for user.

LANGUAGE PROCESSORS

- Computers can understand instructions only when they are written in their own language called the machine language.
- Therefore, a program written in any other language should be translated into machine language. Special programs called *language processor* are available to do this job.
- These special programs accept the user programs and check each statement and if it is grammatically correct, produce a corresponding set of machine code instructions.
- Language processors are known as translators. (Compilers and interpreters) They are usually written and supplied by the hardware vendors.

APPLICATION PROGRAMS

- While an operating system makes the hardware run properly, application programs make the hardware do useful works.
- Application programs are specially prepared to do certain specific tasks.
- They can be classified into two categories
- Standard application programs
- Unique application programs
- Some applications, which are common for many organisations, are called standard application programs. Ready to use software packages for such applications are available from hardware and/or software vendors. Standard packages include among others.

STATISTICAL PACKAGES

- TNSTAT, Microstat, SPSS, MS Excel, Lotus, Supercalc, Quicken98, HG
- GW BASIC, PC CARB, DOS MATE, NE, COREL DRAW, SIGMA PLOT, GRAPHIC LAB, SYSTAT, GENSTAT, CLARIS WORK, ARDIS, SAS
- Pay roll
- Linear Programming - LP88
- Time series Package - TSP
- There are situations where one may have to develop one's own programs to suit one's unique requirement; once developed they come into category of unique application packages

MODULE-33: USES OF COMPUTERS

- Learning objective
- This module deals with uses of computers in various fields.

USES OF COMPUTERS IN STATISTICAL ANALYSIS

- The most advantage of using the computers for statistical analysis is that all the options are accessed through a menu and easy for operation.
- Different statistical packages are available in the market namely, Microstat, MSTAT-C, TNSTAT, RSTAT, SPSS, IRRISTAT, etc. Each package varies in their mode of operation in terms of speed, easiness and accuracy.
- Normally the following options are frequently available in the above statistical packages
    - Data Management Subsystem Descriptive statistics
    - Frequency Distribution Hypothesis tests: Mean, Proportion

    - Analysis of Variance Scatter plot
    - Correlation Matrix Regression Analysis
    - Time Series Analysis Nonparametric statistics
    - Crosstab/chi-square Tests Permutation/Combinations
    - Probability Distribution etc.,
- Besides these standard statistical packages, other spreadsheets programs like LOTUS 1-2-3, Quattro pro, VisiCalc, Supercalc, MS-Excel are available for statistical analysis. In addition to the above statistical procedure, graphs and other financial analysis can be carried using the spreadsheets.

USES OF COMPUTERS IN EPIDEMIOLOGY

- The major purpose of epidemiology is to provide data on which a rational decision for the prevention and/or control of disease in animal population can be based.
- In domestic animal this involves optimizing health (productivity) and not necessarily minimizing the occurrence of diseases. The special contribution of epidemiology is providing information describing frequency and distribution of health and disease, identifying factors influencing the occurrence and severity of disease in the population of concern and quantifying the interrelationships between health and disease.
- Various measurements like disease frequency, morbidity rate, mortality rate, proportional rates, variability rates, etc can be worked using the computers for better prediction.
- Statistical measures like Chi-square test, Student's t-test can be done easily by using the computers to measure the association of various factors for epidemics of certain diseases. Similarly epidemiological measures like Relative risk (RR), Population relative risk (RR), Odds ratio (OR), Population odds ratio (OR), for measuring the strength of the disease incidence, Attributable rate (AR), Attributable fraction (AF), Estimated AF for measuring the effect of disease, Population attributable rate (PAR), Population attributable fraction (PAF), Estimated PAF for measuring the importance of prevention in the population can be performed by use of computers.
- Further, instead of conducting epidemiological trial in live animal population, computer simulation models can be developed to study the course of the disease and its effect in the population.

