Directorate of Distance Education

UNIVERSITY OF JAMMU

JAMMU

SELF LEARNING MATERIAL

PGDBM - II[nd] SEMESTER

PAPER : IV **UNIT : I-V**

SUBJECT : QUANTITAIVE METHODS

Course Co-ordinator
Rohini Gupta Suri

http:/www.distanceeducationju.in

Printed and Published on behalf of the Directorate of Distance Education, University of Jammu, Jammu by the Director, DDE, University of Jammu, Jammu.

QUANTITAIVE METHODS

EDITED & REVIEWED
Rohini Gupta Suri
Co-ordinator

Directorate of Distance Education, University of Jammu, Jammu. 2021

Printed by Printech Printing Press /21/300

PGDBM SEMESTER-II

CourseTitle: Quantitative Methods and Research Methodology **Total Marks :100**
Internal Assessment : 20
Paper -IV **Semester Examination : 80**
Contact Hours :45
Duration of Exam: 3 hrs.

(For the Examination to be held in 2017, 2018,and 2019)

Objective

To appraise the students about the various quantitative techniques and the concept of Research Metholodogy.

UNIT-I

Introduction of Statistics , Role of statistics , Application of statistics in managerial decision-making : Measures of central tendency Mean Median , Quartiles , Deciles Percentiles and Mode

Measures of Dispersion : Range , Mean deviation, Standard deviation , Coefficient of Variation

UNIT -II

Skewness Meaning , Signifinance , Karl Pearson's Coefficient of Skewness, Bowley's Method Kelley's method, Kurtosis , Moments about mean and Moments about origin.

Time series analysis: Concept , Components of time series, Trend analysis Least Square method- Linear and Non linear equations Application in business decision- making.

UNIT -III

Meaning and Signifinance of correlation , types of correlation postive correlation , negative correlation perfect correlation linear and non linear correlation scatter diagram Karl Pearson's coefficient of correlation and Spearman rank correlation

Meaning and Importance of regression types of regression - simple and multiple regression

linear and non linear regression statement of regression lines, regression coefficients Difference between Regression and Correlation.

UNIT-IV

Introduction of Research : Concept Meaning and Significance , Its Application in Various Functions of Management, Types of Research Exploratory Research , Descriptive Research Casual Research.

Process of Research , Steps Involved in Reseach Process. Research Design, Types of Business Problems Encountered by the Researcher

UNIT-V

Concept of Measurement and Scale, Concept Sample , Sample Size and Sampling Procedure, Various Types of Sampling Techniques- Probability and Non -Probability Sampling , Types of Data: Secondary and Primary, Various Methods of Collection and Data Preparation of Questionnaire, Precautions in Preparation of Questionnaire and Collection of Data

Note for paper setting

The question Paper shall contain two questions from each Unit (Total 10 Questions) and the candidates shall be required to answer one question from each unit (total number of questions to be attempted shall be five i.e. there shall be internal choice within each unit)

SUGGESTED READINGS:

1) Business Statistics by Gupta, S.P & Gupta , M.P..Sultan Chand & Sons New Delhi
2) Statistics for Business and Economics by Chandan J S. Vikas 1998 Ist Edition
3) Research Methodology Methods & Techniques by Kothari C.R, New Age International Publishers
4) Research Methods for Business students by Saunders, Prentice hall 2nd Edition 2007
5) Business Research Methods by Cooper and Schindler , Tata Mc Graw Hill 9th Edition.

LESSON No. 1

STATISTICS **UNIT-I**

STATISTICS: ROLE, APPLICATION

STRUCTURE

1.1 INTRODUCTION

Statistics is a very broad subject, with applications in a vast number of different fields. In generally one can say that statistics is the methodology for collecting, analyzing, interpreting and drawing conclusions from information. Putting it in other words, statistics is the methodology which scientists and mathematicians have developed for interpreting and drawing conclusions from collected data. Everything that deals even remotely with the collection, processing, interpretation and presentation of data belongs to the domain of statistics, and so does the detailed planning of that precedes all these activities.

Furthermore, statistics is the science of dealing with uncertain phenomenon and events. Statistics in practice is applied successfully to study the effectiveness of medical treatments, the reaction of consumers to television advertising, the attitudes of young people toward sex and marriage, and much more. It's safe to say that nowadays statistics is used in every field of science.

Statistics consists of a body of methods for collecting and analyzing data (Agresti & Finlay, 1997).

From above, it should be clear that statistics is much more than just the tabulation of numbers and the graphical presentation of these tabulated numbers. Statistics is the science of gaining information from numerical and categorical data. Statistical methods can be used to find answers to the questions like:

o What kind and how much data need to be collected?

o How should we organize and summarize the data?

o How can we analyse the data and draw conclusions from it?

o How can we assess the strength of the conclusions and evaluate their uncertainty?

That is, statistics provides methods for

1. Design: Planning and carrying out research studies.

2. Description: Summarizing and exploring data.

3. Inference: Making predictions and generalizing about phenomena represented by the data.

There are two major types of statistics. The branch of statistics devoted to the summarization and description of data is called descriptive statistics and the branch of statistics concerned with using sample data to make an inference about a population of data is called inferential statistics.

Descriptive Statistics: Descriptive statistics consist of methods for organizing and summarizing information (Weiss, 1999). Descriptive statistics includes the construction of graphs, charts, and tables, and the calculation of various descriptive measures such as averages, measures of variation, and percentiles. In fact, the most part of this course deals with descriptive statistics.

Inferential Statistics: Inferential statistics consist of methods for drawing and measuring the reliability of conclusions about population based on information obtained from a sample of the population. (Weiss, 1999). Inferential statistics includes methods like point estimation, interval estimation and hypothesis testing which are all based on probability theory.

1.2 OBJECTIVES

After studying this chapter students should be able to understand:

- ✓ Understand statistics and its role in business?
- ✓ Explain the applications of statistics in managerial decision making?
- ✓ Understand central tendency is and how it is measured.
- ✓ Outline the various methods to measure central tendency?
- ✓ What is dispersion and what are various methods of dispersion.
- ✓ Which measure of central tendency is most affected by the outlier?
- ✓ Is standard deviation is a measure of central tendency?
- ✓ Comprehend the purpose of obtaining a measure of central tendency?

1.3 Role of Statistics: Applications of Statistics in Managerial Decision Making

There are three major functions in any business enterprise in which the statistical methods are useful. These are as follows:

(i) The planning of operations: This may relate to either special projects or to the recurring activities of a firm over a specified period.

(ii) The setting up of standards: This may relate to the size of employment, volume of sales, fixation of quality norms for the manufactured product, norms for the daily output, and so forth.

(iii) The function of control: This involves comparison of actual production achieved against the norm or target set earlier. In case the production has fallen short of the target, it gives remedial measures so that such a deficiency does not occur again.

A worth noting point is that although these three functions-planning of operations, setting standards, and control-are separate, but in practice they are very much interrelated. Different authors have highlighted the importance of Statistics in business. For instance, Croxton and Cowden give numerous uses of Statistics in business such as project planning, budgetary planning and control, inventory planning and control, quality control, marketing, production and personnel administration. Within these also they have specified certain areas where Statistics is very relevant.

