4th SEM BTTM
CALICUT UNIVERSITY

BUSINESS STATISTICS AND IT-II

2019 ADMISSION

Prepared by

Aliga Rose Many
Rahsina P
Muhammed Nishad. C P
(Department of Tourism)

CPA College of Global of Studies, Puthanathani

Syllabus

TTM4C04: Business Statistics and Information Technology II

Lecture Hours per Week: 5 Credits: 4

Objective:

This course has been devised to give an idea about the use of computer and information technology in the field of tourism and travel industry management. Also, the student should be able to apply various statistical tools in business functions. Pedagogy: A combination of Lecture, Case Analysis, Group Discussion, Seminars, Assignments, Practical's and assigned readings.

Module I Meaning and Definitions of Statistics Scope and Limitations. Statistical enquiries Scope of the problem Methods to be employed types of enquiries Presentation of data by Diagrammatic and Graphical Method Formation of Frequency Distribution. Measures of Central tendency Arithmetic Mean, Median, Mode, Geometric and Harmonic mean, Measures of variation and standard, mean and quartile deviations Skew ness and Kurtosis and Lorenz curve.

Module II Regression and correlation: Simple Correlation Scatter diagram – Karl Pearson's Co efficient of correlation – Rank correlation Regression lines. Analysis of Time Series: Methods of measuring Trend and Seasonal variations Index number Unweighted indices Consumers price and cost of living indices.

Module III MIS and Networking – Management Information System, Types of networks, Different topologies, Concept of DBMS- Database, Characteristics of a Database system, Components of DBMS, Database Users, Database Languages, Database Models.

Module IV IT Systems used in Airlines: Introduction and functions of GDSs-Airline reservation systems, inflight systems, crew scheduling systems, airline scheduling systems, point of sale systems Airport Systems, check in systems, gate scheduling systems, baggage handling and cargo systems-travel distribution systems, online travel agency, other online intermediaries in travel distribution-Disintermediation and reinter mediation : Definition and Concept.

Module V ICT in Destination and Hospitality Management: Introduction-Property Management System Functions and Modules-Guest room systems-F and B Systems- CRSs-Sales and Marketing Systems-Accounting Systems-Guest Information and Entertainment Systems-Destination 64 Management System: Application, uses and functions-Destination Marketing Information Systems-GIS in Destination Management. (Note: About quarter of the hours may be used for practical sessions to demonstrate the use of MS Office applications such as Word, Excel and PowerPoint). Activity: Develop an Amortization Table for Loan Amount – EMI Calculation. Prepare an Overhead Machine / Labour hour rate through matrices. Prepare a Bank Statement using Simple interest and Compound interest. Prepare a Case study. Recommended Practical Study A one /two-week GDS training to the students.

Reference Books:

1. Dileep M.R., 2011, Information Systems in Tourism, Excel Books, New Delhi. ISBN 978-81744-69090

2. Demetrius Buhalis, 2003. eTourism, Prentice Hall: Essex:UK

3.. Sundaresan and Jayaseelan An Introduction to Business Mathematics and Statistical Methods

4. Levine. M. David, Timothy C Krehbiel, Berensen. L. Mark and Viswanathan. P. K, (2011), Business Statistics, A First Course. Pearson Publication, (fifth

5. V. Rajaraman, Introduction to Information Technology, Prentice Hall.

6 Poon A. (1998), Tourism, Technology and Competitive Strategies, CABI.

7 Rayport J.F. & Jaworski B.J. (2002), Introduction to Ecommerce, McGraw-Hill.

8. Management information Systems, (2003). Kenneth C. Laudon and Jane P. Laudon, Pearson Education, New Delhi.

9. Using Microsoft Office, Ed Bott and Woody Leonhard, Prentice Hall of India, New Delhi 1999.

10. Fundamental of Database Systems, Elmasri and Navathe, Addison Wesley, New Delhi.

Module 1

BUSINESS STATISTICS AND INFORMATION TECHNOLOGY II

DEFINITIONS OF STATISTICS

The word statistics is derived from the Latin word 'Status' or Italian word 'Statista' or German word 'Statistik' which means a Political State. It is termed as political state, since in early years, statics indicates a collection of facts about the people in the state for administration or political purpose.

Statistics has been defined either as a singular non or as a plural noun.
Definition of Statistics as Plural noun or as numerical facts:- According to Horace Secrist, 'Statistics are aggregates of facts affected to a marked extent by multiplicity of causes numerically expressed, enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other'.

Characteristics of Statistics

(1) Statistics show be aggregates of facts
(2) They should be affected to a marked extent by multiplicity of causes.
(3) They must be numerically expressed.
(4) They should be enumerated or estimated according to a reasonable standard of accuracy.
(5) They should be collected in a systematic manner.
(6) They should be collected for a predetermined purpose.
(7) They should be placed in relation to each other.

Function of Statistics

The following are the important functions of statistics:
1. It simplifies complexity:- Statistical methods make facts and figures easily understandable form. For this purpose Graphs and Diagrams, classification, averages etc are used.
2. It presents facts in a proper form:- Statistics presents facts in a precise and definite form.
3. It facilitates for comparison:-When date are presented in a simplified form, it is easy to compare date.
4. It facilitates for formulating policies:- Statistics helps for formulating policies for the companies, individuals, Govt. etc. it is possible only with the help of date presented in a suitable form.
4. It tests hypothesis:- Hypothesis is an important concept in research studies. Statistics provides various methods for testing the hypothesis. The important tests are Chi – square, Z-test, T-test and F-test.
5. It helps prediction or forecasting:-Statistical methods provide helpful means of forecasting future events.
6. It enlarges individual's knowledge:-When data are presented in a form of comparison, the individuals try to find out the reasons for the variations of two or more figures. It thereby helps to enlarge the individual's knowledge.
7. It measures the trend behavior:- Statistics helps for predicting the future with the help of

present and past data. Hence plans, programs, and policies are formulated in advance with the help of statistical techniques.

Scope of Statistics or importance or utility of statistics.

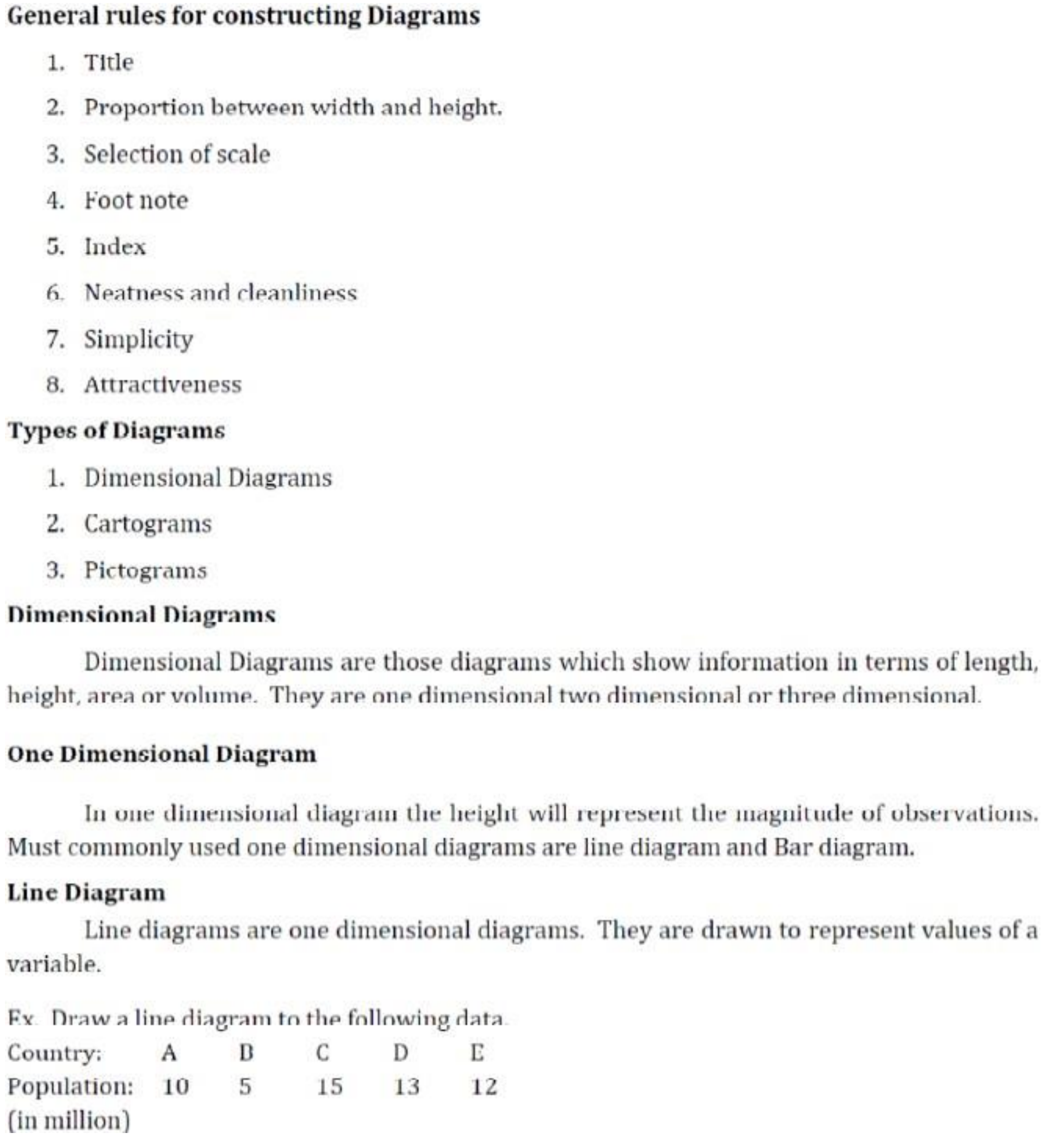
The Scope of Statistics in various field are:
(1) Statistics in Business:- Statistics is most commonly used in business. It helps to take decision making of the business. The statistical data regarding the demand and supply of product can be collected and analyzed to take decisions. The company can also calculate the cost of production and then the selling price. The existing firms can also make a comparative study about their performance with the performance of others through statistical analysis.
(2) Statistics in Management:- Most of the managerial decisions are taken with the help of statistics. The important managerial activities like planning, directing and controlling are properly executed with the help of statistical data and statistical analysis. Statistical techniques can also be used for the payment of wages to the employees of the organization.
(3) Statistics in economics:- Statistical data and methods of statistical analysis render valuable assistance in the proper understanding of the economic problems and the formulation of economic policy.
(4) Statistics in banking and finance:- Banking and financial activities use statistics most commonly.
(5) Statistics in Administration:-The govt. frames polices on the basis of statistical information.
(6) Statistics in research:ı Research work are undertaken with the help of statistics.