TADInfo

- It is being developed by FAO for use at the national, regional or global level.
- This program is designed to allow those making decision on disease control or eradication to be better informed through a systematic collection and multiple manipulation of reports on disease occurrence. Such reports will be geo referenced to allow the full use of GIS in analysing these reports.
- A link within this approach is software program (LABInfo) for tracking laboratory samples from their collection point.
- ***LABInfo*** is to assist laboratories in recording, analysing, interpreting and presenting their data for tracking and managerial purposes. The main objective of this is to assist in daily management of submission, sample tracking and facilitation of reporting.
- The FAHRMS (Food Animal Health and Resource Management System) computer system was created at the University of Michigan. The objective of this is to develop a dairy herd-monitoring program to provide a research database and to serve as a health management tool.
- Dairy Herd Management (DHMP) software programs developed was found to be useful tool due to its efficiency and timeliness of information when implemented as part of dairy herd health program.
- Dairy Com 305 is another program.

USES OF COMPUTERS IN FARMS

- Farm records may be kept in an on-farm computer or in co-operative society computer.
- The records cover cash income and expenses as well as inventory data on quantity and value of feed, crops and livestock at the beginning and end of each year.
- The records also include an inventory and utilisation record of all farmland (owned and rented); breeding records; livestock birth and death records; and members and weights sold, purchased and consumed. All farm records keeping and management activities can be performed by the use of application software available in the market.
- These softwares generally help to prepare balance sheet, cash flow statement, income statement, farm accounting, different farm records like farm employee payroll, farm stock/ inventory, farm products, customers, etc., for measuring the performance of the farm business.
- DHI programs developed in Canada offer in additional to individual cow production records, somatic cell count (SCC) data, reproductive performance parameters, nutrition information and management worklists.

USES OF COMPUTERS IN VETERINARY HOSPITAL

- Computers are really a boon for hospital administration, case sheet record maintenance, retrieval for follow up action, diagnosis, dispensing medicines, problems solving, artificial breeding and its follow up action, creation of animal population and animal health status data base.
- Computerising clinical records would make it possible to store voluminous data with minute details and keep track of performance of patients. Almost all the sophisticated diagnostic equipments (CAD - Computer aided or assisted diagnosis) are provided with computer facilities.
- For example Echo Cardio Gram, CAT, MRI etc, have computer support.

Veterinary Medical Record Maintenance

- To render the service and to run the veterinary practice effectively, proper maintenance of records in the hospital is essential. For efficient operation of a veterinary practice, basically three types of records are to be maintained Viz.
    - Veterinary medical records
    - Financial records
    - Inventory records.

Problem Oriented Veterinary Medical Record: (POVMR)

- The Problem Oriented Veterinary Medical Record is an excellent format for recording and storing medical data. Lawrence L.Wead developed this in early 1960. There are two concepts underlying the POVMR.
- They are
    - Patient Care
    - The basic process of delivering veterinary medical care.
- There are other special programmes namely,

CARDIO

- the ECG diagnostic software for canine and feline and uses a technique of artificial intelligence called "rule based analysis" to examine. Its primary function is determination and evaluation of the heart rate.

HEMO

- the clinico-pathologic diagnostic program, both were written by Dr. Fredric Stevens, USA. It is a program designed to process a patient's databases of laboratory test results and produce a report that provides a diagnostic analysis of these tests for vet's review and the patient's data.

COWCAD

- It contains a database of cattle diseases with all recorded clinical features of each disease and provides a list of diagnostic possibilities in order of priority in case of problematic cases.

CONSULTANT

- The clinical signs are entered and the program responds with a list of possible diagnoses alongwith a list of recent reference.

PROVIDES (PRoblem Oriented Veterinary Information and Decision Support)

- When clinical signs are entered, the program generates a list of differential diagnoses in order of probability, plus a bibliography, lists of relevant diagnostic test, treatment options and prognostic probabilities.

NATIONAL INFORMATIC CENTRE

- National Informatics Centre (NIC), is a premier Information Technology organization in India which is committed to providing state-of-the-art, solutions for the IT needs of the Government of India at all levels. NIC carries the distinction of being the largest IT Organization in the Country and has set up a satellite based nationwide computer communication network, called NICNET, with over 1400 nodes connecting the National Capital, the State Capitals and the District Headquarters to one another.
- The IT services of NIC range from Consultancy, Software Design & Development, Office Automation and Networking Services to Training, Video Conferencing, CAD, EDI, Multimedia and Internet Services including Web Site Development and Hosting. NIC has a nationwide presence, with its offices spread all across the Country, from Leh to Andaman & Nicobar Islands. The website of national informatics centre is www.nic.in