Another author, Irwing W. Burr, dealing with the place of statistics in an industrial organisation, specifies a number of areas where statistics is extremely useful. These are: customer wants and market research, development design and specification, purchasing, production, inspection, packaging and shipping, sales and complaints, inventory and maintenance, costs, management control, industrial engineering and research.

Statistical problems arising in the course of business operations are multitudinous. As such, one may do no more than highlight some of the more important ones to emphasis the relevance of statistics to the business world. In the sphere of production, for example, statistics can be useful in various ways.

Statistical quality control methods are used to ensure the production of quality goods.

Identifying and rejecting defective or substandard goods achieve this. The sale targets can be fixed on the basis of sale forecasts, which are done by using varying methods of forecasting. Analysis of sales affected against the targets set earlier would indicate the deficiency in achievement, which may be on account

of several causes: (i) targets were too high and unrealistic (ii) salesmen's performance has been poor (iii) emergence of increase in competition (iv) poor quality of company's product, and so on. These factors can be further investigated.

Another sphere in business where statistical methods can be used is personnel management. Here, one is concerned with the fixation of wage rates, incentive norms and performance appraisal of individual employee. The concept of productivity is very relevant here. On the basis of measurement of productivity, the productivity bonus is awarded to the workers. Comparisons of wages and productivity are undertaken in order to ensure increases in industrial productivity.

Statistical methods could also be used to ascertain the efficacy of a certain product, say, medicine. For example, a pharmaceutical company has developed a new medicine in the treatment of bronchial asthma. Before launching it on commercial basis, it wants to ascertain the effectiveness of this medicine. It undertakes an experimentation involving the formation of two comparable groups of asthma patients. One group is given this new medicine for a specified period and the other one is treated with the usual medicines. Records are maintained for the two groups for the specified period. This record is then analysed to ascertain if there is any significant difference in the recovery of the two groups. If the difference is really significant statistically, the new medicine is commercially launched.

The other important roles played by statistics are as:

(1) Statistics helps in providing a better understanding and exact description of a phenomenon of nature.

(2) Statistical helps in proper and efficient planning of a statistical inquiry in any field of study.

(3) Statistical helps in collecting an appropriate quantitative data.

(4) Statistics helps in presenting complex data in a suitable tabular, diagrammatic and graphic form for an easy and clear
comprehension of the data.

(5) Statistics helps in understanding the nature and pattern of variability of a phenomenon through quantitative observations.

Applications of Statistics in Managerial Decision Making:

(a) Statistics and business: Statistics is an aid to business and commerce. When a person enters business, he enters into the profession of fore casting. Modern statistical devices have made business forecasting more precise and accurate. A business man needs statistics right from the time he proposes to start business. He should have relevant fact and figures to prepare the financial plan of the proposed business. Statistical methods are necessary for these purposes. In industrial concern statistical devices are being used not only to determine and control the quality of products manufactured by also to reduce wastage to a minimum. The technique of statistical control is used to maintain quality of products.

A manager in a business organization - whether in the top level, or the middle level, or the bottom level - has to perform an important role of decision making. For solving any organizational problem - which most of the times happens to be complex in nature -, he has to identify a set of alternatives, evaluate them and choose the best alternative. The experience, expertise, rationality and wisdom gained by the manager over a period of time will definitely stand in good stead in the evaluation of the alternatives available at his disposal. He has to consider several factors, sometimes singly and sometimes jointly, during the process of decision making. He has to deal with the data of not only his organization but also of other competing organizations.

It would be a challenging situation for a manager when he has to face so many variables operating simultaneously, something internal and something external. Among them, he has to identify the important variables or the dominating factors and he should be able to distinguish one factor from the other. He should be able to find which factors have similar characteristics and which factors stand apart. He should be able to know which factors have an inter play with each other and which factors remain independent. It would be advantageous to him to know whether there is any clear pattern followed by the variables under consideration. At times he may be required to have a good idea of the values that the variables would assume in future occasions. The task of a manager becomes all the more difficult in view of the risks and uncertainties surrounding

the future events. It is imperative on the part of a manager to understand the impact of various policies and programmes on the development of the organization as well as the environment. Also he should be able to understand the impact of several of the environmental factors on his organization. Sometimes a manager has to take a single stage decision and at times he is called for to take a multistage decision on the basis of various factors operating in a situation.

Statistical analysis is a tool for a manager in the process of decision making by means of the data on hand. All managerial activities involve an analysis of data. Statistical approach would enable a manager to have a scientific guess of the future events also. Statistical methods are systematic and built by several experts on firmly established theories and consequently they would enable a manager to overcome the uncertainties associated with future occasions. However, statistical tools have their shortcomings too. The limitations do not reflect on the subject. Rather they shall be traced to the methods of data collection and recording of data. Even with highly sophisticated statistical methods, one may not arrive at valid conclusions if the data collected are devoid of representative character.

In any practical problem, one has to see whether the assumptions are reasonable or not, whether the data represents a wide spectrum, whether the data is adequate, whether all the conditions for the statistical tests have been fulfilled, etc. If one takes care of these aspects, it would be possible to arrive at better alternatives and more reliable solutions, thereby avoiding future shocks. While it is true that a statistical analysis, by itself, cannot solve all the problems faced by an organization, it will definitely enable a manager to comprehend the ground realities of the situation. It will for sure provide a foresight in the identification of the crucial variables and the key areas so that he can locate a set of possible solutions within his ambit. A manager has to have a proper blend of the statistical theories and practical wisdom and he shall always strive for a holistic approach to solve any organizational problem. A manager has to provide some safe-guarding measures against the limitations of the statistical tools. In the process he will be able to draw valid inferences thereby providing a clue as to the direction in which the organization shall move in future. He will be ably guided by the statistical results in the formulation of appropriate strategies for the organization.

Further, he can prepare the organization to face the possible problems of business fluctuations in future and minimize the risks with the help of the early warning signals indicated by the relevant statistical tools.

A marketing manager of a company or a manager in a service organization will have occasions to come across the general public and consumers with several social and psychological variables which are difficult to be measured and quantified.

Depending on the situation and the requirement, a manager may have to deal with the data of just one variable (univariate data), or data on two variables (bivariate data) or data concerning several simultaneous variables (multivariate data).

(b) Statistics and Research: Statistics is an indispensable tool of research. Most of the advancement in knowledge has taken place because of experiments conducted with the help of statistical methods. For example, experiments about crop yield and different types of fertilizers and different types of soils of the growth of animals under different diets and environments are frequently designed and analysed according to statistical methods. Statistical methods are also useful for the research in medicine and public health. In fact there is hardly any research work today that one can find complete without statistical data and statistical methods.