Limitation of statistics
(1) Statistics studies only numerical data
(2) Statistics does not study individual cases
(3) Statistics does not reveal the entire story of the problem.
(4) Statistics in only one of the methods of study a problem.
(5) Statistics can be misused. Statistical result are true only an average

Statistical Enquires or Investigation
Statistical Investigation is concerned with investigation of some problem with the help of statistical methods. It implies search for knowledge about some problems through statistical device.
Different stages in statistical enquiry are:
(1) Planning the enquiry
(2) Collection of data.
(3) Organization of data.
(4) Presentation of data.
(5) Analysis of data.
(6) Interpretation of data.

Graphs and Diagrams
Graphs and diagrams is one of the statistical methods which simplifies the complexity of quantitative data and make them easily understandable.
Importance of Diagrams & Graphs
1. Attract common people
2. Presenting quantitative facts in simple.
3. They have a great memorizing effect.
4. They facilitate comparison of data.

5. Save time in understanding data.
6. Facts can be a understood without mathematical calculations.



General rules for constructing Diagrams

1. Title
2. Proportion between width and height.
3. Selection of scale
4. Foot note
5. Index
6. Neatness and cleanliness
7. Simplicity
8. Attractiveness

Types of Diagrams

1. Dimensional Diagrams
2. Cartograms
3. Pictograms

Dimensional Diagrams

Dimensional Diagrams are those diagrams which show information in terms of length, height, area or volume. They are one dimensional two dimensional or three dimensional.

One Dimensional Diagram

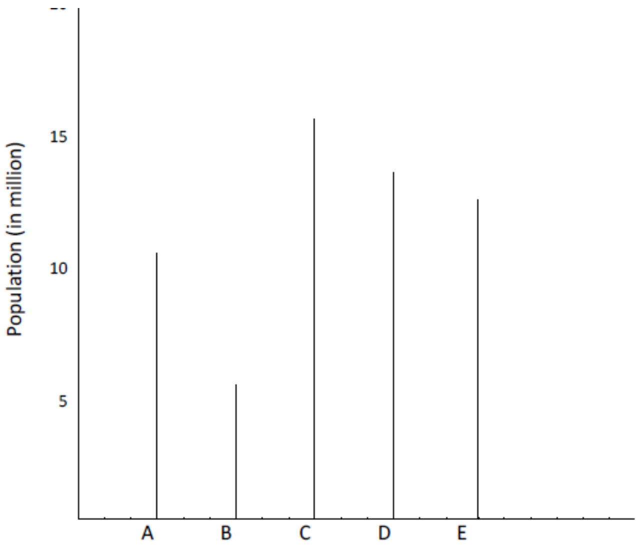
In one dimensional diagram the height will represent the magnitude of observations. Must commonly used one dimensional diagrams are line diagram and Bar diagram.

Line Diagram

Line diagrams are one dimensional diagrams. They are drawn to represent values of a variable.

Ex. Draw a line diagram to the following data.

Country:	A	B	C	D	E
Population: (in million)	10	5	15	13	12



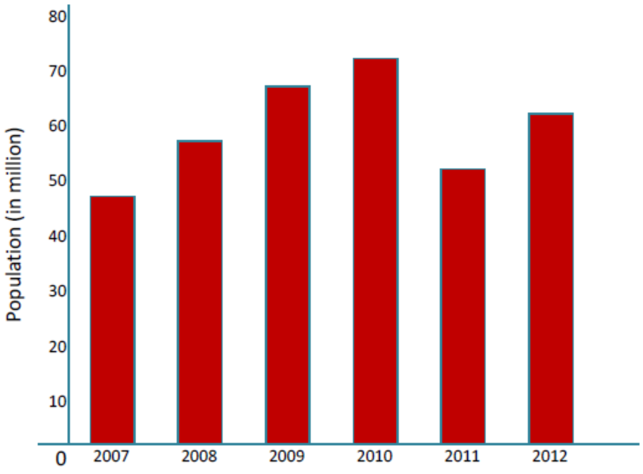
Bar Diagram

In a bar diagram only the length is considered. The width of the bar is not given any importance. Following are the important types of bar diagrams

(1) Simple bar diagram

Simple bar diagram represents only one variable. For example, height, weight, etc.

Year	2007	2008	2009	2010	2011	2012
Sales (In '0000'	45	55	65	70	50	60

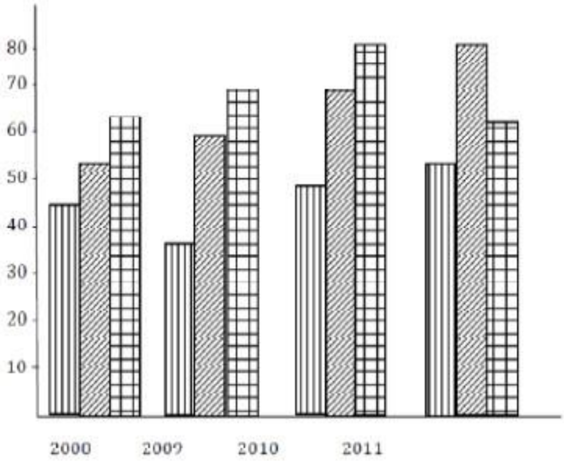


2) Multiple Bar Diagram

Two or more interrelated data are represented in a multiple bar diagram. In order to identity the data, the bars should be differentiated with colors or shades.

Eg:- From the following data draw a suitable diagram.

Year	Production (in units)		
	A	B	C
2008	45	55	65
2009	35	60	70
2010	50	70	00
2011	55	80	60

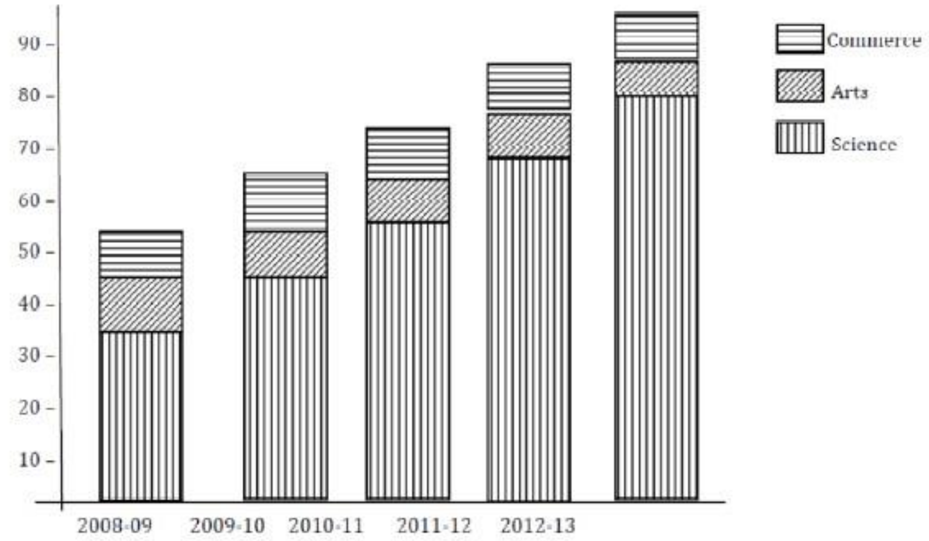


3) Sub Divided Bar Diagram

In the sub divided bar diagram each bar is subdivided into two or more parts. Each part may explain different characters.

Eg:- The number of students in Calicut University are as follows: Represent the date by suitable diagram

Year	Commerce	Arts	Science	Total
2008-09	35000	10000	9000	54000
2009-10	45000	9000	90000	64000
2010-11	55000	7000	8000	69000
2011-12	70000	5000	7000	82000
2012-13	80000	4000	6000	90000



4) Percentage Bar Diagrams

In percentage bar diagram the length of all the base are equal ie each bar represent 100 percent. The component parts are expressed as percentage to the whole.