(c) Statistics and Economics: In the year 1890, Prof. Alfred Marshall, the renowned economist observed that "statistics are the straw out of which I, like every other economist, have to make bricks". This proves the significance of statistics in economics. Economics is concerned with production and distribution of wealth as well as with the complex institutional set-up connected with the consumption, saving and investment of income. Statistical data and statistical methods are of immense help in the proper understanding of the economic problems and in the formulation of economic policies. In fact these are the tools and appliances of an economists laboratory. In the field of economics it is almost impassible to find a problem which does not require an extensive uses of statistical data. As economic theory advances use of statistical methods also increase. The laws of economics like law of demand, law of supply etc can be considered true

and established with the help of statistical methods. Statistics of consumption tells us about the relative strength of the desire of a section of people. Statistics of production describe the wealth of a nation. Exchange statistics through light on commercial development of a nation. Distribution statistics disclose the economic conditions of various classes of people. Therefore, statistical methods are necessary for economics also.

1.4 Measures of Central Tendency

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics. The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode. The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others.

In statistics, a central tendency (or, more commonly, a measure of central tendency) is a central or typical value for a distribution. It may also be called a centre or location of the distribution. Colloquially, measures of central tendency are often called averages. The term central tendency dates from the late 1920s. The central tendency of a distribution is typically contrasted with its dispersion or variability; dispersion and central tendency are the often characterized properties of distributions. Analysts may judge whether data has a strong or a weak central tendency based on its dispersion.

Central tendency is defined as "the statistical measure that identifies a single value as representative of an entire distribution." It aims to provide an accurate description of the entire data. It is the single value that is most typical representative of the collected data. The term "number crunching" is used to illustrate this aspect of data description. The mean, median and mode are the three commonly used measures of central tendency.

Measures of central tendency are a combination of two words i.e. 'measure' and 'central tendency'. Measure means methods and central tendency

means average value of any statistical series. Thus we can say that central tendency means the methods of finding out the central value or average value of a statistical series of quantitative information. A measure of central tendency is a measure that tells us where the middle of a bunch of data lies. A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics. The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode.

J.P. Guilford has pointed out that "an average is a central value of a group of observations or individuals."

According to Clark "average is an attempt to find one single figure to describe whole of figure."

In the words of A.E. Waugh "an average is a single value selected from a group of values to represent them in a same way-a value which is supposed to stand for whole group of which it is a part, as typical of all the values in the group."

Thus it can be said that an average or central tendency is a single figure that is computed from a given distribution to give a central idea about the entire series. The value of the average lies within the maximum and minimum value in the series.

1.4.1 Mean

Mean is the most common measure of central tendency. It is simply the sum of the numbers divided by the number of numbers in a set of data. This is also known as average. For a common man, average means the arithmetic mean. It is most popularly used because of its simplicity, rigidity etc. The arithmetic mean is commonly known as mean. It is a measure of central tendency because other figures of the data congregate around it. Arithmetic mean is obtained by dividing the sum of the values of all observations in the given data set by the number of observations in that set. It is the most commonly used statistical average

in the disciplines such as commerce, management, economics, finance, production, etc. The arithmetic mean is also called as simple Arithmetic Mean.

An arithmetic average is defined as the "quotient obtained by dividing the total of the values of a variable by the total number of their observations or items." II.E. Garett (1985) defines "The arithmetic mean or more simply the mean is the sum of the separate scores or measures divided by their number."

Methods of Calculating Mean:

As you know, the collected data is classified by arranging into different classes or groups on the basis of their similarities and resemblances. Arithmetic mean can be computed for the unclassified or ungrouped 'data (raw data) as well as classified or grouped data. But the methods of computation are different. Now let us understand the methods of computing the arithmetic mean for unclassified data and classified data. Normally, arithmetic mean is denoted by x? which is read as 'X bar'

Ungrouped Data

Method 1: Computation of arithmetic mean is very simple when the data is ungrouped, i.e. when frequency distribution is not done. Just add all the values of the observations and divide it by the number of observations. This can be explained and expressed in the form of a formula as follows:

$$\bar{x} = \frac{XI + X2 + ...+ Xn}{n}$$

Where $\bar{x}$ (X bar) is the arithmetic mean of the variable x and x1, x2..... xn are the various values of the variable x and n is the number of observations. This formula can be simplified as follows:

$$\bar{x} = \frac{\sum X}{n}$$

Where the $\sum$ (read it as sigma) is the Greek symbol denoting the summation over all values of x.

Illustration 1:

The grocery store sells five different products. The profit per unit on the sales of each of these products is given below. Find out the average profit.

Product 1 - Rs. 4

Product 2 - Rs. 9

Product 3 - Rs. 6

Product 4 - Rs. 2

Product 5 - Rs. 9

Solution:

Average profit can be computed as follows:

$$\bar{x} = \frac{\sum X}{n}$$

$$\bar{x} = \frac{4+9+6+2+9}{5}$$

$$\bar{x} = \frac{30}{5}$$

$$\bar{x} = 6$$

Method 2: When the values of the observations the given data are too large or they are in fractions, this method may be followed. This method is based on the fact that the algebraic sum of the deviations of a series of individual observations from their mean is always equal to zero. For example, the arithmetic mean of 8, 14, 16, 12 and 20 is 14. The difference of each of these items from the mean would be -6, 0, +2, - 2, +6 and their total is zero. This is true always. To compute arithmetic mean under this method, the following steps are to be followed.

1) Assume any arbitrary mean (A) to find out the deviations of items from their assumed mean.

2) Compute the deviation (d) of each individual value (x) from the assumed mean i.e., d=x-A.

3) Obtain the sum of all deviations ($\sum d$ called sigma d)

4) Compute the arithmetic mean by using the following formula:

$$\bar{x} = A + \frac{\sum d}{n}$$

Where $\bar{x}$ is the arithmetic mean of the variable x

A is the assumed mean

$\sum d$ is the sum total of the deviations of each individual value from the assumed mean n is the number of observations

Illustration 2: Monthly sales of scooters of 10 dealers are presented below. Calculate the average sales per month:

Dealer	Sales(x)	d=x-A
1	23	-2
2	8	-17
3	14	-11
4	31	6
5	6	-19
6	28	3
7	11	-14
8	27	2
9	32	7
10	46	21
n=10		$\sum d=24$

Assumed mean A =25

n = 10

$\sum d = 24$

$$\bar{x} = A + \frac{\sum d}{n}$$

$$= 25 + \frac{-24}{10}$$

$\bar{x}$ = 25-2.5

$\bar{x}$ = 22.5

Average scooters sold = 22.5

Grouped Data:

Variables can be categorised as discrete variables and continuous variables. The frequency distribution prepared for discrete variable is called discrete distribution and the frequency distribution prepared for continuous variable is called continuous distribution. Methods of computing arithmetic mean for these two types of distributions are different. Now let us study these methods

Arithmetic Mean for Discrete Series:

Method 1: Under this method the mean for grouped data can be obtained by using the following formula:

$$\bar{x} = \frac{f_1 x_1 + f_2 \bar{x}_2 + \ldots\ldots + f_n x_n}{f_1 + f_2 + \ldots\ldots + f_n} = \frac{\Sigma fx}{\Sigma f}$$

Where x_1, x_2, x_3, etc., refer to the values of the variable in classes 1, 2, 3 etc., respectively. Similarly, f1, f2, f3, etc., refer to the frequency of classes 1, 2, 3 etc., respectively. Here f_1x_1 indicates the multiplication of the frequency of the first class (f) by the value of the variable in that class (x), $f_1 x_1$,$f_2 x_2$,. ... ,fn_{xx} indicate the same meaning. Similarly, $\sum f$ is the sum total of f to f_n.

Method 2:

When the number of classes in the given frequency distribution is large, this method is preferred. The procedure followed in this method is almost the same as it is for ungrouped data. Steps to be followed in this method are as follows:

a) Take an assumed mean A.

b) Find the deviations of the variable x from the assumed mean and denote it by d = x - A, Any value can be taken as an assumed mean, but the value of variable x in centrally located class of the given distribution should be chosen

c) Obtain $\sum fd$ by multiplying deviations (d) with their respective class

frequencies (f) and summing it.

d) Take a ratio of ∑fd to ∑f, that is, $\frac{\sum fd}{\sum f}$ is also called a correction factor **(Cf)**.

e) Add this correction factor to the assumed mean to obtain x?

The formula used in computing the arithmetic mean under this method is as follows:

$$\bar{x} = A + \frac{\Sigma fd}{\Sigma f} \quad \text{or} \quad \bar{x} = A + \frac{\Sigma fd}{n}$$

Where A is the assumed mean

∑f denotes the total number of items which can also be denoted by 'n'.

∑fd is the sum total of the deviations (d = x - A) multiplied with their respective class frequencies.

Now let us take an illustration and study how arithmetic mean is computed under these two methods.

Illustration 3:

Calculate the arithmetic mean for the following data by using the two methods:

Marks :	10	20	30	40	50	60	70	80
No of Students :	8	21	23	17	15	9	5	2

Solution.

Calculation of Arithmetic Mean

Marks(x)	**No. of students(f)**	**d= x- 40**	**fd**	**fx**
10	8	-30	-240	80
20	21	-20	-420	420
30	23	-10	-230	690
40	17	0	0	680

50	15	10	150	750
60	9	20	180	540
70	5	30	150	350
80	2	40	80	160
Total	∑f = 100		∑fd = -330	∑fx = 3670

In this case assumed mean (A) is 40.

Method 1:

$$x? = \frac{\sum fx}{\sum f}$$

$$= \frac{3670}{100}$$

$$= 36.70$$

Method 2:

$$x? = A + \frac{\sum fd}{\sum f}$$

$$= 40 + \frac{-330}{100}$$

$$= 40 - 3.30$$

$$= 36.70$$

Combined Mean

The separate means of a number of different series can produce the combined arithmetic mean of all the different series when number of items in each of such series is given. This is calculated by the following formula when the number of groups is n.

$$\mathbf{M_{comb} = \frac{N_1M_1 + N_2M_2 + N_3M_3 + ... + N_nM_n}{N_1 + N_2 + N_3 + ... + N_n}} \quad \text{.... 8.5}$$

Illustration:

Below are given the mean of VI class students of 4 schools. What is the mean of VI class students in general?

We can find out combined mean by applying formula:

$$M_{comb} = \frac{N_1M_1 + N_2M_2 + N_3M_3 + N_4M_4}{N_1 + N_2 + N_3 + N_4}$$

$$= \frac{50 \times 60.53 + 45 \times 55.65 + 60 \times 58 + 30 \times 40.5}{50 + 45 + 60 + 30}$$

$$= \frac{3026.5 + 2497.5 + 3480 + 1216.5}{185}$$

$$= \frac{10220.5}{185} = 55.25$$

So the mean of all the VI class students is 55.25.

Weighted Mean

Simple arithmetic mean gives equal importance to all items. Some times the items in a series may not have equal importance. So the simple arithmetic mean is not suitable for those series and weighted average will be appropriate.

Weighted means are obtained by taking in to account these weights (or importance). Each value is multiplied by its weight and sum of these products is divided by the total weight to get weighted mean.

Weighted average often gives a fair measure of central tendency. In many cases it is better to have weighted average than a simple average. It is invariably used in the following circumstances.

- When the importance of all items in a series are not equal. We associate weights to the items.
- For comparing the average of one group with the average of an other group, when the frequencies in the two groups are different, weighted averages are used.
- When rations percentages and rates are to be averaged, weighted averages

is used.

- It is also used in the calculations of birth and death rate index number etc.
- When average of a number of series is to be found out together weighted average is used.

Formula:

Let $x_1 + x_2 + x_3 - - - - + x_4$ be in values with corresponding weights $w_1 + w_2 + w_3 - - - - + w_4$. Then the weighted average is

$$= \frac{w1x1+w2x2+w3x3+.....wnxn}{w1+w2+w3+.....wn}$$

$$= \frac{\sum wx}{\sum w}$$

Geometric Mean

The geometric mean is a type of mean or average, which indicates the central tendency or typical value of a set of numbers. It is similar to the arithmetic mean, which is what most people think of with the word "average", except that the numbers are multiplied and then the n^{th} root (where n is the count of numbers in the set) of the resulting product is taken.

Geometric mean is defined as the n^{th} root of the product of N items of series. If there are two items, take the square root; if there are three items, we take the cube root; and so on.

Symbolically;

$$GM = \sqrt[n]{(x1)(x2)(x3).....(x4)}$$

Where X_1, X_2 ….. X_n refer the various items of the series.

For instance, the Geometric mean of two numbers, say 2 and 8, is just the square root of their product; that is

$$GM = \sqrt[2]{4 \times 8} = 4$$

As another example, the geometric mean of three numbers 1, ½, ¼ is the cube

root of their product (1/8), which is 1/2; that is

$$GM = \sqrt[3]{1 \times \frac{1}{2} \times \frac{1}{4}}$$

$$GM = \sqrt[3]{\frac{1}{8}}$$

GM = ½

When the number of items are three or more, the task of multiplying the numbers and of extracting the root becomes excessively difficult. To simplify calculations, logarithms are used.

GM then is calculated as follows

$$\text{Log GM} = \frac{\log X1 + \log X2 + \ldots\ldots\ldots\ldots \log Xn}{N}$$

$$\text{Log GM} = \Sigma \frac{\log X}{N}$$

$$GM = \text{Antilog } \Sigma \frac{\log X}{N}$$

$$\text{In discrete series GM} = \text{Antilog } \Sigma \frac{f \log X}{N}$$

$$\text{In continuous series GM} = \text{Antilog } \Sigma \frac{f \log m}{N}$$

Where f = frequency

m = mid point

Merits of GM

1. It is based on each and every item of the series.
2. It is rigidly defined.
3. It is useful in averaging ratios and percentages and in determining rates of

increase and decrease.

4. It is capable of algebraic manipulation.

Limitations of GM:

1. It is difficult to understand
2. It is difficult to compute and to interpret
3. It can't be computed when there are negative and positive values in a series or one or more of values are zero.
4. GM has very limited applications.

Uses of Mean:

There are certain general rules for using mean. Some of these uses are as following:

1. Mean is the centre of gravity in the distribution and each score contributes to the determination of it when the spread of the scores are symmetrically around a central point.
2. Mean is more stable than the median and mode. So that when the measure of central tendency having the greatest stability is wanted mean is used.
3. Mean is used to calculate other statistics like standard deviation, coefficient of correlation, ANOVA, ANCOVA etc.