Eg:- Prepare a subdivided bar diagram on the percentage basis.

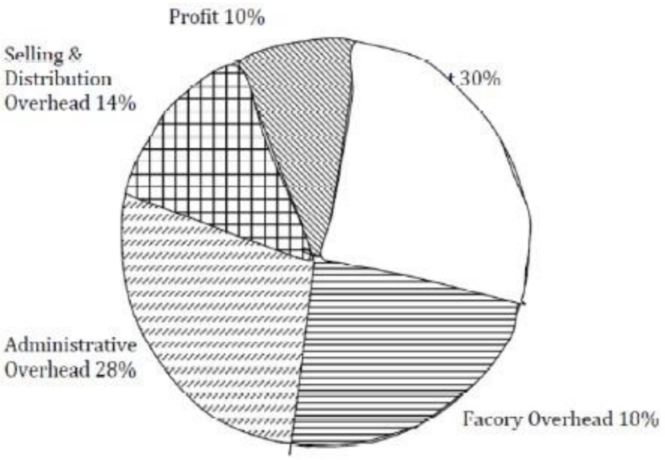
Year	Direct Cost Rs	Indirect Cost Rs	Profit Rs	Sales Rs
2009	35	15	10	60
2010	40	20	12	72
2011	32	22	8	62
2012	25	35	15	75

Answer

Year	Direct Cost in %	Indirect Cost in %	Profit in %	Sales
2009	58	25	17	100
2010	55	28	17	100
2011	52	35	13	100
2012	33	47	20	100

Ans:

Prime Cost	30	108°
Factory over Head	18	65°
Administrative overhead	28	101°
Selling & Distribution overhead	14	50°
Profit	10	36°
	100	360



Three Dimensional Diagrams

Three dimensional diagrams are prepared in the form of cubes, spheres, cylinders etc. In these diagrams width, length and breadth are important.

Cartograms

Cartograms means the presentation of data in a geographical basis. It is otherwise called as statistical maps. The quantities on the map may be shown through shades, dots or colours etc.

Pictograms

Under the pictograms, data are represented in the form of a appropriate pictures most suited for the data.

GRAPHS

Types of Graphs

(1) Graphs of Frequency Distribution

(2) Graphs of Time Series

Graphs of Frequency Distribution

A frequency distribution can be presented graphically in any of the following ways:

(1) Histogram

(2) Frequency Polygon

(3) Frequency Curves

(4) Ogive or cumulative frequency curves.

Histogram

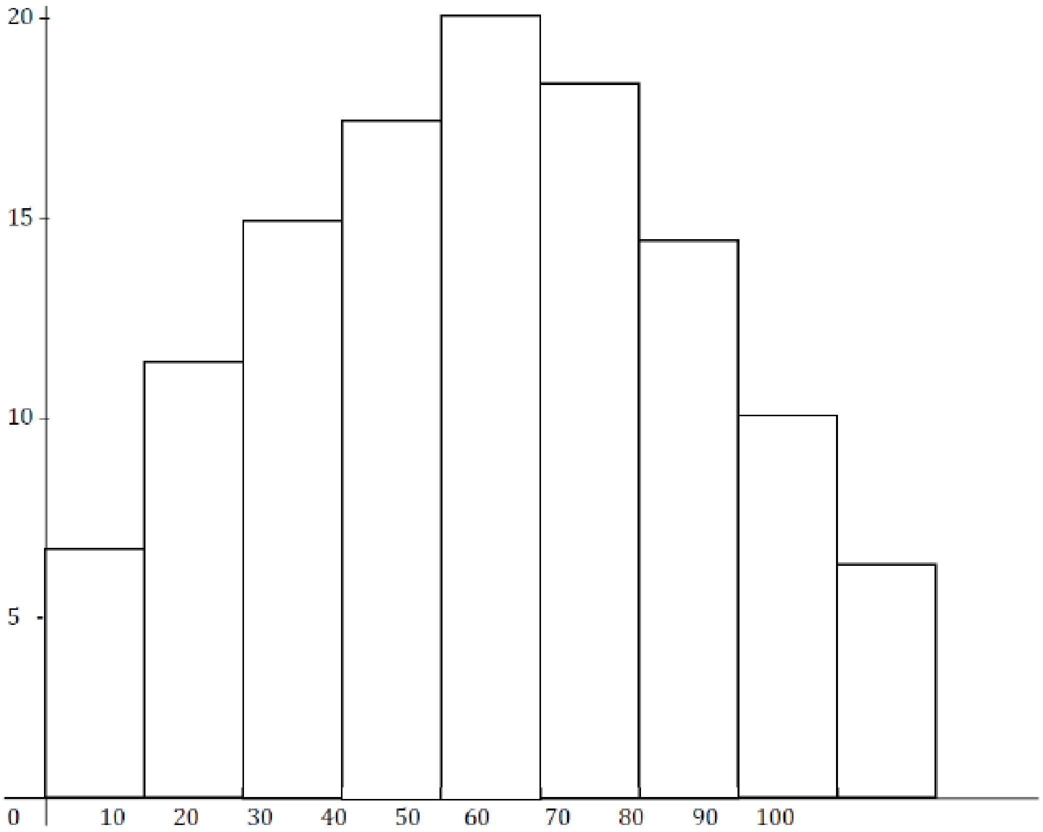
A histogram is a graph of frequency distributions. A histogram consists of bars erected upon the class interval columns.

While constructing histogram, the variable is always taken on the x-axis and the frequency on the y-axis. The width of the bars in the histogram will be proportional to the class interval.

Histogram for frequency Distribution having equal Class interval

1) Draw a histogram from the following information

Marks	No. of Students
0-10	7
10-20	12
20-30	15
30-40	17
40-50	20
60-70	14
70-80	10
80-90	4



Histogram for unequal Class Interval

Unequal class intervals must be corrected.

$$\text{Unequal class intervals} = \frac{\text{Frequency unequal class intervals}}{\text{width of the unequal class intervals}} \times \text{width of the lowest class interval}$$

Draw a histogram from the following data

Daily wages	No. of workers
15-20	4
20-25	9
25-30	12
30-40	20
40-50	16
50-55	7
55-60	6
60-75	15
75-80	4
80-95	9
95-100	2

Frequency Polygon

It is a curve instead of bars. There are two methods for constructing frequency polygon. First, histogram should be drawn and mark mid point of upper side of each bar and join such joints by a curve.

In the second method, first of all plot the frequencies corresponding to midpoints of various class intervals. Then join all the plotted points to get the frequency polygon curve.

3) Ogive or Cumulative Frequency Curve

A frequency distribution when cumulated, we get cumulative frequency distribution and curve drawn is known as ogive. An ogive can either less than ogive or more than ogive. Less than ogive curve is drawn on the basis of less than cumulative frequency distribution and more than ogive is drawn on the basis of more than cumulative frequency distribution.

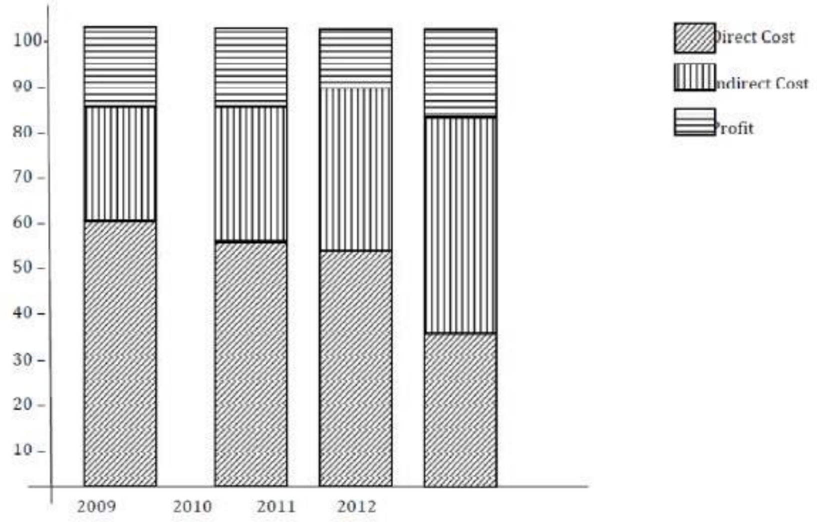
Example :-

From the following data drawn less than and more than ogives

Marks :	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of Students	10	20	35	30	20	15	10	10

Answer :

Less than CF	F	More than CF	F
Less than 0	0	More than 0	150
Less than 10	10	More than 10	140
Less than 20	30	More than 20	120
Less than 30	65	More than 30	95
Less than 40	95	More than 40	55
Less than 50	125	More than 50	35
Less than 60	130	More than 60	20
Less than 70	140	More than 70	10
Less than 80	150	More than 80	0



Two Dimensional Diagram

In two dimensional diagram the length as well as width have to be considered. The most commonly used two dimensional diagrams is pie diagram, Rectangles, Squares, Circles etc are also two dimensional diagrams.

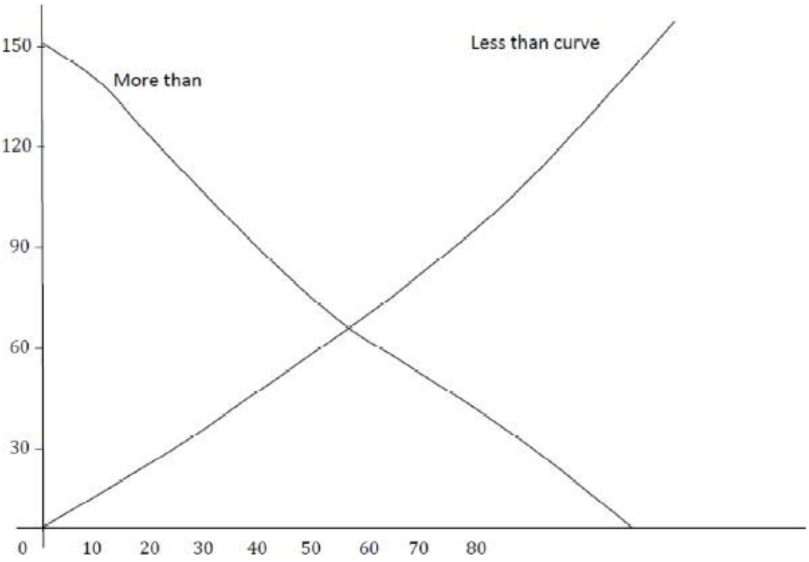
Pie Diagrams

Pie diagrams are used when the aggregate and their divisions are to be shown together. The aggregate is shown by means of a circle and divisions by the sectors of the circle. For example, the selling price of a product can be divided into various segments like factory cost, administrative cost, selling cost and profit. These segments are converted into percentage in order to represent in the pie diagram.

In order to prepare the pie diagram, each percentage outlay must be multiplied by 3.6, since the pie diagram contain 360° scale.

Eg:- Draw a pie diagram from the following data

Prime Cost	30%
Factory over Head	18%
Administrative overhead	28%
Selling & Distribution overhead	14%
Profit	10%



Limitations

1. They can present only approximate values.
2. They can represent only limited amount of information.
3. They can be misused very easily.
4. They are not capable of further mathematical treatment.
5. They are generally useful for comparison purpose only

Various Statistical Techniques

A brief comment on certain standard techniques of statistics which can be helpful to a decision- maker in solving problems is given below.

i) **Measures of Central Tendency**: Obviously for proper understanding of quantitative data, they should be classified and converted into a frequency distribution (number of times or frequency with which a particular data occurs in the given mass of data.). This type of condensation of data reduces their bulk and gives a clear picture of their structure. If you want to know any specific characteristics of the given data or if frequency distribution of one set of data is to be compared with another, then it is necessary that the frequency distribution help us to make useful inferences about the data and also provide yardstick for comparing different sets of data. Measures of average or central tendency provide one such yardstick. Different methods of measuring central tendency, provide us with different kinds of averages. The main three types of averages commonly used are:
a) **Mean**: the mean is the common arithmetic average. It is computed by dividing the sum of the values of the observations by the number of items observed.
b)**Median**: the median is that item which lies exactly half-way between the lowest and highest value when the data is arranged in an ascending or descending order. It is not affected by the value of the observation but by the number of observations. Suppose you

have the data on monthly income of households in a particular area. The median value would give you that monthly income which divides the number of households into two equal parts. Fifty per cent of all the households have a monthly income above the median value and fifty per cent of households have a monthly income below the median income.
c) **Mode**: the mode is the central value (or item) that occurs most frequently. When the data organised as a frequency distribution the mode is that category which has the maximum number of observations. For example, a shopkeeper ordering fresh stock of shoes for the season would make use of the mode to determine the size which is most frequently sold. The advantages of mode are that (a) it is easy to compute, (b) is not affected by extreme values in the frequency distribution, and (c) is representative if the observations are clustered at one particular value or class.

ii) **Measures of Dispersion:** the measures of central tendency measure the most typical value around which most values in the distribution tend to coverage. However, there are always extreme values in each distribution. These extreme values indicate the spread or the dispersion of the distribution. The measures of this spread are called 'measures of dispersion' or 'variation' or 'spread'. Measures of dispersion would tell you the number of values which are substantially different from the mean, median or mode. The commonly used measures of dispersion are range, mean deviation and standard deviation. The data may spread around the central tendency in a symmetrical or an asymmetrical pattern. The measures of the direction and degree of symmetry are called measures of the skewness. Another characteristic of the frequency distribution is the shape of the peak, when it is plotted on a graph paper. The measures of the peakedness are called measures of Kurtosis.
iii) **Correlation:** Correlation coefficient measures the degree to which the charge in one variable (the dependent variable) is associated with change in the other variable (independent one). For example, as a marketing manager, you would like to know if there is any relation between the amount of money you spend on advertising and the sales you achieve. Here, sales is the dependent variable and advertising budget is the independent variable. Correlation coefficient, in this case, would tell you the extent or relationship between these two variables,' whether the relationship is directly proportional (i.e. increase or decrease in advertising is associated with decrease in sales) or it is an inverse relationship (i.e. increasing advertising is associated with decrease in sales and vice- versa) or there is no relationship between the two variables. However, it is important to note that correlation coefficient does not indicate a casual relationship, Sales is not a direct result of advertising alone, there are many other factors which affect sales. Correlation only indicates that there is some kind of association-whether it is casual or causal can be determined only after further investigation. Your may find a correlation between the height of your salesmen and the sales, but obviously it is of no significance.

iv) **Regression Analysis**: For determining causal relationship between two variables you may use regression analysis. Using this technique you can predict the dependent variables on the basis of the independent variables. In 1970, NCAER (National Council of Applied and Economic Research) predicted the annual stock of scooters using a regression model in which real personal disposable income and relative weighted price index of scooters were used as independent variable. The correlation and regression analysis are suitable techniques to find relationship between two variables only. But in reality you would rarely find a one-to-one causal relationship, rather you would find that the dependent variables are affected by a number of independent variables. For example, sales affected by the advertising budget, the media plan, the content of the advertisements, number of salesmen, price of the product, efficiency of the distribution network and a host of other variables. For determining causal relationship involving two or more variables, multi- variable statistical techniques are applicable. The most important of these are the multiple regression analysis deiscriminant analysis and factor analysis.
v) **Time Series Analysis** : A time series consists of a set of data (arranged in some desired

manner) recorded either at successive points in time or over successive periods of time. The changes in such type of data from time to time are considered as the resultant of the combined impact of a force that is constantly at work. This force has four components: (i) Editing time series data, (ii) secular trend, (iii) periodic changes, cyclical changes and seasonal variations, and (iv) irregular or random variations. With time series analysis, you can isolate and measure the separate effects of these forces on the variables. Examples of these changes can be seen, if you start measuring increase in cost of living, increase of population over a period of time, growth of agricultural food production in India over the last fifteen years, seasonal requirement of items, impact of floods, strikes, wars and so on

vi) **Index Numbers**: Index number is a relative number that is used to represent the net result of change in a group of related variables that has some over a period of time. Index numbers are stated in the form of percentages. For example, if we say that the index of prices is 105, it means that prices have gone up by 5% as compared to a point of reference, called the base year. If the prices of the year 1985 are compared with those of 1975, the year 1985 would be called “given or current year” and the year 1975 would be termed as the “base year”. Index numbers are also used in comparing production, sales price, volume employment, etc. changes over period of time, relative to a base

.

vii) **Sampling and Statistical Inference**: In many cases due to shortage of time, cost or nonavailability of data, only limited part or section of the universe (or population) is examined to (i)get information about the universe as clearly and precisely as possible, and (ii) determine the reliability of the estimates. This small part or section selected from the universe is called the sample, and the process of selection such a section (or past) is called sampling.

Schemes of drawing samples from the population can be classified into two broad categories:

1 .Random sampling schemes: In these schemes drawing of elements from the population is random and selection of an element is made in such a way that every element has equal change (probability) of being selected.

2 Non-random sampling schemes: in these schemes, drawing of elements for the population is based on the choice or purpose of selector

Measures of central tendency are also usually called as the averages

. • They give us an idea about the concentration of the values in the central part of the distribution.

• The following are the five measures of central tendency that are in common use: • (i) Arithmetic mean, (ii) Median, (iii) Mode, (iv) Geometric mean, and (v) Harmonic mean (vi) weighted mean

measures of variation and dispersion

1. Dispersion measures the extent to which the items vary from some central value. It may be noted that the measures of dispersion measure only the degree (the amount of variation) but not the direction of variation. The various measures of central value give us one single figure that represents the entire data. But the average alone cannot adequately describe a set of observations, unless all the observations are the same. It is necessary to describe the variability or dispersion of the observations.
2. 3. A good measure of dispersion should possess the following properties } It should be simple to understand. } It should be easy to compute. } It should be rigidly defined. } It should be based on each and every item of the distribution. } It should be amenable to further algebraic treatment. } It should have sampling stability. } Extreme items should not unduly affect it.
3. 4. } Range } Inter-quartile range or Quartile Deviation } Mean deviation or Average Deviation } Standard Deviation } Lorenz curve

Measures of Skewness and Kurtosis

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers. A uniform distribution would be the extreme case.

What Is a Lorenz Curve?