Merits of Mean:

1. Mean is rigidly defined so that there is no question of misunderstanding about its meaning and nature.
2. It is the most popular central tendency as it is easy to understand.
3. It is easy to calculate.
4. It includes all the scores of a distribution.
5. It is not affected by sampling so that the result is reliable.
6. Mean is capable of further algebraic treatment so that different other statistics like dispersion, correlation, skewness requires mean for calculation.

Demerits of Mean:

1. Mean is affected by extreme scores.
2. Sometimes mean is a value which is not present in the series.
3. Sometimes it gives absurd values. For example there are 41, 44 and 42 students in class VIII, IX and X of a school. So the average students per class are 42.33. It is never possible.
4. In case of open ended class intervals, it cannot be calculated without assuming the size of the open end classes.

1.4.2 Median

The median is also a frequently used measure of central tendency. The median is the midpoint of a distribution: the same numbers of data points are above the median as below it. The median is the middle score for a set of data that has been arranged in order of magnitude.

The median is determined by sorting the data set from lowest to highest values and taking the data point in the middle of the sequence. There is an equal number of points above and below the median. For example, in the data 7,8,9,10,11, the median is 9; there are two data points greater than this value and two data points less than this value. Thus to find the median, we arrange the observations in order from smallest to largest value. If there is an odd number of an observation, the median is the middle value.

If there is an even number of observations, the median is the average of the two middle values. Thus, the median of the numbers 2, 4, 7, 12 is (4+7)/2 = 5.5.

In certain situations the mean and median of the distribution will be the same, and in some situations it will be different. For example, in the data 1, 2, 3, 4, 5 the median is 3; there are two data points greater than this value and two data points less than this value. In this case, the median is equal to the mean. But consider the data 1, 2, 3, 4, 10. In this dataset, the median still is three, but the mean is equal to 4.

The median can be determined for ordinal data as well as interval and ratio data. Unlike the mean, the median is not influenced by outliers at the

extremes of the data set. For this reason, the median often is used when there are a few extreme values that could greatly influence the mean and distort what might be considered typical. For data which is much skewed, the median often is used instead of the mean.

Calculation of Median: Discrete series

Steps:

- Arrange the date in ascending or descending order
- Find cumulative frequencies
- Apply the formula Median

$$\text{Median} = \text{Size of } \left[\frac{N+1}{2}\right]^{th} \text{ item}$$

Example: calculate median from the following

Size of shoes:	5	5.5	6	6.5	7	7.5	8
Frequency	10	16	28	15	30	40	34

Solution:

Size	f	Cumulative f (f)
5	10	10
5.5	16	26
6	28	54
6.5	15	69
7	30	99
7.5	40	139
8	34	173

$$\text{Median} = \text{Size of } \left[\frac{N+1}{2}\right]^{th} \text{ item}$$

N = 173

$$\text{Median} = \frac{173+1}{2} = 87^{th} \text{ item} = 7$$

Median = 7

Calculation of median - Continuous frequency distribution

Steps:

- Find out the median by using N/2
- Find out the class which median lies
- Apply the formula

$$\text{Median} = L + \frac{h}{f}\left(\frac{N}{2} - C\right)$$

Where L = lower limit of the median class

h = class interval of the median class

f = frequency of the median class

N = ∑f is the total frequency

C = Cumulative frequency of the preceding median class

Example: Calculate median from the following data

Age in years	Below 10	Below 20	Below 30	Below 40	Below 50	Below 60	Below 70	Above 70
No. of persons	2	5	9	12	14	15	15.5	15.6

Solution:

First we have to convert the distribution to a continuous frequency distribution as in the following table and then compute median.

Age in years	No. of persons (f)	Cumulative frequency (cf) – less than
0-10	2	2
10-20	5-2=3	5
20-30	9-5=**4**	9
30-40	12-9=3	12
40-50	14-12=2	14
50-60	15-14=1	15
60-70	15.5-15=0.5	15.5
70 and above	15.6-15.5=0.1	15.6

Median item = N/2 = 15.6/2 = 7.8

Find the cumulative frequency (*c.f*) greater than 7.8 is 9. Thus the corresponding class 20-30 is the median class.

Here L = 20, h = 10, f = 4, N = 15.6, C = 5

Using the formula

$$\text{Median} = L + \frac{h}{f}\left(\frac{N}{2} - c\right)$$

$$\text{Median} = 20 + \frac{10}{4}\left(\frac{15.6}{2} - 5\right)$$

Median = 20 + 2.5 (7.8-5)

Median = 20 + 2.5 ×2.8

Median = 27

So the median age is 27

The Mean vs. the Median

As measures of central tendency, the mean and the median each have advantages and disadvantages. Some pros and cons of each measure are summarized below. The median may be a better indicator of the most typical

value if a set of scores has an outlier. An outlier is an extreme value that differs greatly from other values.

However, when the sample size is large and does not include outliers, the mean score usually provides a better measure of central tendency.

1.4.3 Quartiles

All of us are aware of the median, which is the middle value or the mean of the two middle values, of an array. We have learned that the median divides a set of data into two equal parts. In the same way, there are also certain other values which divide a set of data into four, ten or hundred equal parts. Such values are referred as quartiles, deciles and percentiles respectively. Collectively, the quartiles, deciles and percentiles and other values obtained by equal sub-division of the data are called Quartiles.

Quartiles: The values which divide an array (a set of data arranged in ascending or descending order) into four equal parts are called quartiles. The first, second and third quartiles are denoted by Q_1, Q_2 and Q_3 respectively. The first and third quartiles are also called the lower and upper quartiles respectively. The second quartile represents the median, the middle value.

Quartiles for Ungrouped Data:

Quartiles for ungrouped data are calculated by the following formulae.

$$Q_1 = Value\ of\ \frac{(n+1)}{4}\ th\ item$$

$$Q_2 = Value\ of\ \frac{2(n+1)}{4}\ th\ item\ or\ \frac{(n+1)}{2}\ th\ item$$

$$Q_3 = Value\ of\ \frac{3(n+1)}{4}\ th\ item$$

Example:

Following is the data is of marks obtained by 20 students in a test of statistics;

53	74	82	42	39	20	81	68	58	28
67	54	93	70	30	55	36	37	29	61

In order to apply formulae we need to arrange the above data into ascending order i.e. in the form of an array.

20	28	29	30	36	37	39	42	53	54
55	58	61	67	68	70	74	81	82	93

Here, n = 20

$$Q_1 = \text{Value of } \frac{(n+1)}{4} \text{th item}$$

$$= \frac{(20+1)}{4} \text{th item}$$

$$= 5.25\text{th item from below}$$

The value of 5th item is 36 and that of the 6th item is 37. Thus the first quartile is a value 0.25th of the way between 36 and 37, which are 36.25. Therefore, Q_1 = 36.25. Similarly,

II. $$Q_2(\text{median}) = \text{Value of } \frac{2(n+1)}{4} \text{th item}$$

$$= \frac{2(20+1)}{4} \text{th item}$$

$$= 10.5\text{th item from below}$$

The value of the 10th item is 54 and that of the 11th item is 55. Thus the second quartile is the 0.5th of the value 54 and 55. Since the difference between 54 and 55 is of 1, therefore 54 + 1(0.5) = 54.5. Hence, = 54.5. Likewise,

$$Q_3 = \text{Value of } \frac{3(n+1)}{4} \text{th item}$$

$$= \frac{3(20+1)}{4} \text{th item}$$

$$= 15.75\text{th item from below}$$

The value of the 15th item is 68 and that of the 16th item is 70. Thus the third quartile is a value 0.75th of the way between 68 and 70. As the difference between

68 and 70 is 2, so the third quartile will be 68 + 2(0.75) = 69.5. Therefore, Q3 = 69.5.