A Lorenz curve is a graphical representation of income inequality or wealth inequality developed by American economist Max Lorenz in 1905. The graph plots percentiles of the population on the horizontal axis according to income or wealth. It plots cumulative income or wealth on the vertical axis, so that an x-value of 45 and a y- value of 14.2 would mean that the bottom 45% of the population controls 14.2% of the total income or wealth. In practice, a Lorenz curve is usually a mathematical function estimated from an incomplete set of observations of income or wealth.

.

Module 2

CORRELEATION ANALYSIS

Introduction:

- In practice, we may come across with lot of situations which need statistical analysis of either one or more variables. The data concerned with one variable only is called univariate data. For Example: Price, income, demand, production, weight, height marks etc are concerned with one variable only. The analysis of such data is called univariate analysis.
- The data concerned with two variables are called bivariate data. For example: rainfall and agriculture; income and consumption; price and demand; height and weight etc. The analysis of these two sets of data is called bivariate analysis.
- The date concerned with three or more variables are called multivariate date. For example: agricultural production is influenced by rainfall, quality of soil, fertilizer etc.

Definition:

Two or more variables are said to be correlated if the change in one variable results in a corresponding change in the other variable.

According to Simpson and Kafka, "Correlation analysis deals with the association between two or more variables".

Lun chou defines, "Correlation analysis attempts to determine the degree of relationship between variables".

Boddington states that "Whenever some definite connection exists between two or more groups or classes of series of data, there is said to be correlation."

Correlation Coefficient:

Correlation analysis is actually an attempt to find a numerical value to express the extent of relationship exists between two or more variables. The numerical measurement showing the degree of correlation between two or more variables is called correlation coefficient. Correlation coefficient ranges between -1 and +1.

SIGNIFICANCE OF CORRELATION ANALYSIS

Correlation analysis is of immense use in practical life because of the following reasons:

1. Correlation analysis helps us to find a single figure to measure the degree of relationship exists between the variables.
2. Correlation analysis helps to understand the economic behavior
3. Correlation analysis enables the business executives to estimate cost, price and other variables.
4. Correlation analysis can be used as a basis for the study of regression. Once we know that two variables are closely related, we can estimate the value of one variable if the value of other is known.
5. Correlation analysis helps to reduce the range of uncertainty associated with decision making. The prediction based on correlation analysis is always near to reality.
6. It helps to know whether the correlation is significant or not. This is possible by comparing the correlation co-efficient with 6PE. It 'r' is more than 6 PE, the correlation is significant.

Classification of Correlation

Correlation can be classified in different ways. The following are the most important classifications

1. Positive and Negative correlation
2. Simple, partial and multiple correlation
3. Linear and Non-linear correlation

Positive and Negative Correlation

Positive Correlation

When the variables are varying in the same direction, it is called positive correlation. In other words, if an increase in the value of one variable is accompanied by an increase in the value of other variable or if a decrease in the value of one variable is accompanied by a decree se in the value of other variable, it is called positive correlation.

Eg: 1)					
	A: 10	20	30	40	50
	B: 80	100	150	170	200

When the variables are moving in opposite direction, it is called negative correlation. In other words, if an increase in the value of one variable is accompanied by a decrease in the value of other variable or if a decrease in the value of one variable is accompanied by an increase in the value of other variable, it is called negative correlation.

Simple, Partial and Multiple correlation

Simple Correlation

In a correlation analysis, if only two variables are studied it is called simple correlation. Eg. the study of the relationship between price & demand, of a product or price and supply of a product is a problem of simple correlation.

Multiple correlation

In a correlation analysis, if three or more variables are studied simultaneously, it is called multiple correlations. For example, when we study the relationship between the yield of rice with both rainfall and fertilizer together, it is a problem of multiple correlation.

Partial correlation

In a correlation analysis, we recognize more than two variable, but consider one dependent variable and one independent variable and keeping the other Independent variables as constant. For example yield of rice is influenced b the amount of rainfall and the amount of fertilizer used. But if we study the correlation between yield of rice and the amount of rainfall by keeping the amount of fertilizers used as constant, it is a problem of partial correlation.

Linear and Non-linear correlation

Linear Correlation

In a correlation analysis, if the ratio of change between the two sets of variables is same, then it is called linear correlation.

For example when 10% increase in one variable is accompanied by 10% increase in the other variable, it is the problem of linear correlation.

X:	10	15	30	60
Y:	50	75	150	300

Here the ratio of change between X and Y is the same. When we plot the data in graph paper, all the plotted points would fall on a straight line.

Non-linear correlation

In a correlation analysis if the amount of change in one variable does not bring the same ratio of change in the other variable, it is called nonlinear correlation.

X:	2	4	6	10	15
Y:	8	10	18	22	26

Here the change in the value of X does not being the same proportionate change in the value of Y

Degrees of correlation:

Correlation exists in various degrees

1. **Perfect positive correlation**

If an increase in the value of one variable is followed by the same proportion of increase in other related variable or if a decrease in the value of one variable is followed by the same proportion of decrease in other related variable, it is perfect positive correlation. eg: if 10% rise in price of a commodity results in 10% rise in its supply, the correlation is perfectly positive. Similarly, if 5% full in price results in 5% fall in supply, the correlation is perfectly positive.

2. **Perfect Negative correlation**

If an increase in the value of one variable is followed by the same proportion of decrease in other related variable or if a decrease in the value of one variable is followed by the same proportion of increase in other related variably it is Perfect Negative Correlation. For example if 10% rise in price results in 10% fall in its demand the correlation is perfectly negative. Similarly if 5% fall in price results in 5% increase in demand, the correlation is perfectly negative.

3. **Limited Degree of Positive correlation**:

When an increase in the value of one variable is followed by a non-proportional increase in other related variable, or when a decrease in the value of one variable is followed by a non-proportional decrease in other related variable, it is called limited degree of positive correlation.

For example, if 10% rise in price of a commodity results in 5% rise in its supply, it is limited degree of positive correlation. Similarly if 10% fall in price of a commodity results in 5% fall in its supply, it is limited degree of positive correlation.

4. **Limited degree of Negative correlation**

When an increase in the value of one variable is followed by a non-proportional decrease in other related variable, or when a decrease in the value of one variable is followed by a non-proportional increase in other related variable, it is called limited degree of negative correlation.

For example, if 10% rise in price results in 5% fall in its demand, it is limited degree of negative correlation. Similarly, if 5% fall in price results in 10% increase in demand, it is limited degree of negative correlation.

5. **Zero Correlation (Zero Degree correlation)**

If there is no correlation between variables it is called zero correlation. In other words, if the values of one variable cannot be associated with the values of the other variable, it is zero correlation.

Methods of measuring correlation

Correlation between 2 variables can be measured by graphic methods and algebraic methods.

I Graphic Methods

1) Scatter Diagram

2) Correlation graph

II Algebraic methods (Mathematical methods or statistical methods or Co-efficient of correlation methods):

1) Karl Pearson’s Co-efficient of correlation

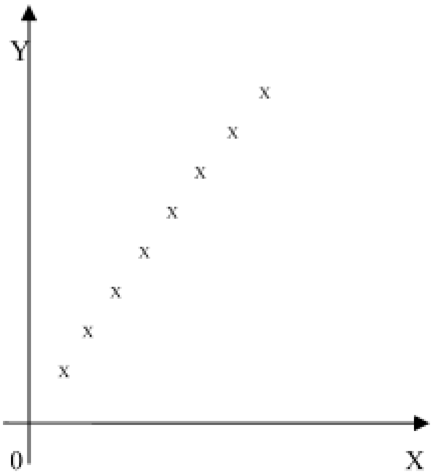
2) Spear man’s Rank correlation method

3) Concurrent deviation method

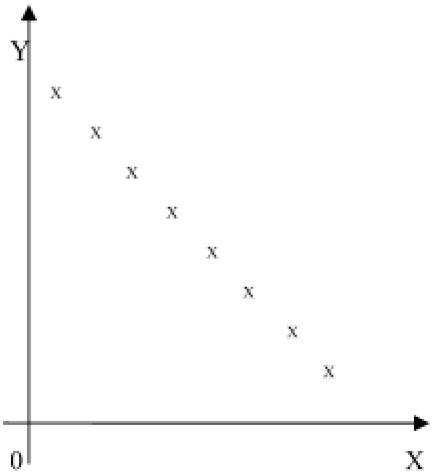
Scatter Diagram

This is the simplest method for ascertaining the correlation between variables. Under this method all the values of the two variable are plotted in a chart in the form of dots. Therefore, it is also known as dot chart. By observing the scatter of the various dots, we can form an idea that whether the variables are related or not.