Quartiles for Grouped Data:

The quartiles may be determined from grouped data in the same way as the median except that in place of n/2 we will use n/4. For calculating quartiles from grouped data we will form cumulative frequency column. Quartiles for grouped data will be calculated from the following formulae;

$$Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - C.F\right)$$

$$Q_3 = l + \frac{h}{f}\left(\frac{3n}{4} - C.F\right)$$

Q_2 = Median.

Where,

l = lower class boundary of the class containing the Q_1 or Q_3, i.e. the class corresponding to the cumulative frequency in which n/4 or 3n/4 lies

h = class interval size of the class containing Q_1 or Q_3.

f = frequency of the class containing Q_1 or Q_3.

n = number of values, or the total frequency.

C.F = cumulative frequency of the class preceding the class containing Q_1 or Q_3.

Example:

We will calculate the quartiles from the frequency distribution for the weight of 120 students as given in the following Table 1;

Table 1

Weight (lb)	Frequency (f)	Class Boundaries	Cumulative Frequency
110 – 119	**1**	**109.5 – 119.5**	**0**
120 – 129	**4**	**119.5 – 129.5**	**5**
130 – 139	**17**	**129.5 – 139.5**	**22**
140 – 149	**28**	**139.5 – 149.5**	**50**
150 – 159	**25**	**149.5 – 159.5**	**75**
160 – 169	**18**	**159.5 – 169.5**	**93**
170 – 179	**13**	**169.5 – 179.5**	**106**
180 – 189	**6**	**179.5 – 189.5**	**112**
190 – 199	**5**	**189.5 – 199.5**	**117**
200 – 209	**2**	**195.5 – 209.5**	**119**
210 – 219	**1**	**209.5 – 219.5**	**120**
	$\sum f$ **= n = 120**		

i. The first quartile Q_1 the value of $\frac{n}{4}th = \frac{120}{4}th$ or the 30th item from the lower end. From Table 18 we see that cumulative frequency of the third class is 22 and that of the fourth class is 50. Thus Q_1 lies in the fourth class i.e. 140 - 149.

$$Q_1 = l + \frac{h}{f}\left(\frac{n}{4} - C.F\right)$$

$$= 139.5 + \frac{10}{28}(30 - 22)$$

$$= 139.5 + 2.86$$

$$= 142.36 \ pounds.$$

ii. The thirds quartile Q_3 is the value of $\frac{3n}{4}th = \frac{3(120)}{4}th$ or 90^{th} item from the lower end. The cumulative frequency of the fifth class is 75 and that of the sixth class is 93. Thus, Q_3 lies in the sixth class i.e. 160 - 169.

$$Q_3 = l + \frac{h}{f}\left(\frac{3n}{4} - C.F\right)$$

$$= 159.5 + \frac{10}{18}(90 - 75)$$

$$= 159.5 + 8.33$$

$$= 167.83 \ pounds.$$

Conclusion

From Q_1 and Q_3 we conclude that 25% of the students weigh 142.36 pounds or less and 75% of the students weigh 167.83 pounds or less.

1.4.4 Deciles:

The values which divide an array into ten equal parts are called deciles. The first, second …and ninth deciles are represented by D_1, D_2,.....D_9 respectively. The fifth D_5 corresponds to median. The second, fourth, sixth and eighth deciles which collectively divide the data into five equal parts are called quintiles.

Deciles for Ungrouped Data:

Deciles for ungrouped data will be calculated from the following formulae;

$$D_1 = Value\ of\ \frac{(n+1)}{10}th\ item$$

$$D_2 = Value\ of\ \frac{2(n+1)}{10}th\ item$$

$$\vdots$$

$$\vdots$$

$$D_9 = Value\ of\ \frac{9(n+1)}{10}th\ item$$

20	28	29	30	36	37	39	42	53	54
55	58	61	67	68	70	74	81	82	93

i. $D_2 =$ Value of $\frac{2(n+1)}{10}$ th item

$= \frac{2(20+1)}{10}$ th item

$=$ 4.25th item from below

The value of the 4th item is 30 and that of the 5th item is 36. Thus the second decile is a value 0.2th of the way between 30 and 36. The fifth decile will be 30 + 6(0.2) = 31.2. Therefore, $D_2 = 31.2$.

ii. $D_2 =$ Value of $\frac{2(n+1)}{10}$ th item

$= \frac{3(20+1)}{10}$ th item

$=$ 6.3th item from below

The value of the 6th item is 37 and that of the 7th item is 39. Thus the third decile is 0.3th of the way between 37 and 39. The third decile will be 37 + 2(0.3) = 37.6. Hence, $D_3 = 37.6$.

iii. $D_7 =$ Value of $\frac{7(n+1)}{10}$ th item

$= \frac{7(20+1)}{10}$ th item

$=$ 14.7th item from below

The value of the 14th item is 67 and that of the 15th item is 68. Thus the 7th decile is 0.7th of the way between 67 and 68, which will be as 37 + 0.7 = 67.7. Therefore, $D_7 = 67.7$.

Decile for Grouped Data

Decile for grouped data can be calculated from the following formulae;

$$D_1 = l + \frac{h}{f}\left(\frac{n}{10} - C.F\right)$$

$$D_2 = l + \frac{h}{f}\left(\frac{2n}{10} - C.F\right)$$

$$\vdots$$

$$\vdots$$

$$D_9 = l + \frac{h}{f}\left(\frac{9n}{10} - C.F\right)$$

Where,

l = lower class boundary of the class containing the D_2 or D_9, i.e. the class corresponding to the cumulative frequency in which 2n/10 or 9n/10 lies

h = class interval size of the class containing D_2 or D_9.

f = frequency of the class containing D_2 or D_9.

n = number of values, or the total frequency.

C.F = cumulative frequency of the class preceding the class containing D_2 or D_9.

Example:

We will calculate fourth, seventh and ninth deciles from the frequency distribution of weights of 120 students, as provided in Table 1.

i. $D_4 = l + \frac{h}{f}\left(\frac{4n}{10} - C.F\right)$

$= 139.5 + \frac{10}{28}(48 - 22)$

$= 148.79\ pounds.$

ii. $D_7 = l + \frac{h}{f}\left(\frac{7n}{10} - C.F\right)$

$= 159.5 + \frac{10}{18}(84 - 75)$

$$= 164.5 \ pounds.$$

iii. $$D_9 = l + \frac{h}{f}\left(\frac{9n}{10} - C.F\right)$$

$$= 179.5 + \frac{10}{6}(108 - 106)$$

$$= 182.83 \ pounds.$$

Conclusion:

From D_1, D_4, and D_9 we conclude that 40% students weigh 148.79 pounds or less, 70% students weigh 164.5 pounds or less and 90% students weigh 182.83 pounds or less.