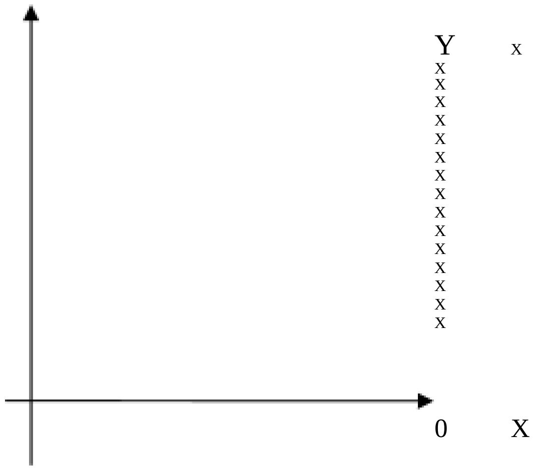
A scatter diagram indicates the direction of correlation and tells us how closely the two variables under study are related. The greater the scatter of the dots, the lower is the relationship



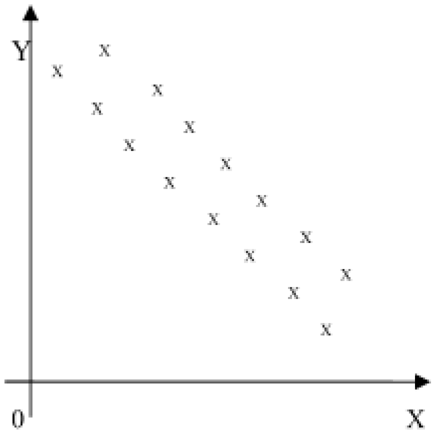
Perfect Positive Correlation



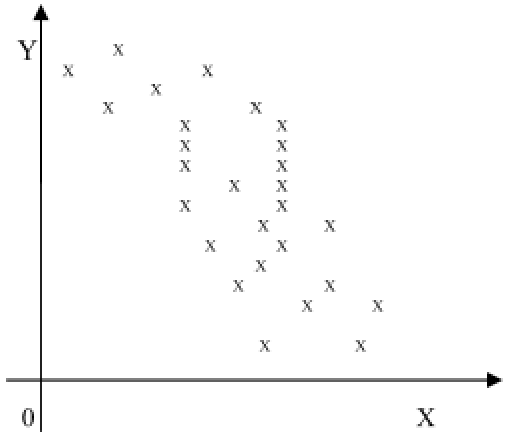
Perfect Negative Correlation



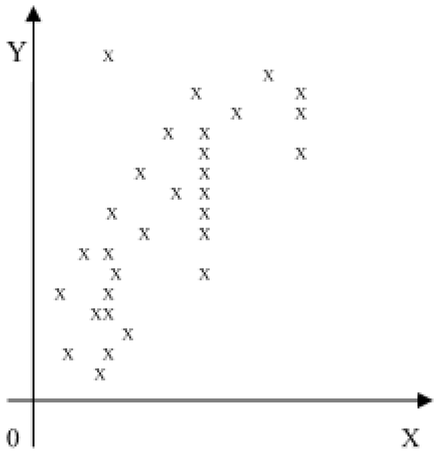
High Degree of Positive Correlation



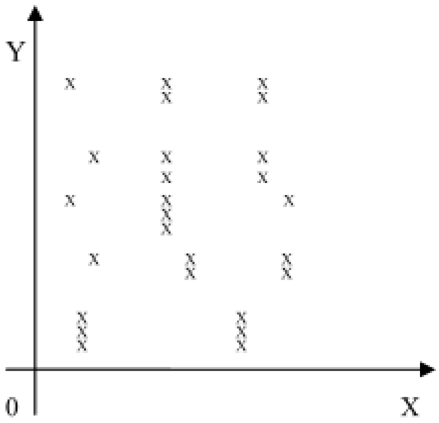
High Degree of Negative Correlation



Low Degree of Positive Correlation



Low Degree of Negative Correlation



No Correlation (r = 0)

Merits of Scatter Diagram method

1. It is a simple method of studying correlation between variables.
2. It is a non-mathematical method of studying correlation between the variables. It does not require any mathematical calculations.
3. It is very easy to understand. It gives an idea about the correlation between variables even to a layman.
4. It is not influenced by the size of extreme items.
5. Making a scatter diagram is, usually, the first step in investigating the relationship between two variables.

Demerits of Scatter diagram method

1. It gives only a rough idea about the correlation between variables.
2. The numerical measurement of correlation co-efficient cannot be calculated under this method.
3. It is not possible to establish the exact degree of relationship between the variables.

Correlation graph Method

Under correlation graph method the individual values of the two variables are plotted on a graph paper. Then dots relating to these variables are joined separately so as to get two curves. By examining the direction and closeness of the two curves, we can infer whether the variables are related or not. If both the curves are moving in the same direction(either upward or downward) correlation is said to be positive. If the curves are moving in the opposite directions, correlation is said to be negative.

Merits of Correlation Graph Method

1. This is a simple method of studying relationship between the variable
2. This does not require mathematical calculations.
3. This method is very easy to understand

Demerits of correlation graph method:

1. A numerical value of correlation cannot be calculated.
2. It is only a pictorial presentation of the relationship between variables.
3. It is not possible to establish the exact degree of relationship between the variables.

Karl Pearson’s Co-efficient of Correlation

Karl Pearson’s Coefficient of Correlation is the most popular method among the algebraic methods for measuring correlation. This method was developed by Prof. Karl Pearson in 1896. It is also called product moment correlation coefficient.

$$X(s-\bar{s})(y-\bar{y})$$

$$r = fX(x-\bar{x})^2X(y-\bar{y})$$

Interpretation of Co-efficient of Correlation

Pearson’s Co-efficient of correlation always lies between +1 and -1. The following general rules will help to interpret the Co-efficient of correlation:

1. When r - +1, It means there is perfect positive relationship between variables.
2. When r = -1, it means there is perfect negative relationship between variables.
3. When r = 0, it means there is no relationship between the variables.
4. When ‘r’ is closer to +1, it means there is high degree of positive correlation between variables.
5. When ‘r’ is closer to – 1, it means there is high degree of negative correlation between variables.
6. When ‘r’ is closer to ‘O’, it means there is less relationship between variables.

Properties of Pearson’s Co-efficient of Correlation

1. If there is correlation between variables, the Co-efficient of correlation lies between +1 and -1.
2. If there is no correlation, the coefficient of correlation is denoted by zero (ie r=0)
3. It measures the degree and direction of change
4. If simply measures the correlation and does not help to predict cansation.
5. It is the geometric mean of two regressions co-efficient.

i.e $r = \underline{Jb_{sy} \cdot b_{yx}}$

Probable Error and Coefficient of Correlation

Probable error (PE) of the Co-efficient of correlation is a statistical device which measures the reliability and dependability of the value of co-efficient of correlation.

$$¤\ PE = 0.6745 \times \frac{1-r^2}{\sqrt{n}}$$

If the value of coefficient of correlation (r) is less than the PE, then there is no evidence of correlation.

If the value of ‘r’ is more than 6 times of PE, the correlation is certain and significant.

By adding and submitting PE from coefficient of correlation, we can find out the upper and lower limits within which the population coefficient of correlation may be expected to lie.

Uses of PE:

1) PE is used to determine the limits within which the population coefficient of correlation may be expected to lie.
2) It can be used to test whether the value of correlation coefficient of a sample is significant with that of the population

If r = 0.6 and N = 64, find out the PE and SE of the correlation coefficient. Also determine the limits of population correlation coefficient.

Merits of Pearson’s Coefficient of Correlation:-

1. This is the most widely used algebraic method to measure coefficient of correlation.
2. It gives a numerical value to express the relationship between variables
3. It gives both direction and degree of relationship between variables
4. It can be used for further algebraic treatment such as coefficient of determination coefficient of non-determination etc.
5. It gives a single figure to explain the accurate degree of correlation between two variables

Demerits of Pearson’s Coefficient of correlation

1. It is very difficult to compute the value of coefficient of correlation.
2. It is very difficult to understand

Spearman’s Rank Correlation Method

Pearson’s coefficient of correlation method is applicable when variables are measured in quantitative form. But there were many cases where measurement is not possible because of the qualitative nature of the variable. For example, we cannot measure the beauty, morality, intelligence, honesty etc in quantitative terms. However it is possible to rank these qualitative characteristics in some order.

$$(R) = 1- \frac{6XD2}{N3—N}$$

Merits of Rank Correlation method

1. Rank correlation coefficient is only an approximate measure as the actual values are not used for calculations
2. It is very simple to understand the method.
3. It can be applied to any type of data, ie quantitative and qualitative
4. It is the only way of studying correlation between qualitative data such as honesty, beauty etc.
5. As the sum of rank differences of the two qualitative data is always equal to zero, this method facilitates a cross check on the calculation.

Demerits of Rank Correlation method

1. Rank correlation coefficient is only an approximate measure as the actual values are not used for calculations.
2. It is not convenient when number of pairs (ie. N) is large
3. Further algebraic treatment is not possible.
4. Combined correlation coefficient of different series cannot be obtained as in the case of mean and standard deviation. In case of mean and standard deviation, it is possible to compute combine arithmetic mean standard deviation.

Concurrent Deviation Method:

Concurrent deviation method is a very simple method of measuring correlation. Under this method, we consider only the directions of deviations. The magnitudes of the values are completely ignored. Therefore, this method is useful when we are interested in studying correlation between two variables in a casual manner and not interested in degree (or precision).

Under this method, the nature of correlation is known from the direction of deviation in the values of variables. If deviations of 2 variables are concurrent, then they move in the same direction, otherwise in the opposite direction.

The formula for computing the coefficient of concurrent deviation is: -

$$r = {}^{\pm}J\pm \overline{\frac{(2c-N)}{N}}$$

Where N = No. of pairs of symbol

C = No. of concurrent deviations (ie, No. of + signs in ‘dx dy’ column)

Steps:

1. Every value of ‘X’ series is compared with its proceeding value. Increase is shown by ‘+’ symbol and decrease is shown by ‘-‘
2. The above step is repeated for ‘Y’ series and we get ‘dy’
3. Multiply ‘dx’ by ‘dy’ and the product is shown in the next column. The column heading is ‘dxdy’.
4. Take the total number of ‘+’ signs in ‘dxdy’ column. ‘+’ signs in ‘dxdy’ column denotes the concurrent deviations, and it is indicated by ‘C’.
5. Apply the formula:

$$r = \pm J\pm \overline{\frac{(2c-N)}{N}}$$

If 2cΣ N, then r = +ve and if 2c € N, then r = –ve.

REGRESSION ANALYSIS

Introduction:-

Correlation analysis analyses whether two variables are correlated or not. After having established the fact that two variables are closely related, we may be interested in estimating the value of one variable, given the value of another. Hence, regression analysis means to analyses the average relationship between two variables and thereby provides a mechanism for estimation or predication or forecasting.

The term ‘Regression” was firstly used by Sir Francis Galton in 1877. The dictionary meaning of the term ‘regression” is “stepping back” to the average.

Definition:

“Regression is the measure of the average relationship between two or more variables in terms of the original units of the date”.

“Regression analysis is an attempt to establish the nature of the relationship between variables-that is to study the functional relationship between the variables and thereby provides a mechanism for prediction or forecasting”.

Types of Regression:-

There are two types of regression. They are linear regression and multiple regressions.

Linear Regression:

It is a type of regression which uses one independent variable to explain and/or predict the dependent variable.

Multiple Regression:

It is a type of regression which uses two or more independent variable to explain and/or predict the dependent variable.

Regression Lines:

Regression line is a graphic technique to show the functional relationship between the two variables X and Y. It is a line which shows the average relationship between two variables X and Y.

If there is perfect positive correlation between 2 variables, then the two regression lines are winding each other and to give one line. There would be two regression lines when there is no perfect correlation between two variables. The nearer the two regression lines to each other, the higher is the degree of correlation and the farther the regression lines from each other, the lesser is the degree of correlation.

Properties of Regression lines:-

1. The two regression lines cut each other at the point of average of X and average of Y (i.e $\bar{X}$ and $\bar{Y}$)
2. When r = 1, the two regression lines coincide each other and give one line.
3. When r = 0, the two regression lines are mutually perpendicular.

Regression Equations (Estimating Equations)

Regression equations are algebraic expressions of the regression lines. Since there are two regression lines, therefore two regression equations. They are :-

1. Regression Equation of X on Y:- This is used to describe the variations in the values of X for given changes in Y.
2. Regression Equation of Y on X :- This is used to describe the variations in the value of Y for given changes in X.

Regression Equation of Y on X:-

$$Y = a + bx$$

The normal equations to compute ‘a’ and ‘b’ are: -

$$\Sigma y \quad = Na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

Regression Equation of X on Y:-

$$X = a + by$$

The normal equations to compute ‘a’ and ‘b’ are:-

$$\Sigma x \quad = Na + n\Sigma y$$

$$\Sigma xy = a\Sigma y + b\Sigma y^2$$

Properties of Regression Coefficient:

1. There are two regression coefficients. They are b_{xy} and b_{yx}
2. Both the regression coefficients must have the same signs. If one is +ve, the other will also be a +ve value.
3. The geometric mean of regression coefficients will be the coefficient of correlation. $r = f\ b_{sy.}\ b_{ys.}$
4. If x $\tilde{x}$ and $\tilde{y}$ are the same, then the regression coefficient and correlation coefficient will be the same.

Computation of Regression Co-efficient

Regression co-efficient can be calculated in 3 different ways:

1. Actual mean method

- Regression coefficient x on y (b_{xy}) = $\frac{\Sigma sy}{\Sigma y^2}$

- Regression coefficient y on x (b $_{yx}$) = $\frac{\Sigma sy}{\Sigma s^2}$

Correlation	Regression
It studies degree of relationship between variables	It studies the nature of relationship between variables
It is not used for prediction purposes	It is basically used for prediction purposes
It is basically used as a tool for determining the degree of relationship	It is basically used as a tool for studying cause and effect relationship
There may be nonsense correlation between two variables	There is no such nonsense regression
There is no question of dependent and independent variables	There must be dependent and independent variables

Module 3

MIS AND DBMS

MIS-Management Information System

- MIS provides for the identification of relevant information needs, the collection of relevant information, processing the same to become useable by the business managers, and timely dissemination of processed information to the users of information for properly managing the affairs of the enterprise.
- it is a system to convert data from internal and external into and to communicate that information, in an appropriate form, to mangers at all levels in all functions to enable them to make timely and effective decision for planning, directing and controlling the activities for which they are responsible.

Definition

Jerome Kanter defines MIS as “a system that aids management in making, carrying out and controlling decisions”

- ➢ user – machine system
- ➢ integrated system
- ➢ need for a database
- ➢ utilisation of models

Components of MIS

- data gathering
- data entry
- data transformation
- information utilisation

Characteristics of MIS

- management oriented
- management directed
- integrated concept
- common database
- avoids redundancy in data storage
- heavy planning
- subsystem concept
- common data flow
- flexibility and ease of use
- distributed data processing
- information as a resource

Importance of MIS in organization

- data processing

- decision making
- optimum use of resources
- effective communication
- planning and control
- decentralization
- coordination

Elements of MIS

- management
- information
- system

Need for MIS

- management oriented
- integrated system
- to make plans
- to achieve control
- latest information
- greater accuracy
- fulfilment of statutory obligations
- decision making
- strategic planning
- to practice management by exception

Limitations of MIS

- quality of output
- not a substitute for judgements
- lack of flexibility
- no tailor made packages
- ignoring of non-quantitative factors
- not suitable for non-programmed decisions
- costly affair
- greater chance for failures
- frequent changes in top management
- hoarding of information

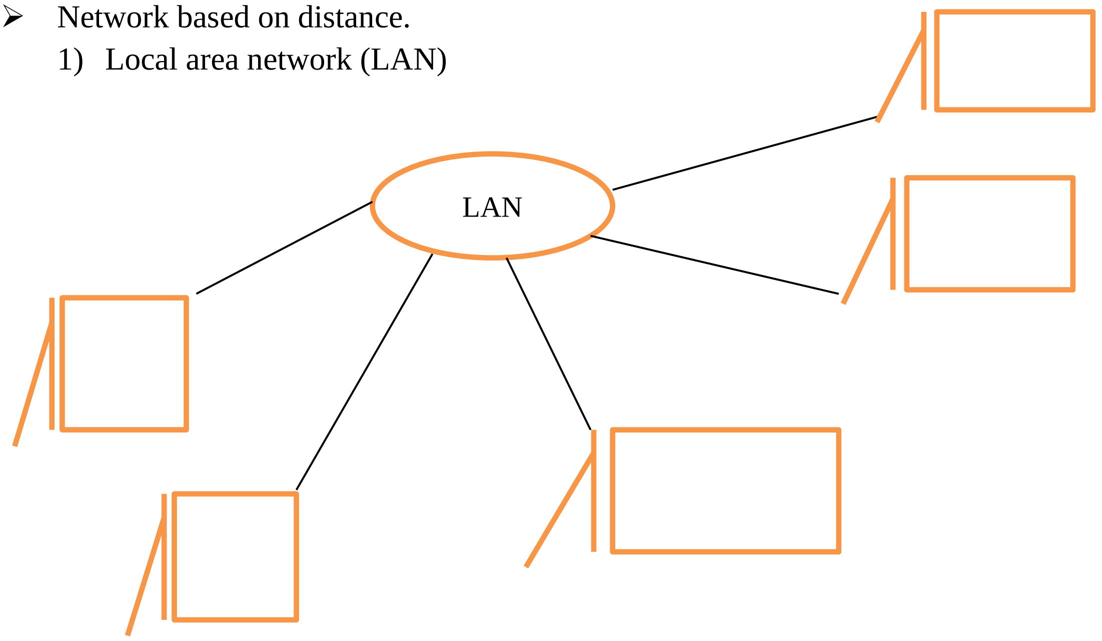
NETWORKS

- A network is a collection of computers, servers, mainframes, network devices, peripherals, or other devices connected to one another to allow the sharing of data.
- An example of a network is the internet, which connects millions of people all over the world.

Types of network

A computer network is a data communication system which interconnects computer system at various locations with the help of communication devices like hubs, routers, cables and NICs.

- Network based on distance.
  1) Local area network (LAN)



  2) Wide area network (WAN)
  3) Metropolitan area network (MAN)
  4) Personal area network (PAN)
  5) Virtual private network (VPN)
- Network based on Administration
  1) peer-to-peer network
     a) Wired Ethernet Networks
     b) wireless Ethernet Networks
     c) power line networks
  2) client server network
- Uses of Computer Network
  - Information sharing
  - sharing hardware
  - sharing software and application
  - centralized administration

- Email
- internet Relay Chat –IRC
- audio/video conferencing
- internet phone

Network Topologies

- The geometric arrangement of computer system is called Topology.
- It describes the method used to the physical configuration of cables, computer and network devices.
- the choice of topology is dependent upon:
  - type of number of equipment being used
  - planned applications and rate of data transfers
  - required response times
  - cost

Common network topologies are;

1) Bus Topology
2) Star Topology
3) Ring Topology
4) Tree Topology
5) Mesh Topology

INTRODUCTION TO DBMS

Database concept

- A database is an integrated collection of logically related records and files. It is a collection of interrelated data items that can be processed by one or more application system.
- Database is a collection of data, integrated and organised into a single comprehensive file system.
- It is designed to minimise duplication of data within that system to satisfy a wide variety of user needs
- A centrally controlled and integrated collection data is called database.
- The database is either a flat file or relational. In a flat file system all data is arrange in a single table. Relational database split the data into several tables, with each table holding some portion of total data.