1.4.5 Percentiles and Mode:

The values which divide an array into one hundred equal parts are called percentiles. The first, second,.....,Ninety-ninth percentile are denoted by P_1, P_2,.....and P_{99}. The 50^{th} percentile P_{50} corresponds to the median. The 25^{th} percentile P_{25} corresponds to the first quartile and the 75^{th} percentile P_{75} corresponds to the third quartile.

Percentiles for Ungrouped Data:

Percentile from ungrouped data could be calculated from the following formulae;

$$P_1 = Value\ of\ \frac{(n+1)}{100}\ th\ item$$

$$P_2 = Value\ of\ \frac{2(n+1)}{100}\ th\ item$$

$$\vdots$$

$$\vdots$$

$$P_{99} = Value\ of\ \frac{99(n+1)}{100}\ th\ item$$

Example:

We will calculate fifteenth, thirty-seventh and sixty-fourth percentile from the following array;

20 28 29 30 36 37 39 42 53 54

55 58 61 67 68 70 74 81 82 93

P_{15} = Value of $\frac{15(n+1)}{100}$ th item

$= \frac{15(20+1)}{100}$ th item

= 3.15th item from below

The value of the 3rd item is 29 and that of the 4th item is 30. Thus the 15th percentile is 0.15th item the way between 29 and 30, which will be calculated as 29 + 0.15 = 29.15. Hence, = 29.15.

ii. P_{37} = Value of $\frac{37(n+1)}{100}$ th item

$= \frac{37(20+1)}{100}$ th item

= 7.77th item from below

The value of 7th item is 39 and that of the 8th item is 42. Thus the 37th percentile is 0.77th of the between 39 and 42, which will be calculate as 39 + 3(0.77) = 41.31. Hence, = 41.31.

Percentiles for Grouped Data:

Percentiles can also be calculated for grouped data which is done with the help of following formulae;

$$P_1 = l + \frac{h}{f}\left(\frac{n}{100} - C.F\right)$$

$$P_2 = l + \frac{h}{f}\left(\frac{2n}{100} - C.F\right)$$

⋮

$$P_{99} = l + \frac{h}{f}\left(\frac{99n}{100} - C.F\right)$$

Where,

l = lower class boundary of the class containing the, P_{35} or P_{99} i.e. the class corresponding to the cumulative frequency in which 35n/100 or 99n/100 lies

h = class interval size of the class containing. P_{35} or P_{99}

f = frequency of the class containing . P_{35} or P_{99}

n = number of values, or the total frequency.

C.F = cumulative frequency of the class preceding the class containing . P_{35} or P_{99}

Example:

We will calculate thirty-seventh, forty-fifth and ninetieth percentile from the frequency distribution of weights of 120 students, by using the Table 1.

i. $$P_{37} = l + \frac{h}{f}\left(\frac{37n}{100} - C.F\right)$$

$$= 139.5 + \frac{10}{28}(44.4 - 22)$$

$$= 147.5 \ pounds.$$

$$P_{45} = l + \frac{h}{f}\left(\frac{45n}{100}0 - C.F\right)$$

$$= 149.5 + \frac{10}{25}(54 - 50)$$

$$= 151.1 \ pounds.$$

Mode:

The mode of a data set is the value that occurs with the most frequency. This measurement is crude, yet is very easy to calculate. Suppose that a history class of eleven students scored the following (out of 100) on a test: 60, 64, 70, 70, 70, 75, 80, 90, 95, 95, and 100. We see that 70 is in the list three times, 95 occurs twice, and each of the other scores are each listed only once. Since 70 appears in the list more than any other score, it is the mode. If there are two values that tie for the most frequency, then the data is said to be bimodal.

The mode can be very useful for dealing with categorical data. For example, if a pizza shop sells 10 different types of sandwiches, the mode would

represent the most popular pizza. The mode also can be used with ordinal, interval, and ratio data. However, in interval and ratio scales, the data may be spread thinly with no data points having the same value. In such cases, the mode may not exist or may not be very meaningful.

To find mode in the case of a continuous frequency distribution, mode is found using the formula

$$\text{Mode} = L + \frac{h\,(f1-f2)}{(f1-f0)(f2-f1)}$$

Rearranging we get

$$\text{Mode} = L + \frac{h\,(f1-fo)}{2f1-fo-f2}$$

Where

L is the lower limit of the model class

f1 is the frequency of the model class

fo is the frequency of the class preceding the model class

f2 is the frequency of the class succeeding the model class

h is the class interval of the model class

When to use Mean, Median, and Mode

The following table summarizes the appropriate methods of determining the middle or typical value of a data set based on the measurement scale of the data.

Measurement Scale	Best Measure
Nominal (Categorical)	Mode
Ordinal	Median
Interval	Symmetrical data: Mean
	Skewed data: Median
Ratio	Symmetrical data: Mean
	Skewed data: Median

Merits and demerits of median and mode

Merits and demerits of arithmetic mean have already been discussed. Please refer to that. Here we discuss only median and mode.

Median:

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

Merits of median

(1) Simplicity: - It is very simple measure of the central tendency of the series. I the case of simple statistical series, just a glance at the data is enough to locate the median value.

(2) Free from the effect of extreme values: - Unlike arithmetic mean, median value is not destroyed by the extreme values of the series.

(3) Certainty: - Certainty is another merits is the median. Median values are always a certain specific value in the series.

(4) Real value: - Median value is real value and is a better representative value of the series compared to arithmetic mean average, the value of which may not exist in the series at all.

(5) Graphic presentation: - Besides algebraic approach, the median value can be estimated also through the graphic presentation of data.

(6) Possible even when data is incomplete: - Median can be estimated even in the case of certain incomplete series. It is enough if one knows the number of items and the middle item of the series.

Demerits of median:

Following are the various demerits of median:

(1) Lack of representative character: - Median fails to be a representative measure in case of such series the different values of which are wide apart from each other. Also, median is of limited representative character as it is not based on all the items in the series.

(2) Unrealistic: - When the median is located somewhere between the two middle

values, it remains only an approximate measure, not a precise value.

(3) Lack of algebraic treatment: - Arithmetic mean is capable of further algebraic treatment, but median is not. For example, multiplying the median with the number of items in the series will not give us the sum total of the values of the series.

However, median is quite a simple method finding an average of a series. It is quite a commonly used measure in the case of such series which are related to qualitative observation as and health of the student.

Mode: The value of the variable which occurs most frequently in a distribution is called the mode.

Merits of mode:

Following are the various merits of mode:

(1) Simple and popular: - Mode is very simple measure of central tendency. Sometimes, just at the series is enough to locate the model value. Because of its simplicity, it s a very popular measure of central tendency.

(2) Less effect of marginal values: - Compared top mean, mode is less affected by marginal values in the series. Mode is determined only by the value with highest frequencies.

(3) Graphic presentation:- Mode can be located graphically, with the help of histogram.

(4) Best representative: - Mode is that value which occurs most frequently in the series. Accordingly, mode is the best representative value of the series.