Necessity of database

- Reduced data redundancy
- Reduced programming effort
- Faster response time
- Data independence
- The ability to change
- Cost reductions
- Information protection
- Multi user support and distributed processing

Characteristics of database system

- Data abstraction
- Reliability
- Efficiency

Database management system –DBMS

- A database management system is also known as database system is a collection of prewritten, integrated programs. Its major function is to assist users in all aspects of data manipulation and utilization.
- A DBMS is a software that organize data into a database, providing information storage and retrievals. It helps to access multiple databases simultaneously.
- The DBMS stores and process data so that records can be accessed through their relationship to other records.

Components of Database

- The database file
- The users
- A host language interface system
- The application programs
- Natural language interface system
- The data dictionary
- Online access and update terminals
- The output system or report generator.

Functions of DBMS

- transaction processing
- concurrency management
- recovery
- security
- data dictionary

Database administrator

- the responsibility of database administration is assigned to an individual called a database administrator (DBA)
- he is the person who is responsible for defining, updating and controlling access to database

➢ functions of database administrator
  - communicating with users
  - establishing standards and procedures
  - servicing end user requirements
  - ensuring database security and integrity
  - backup and recovery

Data Definition Language- DDL

A database management system contains two languages namely Data definition language (DDL) and Data manipulation language (DML)

- Data definition language is used to define the structure of the database.
- Structure of the database is called Schema of the database outlines the data to be included in the database. In the schema there are several fields.
- DDL establishes the connection between logical and physical structures of the database. Here logical refers to the way the user views data. On the other hand, physical refers to the way the data is physically stored.
- The DDL is used to define the physical characteristics of each record such as field name in the record, the length of each field and its data type.

➢ Functions of DDL

- Description of the schema and sub schemas
- Description of fields in each record and the logical name of the record
- Description of the data type and name of each field
- Description of the keys of record
- Provide protection to the data
- Provide physical and logical data independence

Data Manipulation Language –DML

- It is used to manipulate data in the database.
- It includes all the commands that enable the users to manipulate the data and the users can view the data, add new data, delete existing data and modify selected fields in a record.

➢ Functions of DML

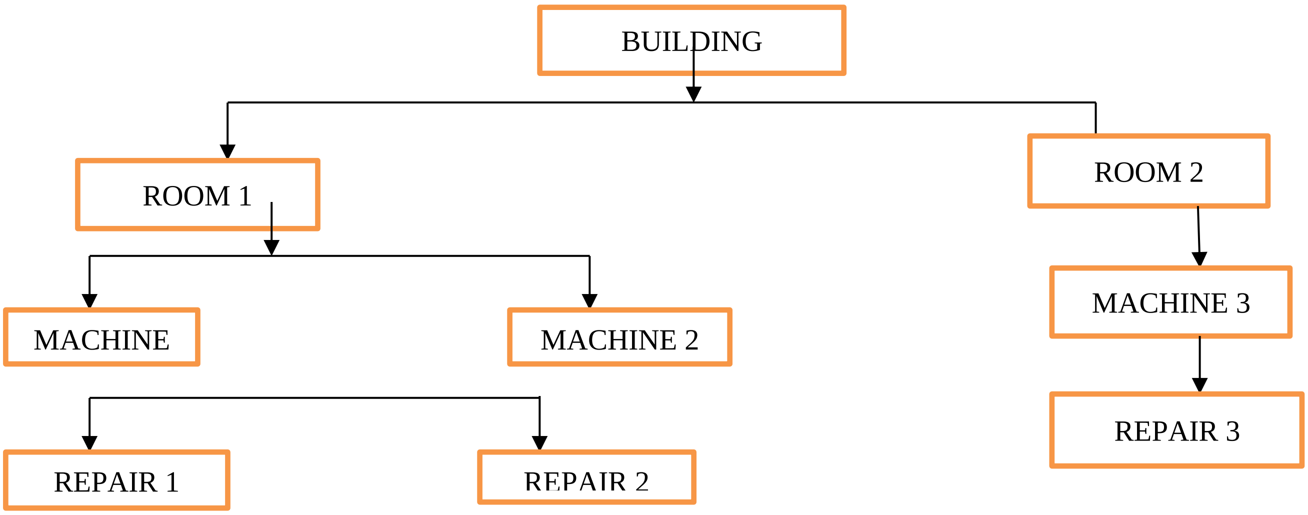
- Provide the techniques of data manipulation such as deletion, addition, retrieval of data records.
- It permits the users and application programs to process data on a symbolic logical basis rather than on physical location basis
- Provides for independence of programming languages
- Provide the relationship between different records
- It also allows the user and application programs to be independent of physical data structure and database structure maintenance.

Types of Databases

- Database models are also known as database structures or architecture.
- The structure of data refers to the view of data accessed by a user from the database.

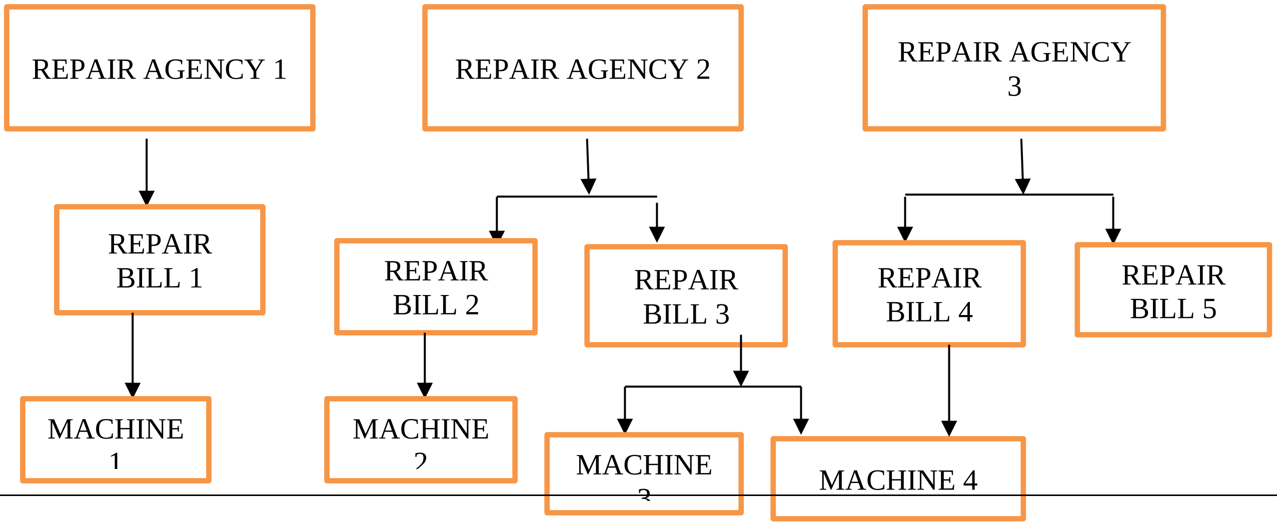
1. hierarchical database model
    - it employees a tree structure to represent relationship among data elements
    - Here, the relationship between records is that of parent and child. One record is connected to only one record at the higher level. data at the lower level (child) can be accessed through a higher level record (parent)
    - All records in hierarchy are called nodes. Each node is related to the others in a parent child relationship. Each parent record may have one or mode child records,

but no child record may have more than one parent record. Thus the hierarchical database structure implements one-to-one and one-to-many relationship.



- Advantages
  - simplicity
  - data security
  - data integrity
  - efficiency
- Disadvantages
  - implementation complexity
  - database management problems
  - Lack of structural independence.

2. network database model
   - It allows more general connections among data elements.
   - A group of interconnected node is called a network.
   - This structure allows multiple relations between data items.
   - Here records are not confined to only one superior. a record may have many superior records and subordinate records.
   - A network model allows a record to be a member of more than one set at a time. namely it shows two types of relationship namely many-to-one and many-to-many relationship



- Advantages
  - capability to handle more relationship types
  - ease of data access
  - data independence
  - data integrity
- Disadvantages
  - system complexity
  - absence of structural independence

3. Relational database model
   - It is designed just like a two dimensional table. Thus the basic structure of relational database design is the table, known as relation.
   - This is highly beneficial to the managers of business organization because they often handle financial data in tabular forms.
   - In relational database system the table row is called a tuple. The columns of a table divide each record into different data fields. These columns in a table are called attributes.
   - Here the data stored in different tables can be related so long as these tables shares common data elements. Moreover information in different files can be taken and combined into a new table.

Roll no	Name	Mark	Average
501	Soorya	534	90
502	John	456	75

- advantages
  - structural independence
  - conceptual simplicity
  - design, implementation, maintenance and usage ease
- disadvantages
  - hardware over heads
  - Ease of design can lead to bad design.

Limitations of database

- concurrency problems
- ownership problems
- more resources required
- security problems