(5) No need of knowing all the items or frequencies: - The calculation of mode does not require knowledge of all the items and frequencies of a distribution. In simple series, it is enough if one knows the items with highest frequencies in the distribution.

Demerits of mode:

(1) Uncertain and vague: - Mode is an uncertain and vague measure of the central tendency.

(2) Not capable of algebraic treatment: - Unlike mean, mode is not capable of

further algebraic treatment.

(3) Difficult: - With frequencies of all items are identical, it is difficult to identify the modal value.

(4) Complex procedure of grouping: - Calculation of mode involves cumbersome procedure of grouping the data. If the extent of grouping changes there will be a change in the model value.

(5) Ignores extreme marginal frequencies: - It ignores extreme marginal frequencies. To that extent model value is not a representative value of all the items in a series.

Besides, one can question the representative character of the model value as its calculation does not involve all items of the series.

$$\text{Mean} = \frac{\text{sum of all values}}{\text{total number of values}}$$

$$\text{Median} = \text{middle value (when the data are arranged in order)}$$

$$\text{Mode} = \text{most common value}$$

1.5 Measures of Dispersion

The terms variability, spread, and dispersion are synonyms, and refer to how spread out a distribution is. Just as in the section on central tendency where we discussed measures of the centre of a distribution of scores, here we discuss measures of the variability of a distribution. Measures of variability provide information about the degree to which individual scores are clustered about or

deviate from the average value in a distribution.

Quite often students find it difficult to understand what is meant by variability or dispersion and hence they find the measures of dispersion difficult. So we will discuss the meaning of the term in detail. First one should understand that dispersion or variability is a continuation of our discussion of measure of central tendency. So for any discussion on measure of dispersion we should use any of the measure of central tendency. We continue this discussion taking mean as an example. The mean or average measures the centre of the data. It is one aspect observations. Another feature of the observations is as to how the observations are spread about the centre. The observation may be close to the centre or they may be spread away from the centre. If the observation are close to the centre (usually the arithmetic mean or median), we say that dispersion or scatter or variation is small. If the observations are spread away from the centre, we say dispersion is large.

Let us make this clear with the help of an example. Suppose we have three groups of students who have obtained the following marks in a test. The arithmetic means of the three groups are also given below:

Group A: 46, 48, 50, 52, 54, for this the mean is 50.

Group B: 30, 40, 50, 60, 70, for this the mean is 50.

Group C: 40, 50, 60, 70, 80, for this the mean is 60.

In a group A and B arithmetic means are equal i.e. mean of Group A = Mean of Group B = 50. But in group A the observations are concentrated on the centre. All students of group A have almost the same level of performance. We say that there is consistence in the observations in group A. In group B the mean is 50 but the observations are not closed to the centre. One observation is as small as 30 and one observation is as large as 70. Thus there is greater dispersion in group B. In group C the mean is 60 but the spread of the observations with respect to the centre 60 is the same as the spread of the observations in group B with respect to their own centre which is 50. Thus in group B and C the means

are different but their dispersion is the same. In group A and C the means are different and their dispersions are also different. Dispersion is an important feature of the observations and it is measured with the help of the measures of dispersion, scatter or variation. The word variability is also used for this idea of dispersion.

The study of dispersion is very important in statistical data. If in a certain factory there is consistence in the wages of workers, the workers will be satisfied. But if some workers have high wages and some have low wages, there will be unrest among the low paid workers and they might go on strikes and arrange demonstrations. If in a certain country some people are very poor and some are very high rich, we say there is economic disparity. It means that dispersion is large. The idea of dispersion is important in the study of wages of workers, prices of commodities, standard of living of different people, distribution of wealth, distribution of land among framers and various other fields of life. Some brief definitions of dispersion are:

The degree to which numerical data tend to spread about an average value is called the dispersion or variation of the data.

Dispersion or variation may be defined as a statistics signifying the extent of the scatteredness of items around a measure of central tendency.

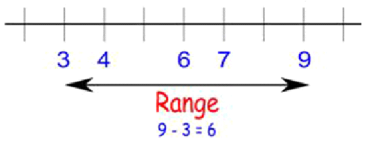
Dispersion or variation is the measurement of the scatter of the size of the items of a series about the average.

There are five frequently used measures of variability: the Range, Interquartile range or quartile deviation, Mean deviation or average deviation, Standard deviation and Lorenz curve.

1.5.1 Range

The range is the simplest measure of variability to calculate, and one you have probably encountered many times in your life. The range is simply the highest score minus the lowest score.

Range: R = maximum - minimum



Let's take a few examples.

What is the range of the following group of numbers: 10, 2, 5, 6, 7, 3, 4. Well, the highest number is 10, and the lowest number is 2, so 10 - 2 = 8. The range is 8.

Let's take another example.

Here's a dataset with 10 numbers: 99, 45, 23, 67, 45, 91, 82, 78, 62, and 51. What is the range? The highest number is 99 and the lowest number is 23, so 99 - 23 equals 76; the range is 76.

Example: Ms. Kesavan listed 9 integers on the blackboard. What is the range of these integers? 14, -12, 7, 0, -5, -8, 17, -11, 19

Ordering the data from least to greatest, we get: -12, -11, -8, -5, 0, 7, 14, 17, and 19

Range: R = highest - lowest = 19 - - 12 = 19 + 12 = +31

The range of these integers is +31

Merits and Limitations of range

Merits

- Amongst all the methods of studying dispersion, range is the simplest to understand easiest to compute.
- It takes minimum time to calculate the value of range. Hence if one is interested in getting a quick rather than very accurate picture of variability one may compute range.

Limitation

- Range is not based on each and every item of the distribution.
- It is subject to fluctuation of considerable magnitude from sample to sample.
- Range can't tell us anything about the character of the distribution with the two.
- According to kind "Range is too indefinite to be used as a practical measure of dispersion

Uses of Range

Range is useful in studying the variations in the prices of stocks, shares and other commodities that are sensitive to price changes from one period to another period.

The meteorological department uses the range for weather forecasts since public is interested to know the limits within which the temperature is likely to vary on a particular day.

1.5.2 Mean deviation

The mean deviation is also known as the average deviation. As the name implies, it is the average of absolute amounts by which the individual items deviate from the mean. Since the positive deviations from the mean are equal to the negative deviations, while computing the mean deviation, we ignore positive and negative signs.

Symbolically,

$MD = n \sum| x |$

Where MD = mean deviation, |x| = deviation of an item from the mean ignoring positive and negative signs, n = the total number of observations.

Example

Size of Item	Frequency
2-4	20
4-6	40
6-8	30
8-10	10

Solution:

Size of Item	Mid-points (m)	Frequency (f)	fm	d from $\bar{x}$	f |d|
2-4	3	20	60	-2.6	52
4-6	5	40	200	-0.6	24
6-8	7	30	210	1.4	42
8-10	9	10	90	3.4	34

Total	100	560	152

$\bar{x} = \sum fm/n = 560/100 = 5.6$

$MD(\bar{x}) = \sum f|d|/n = 152/100 = 1.52$

Merits of Mean Deviation

1. A major advantage of mean deviation is that it is simple to understand and easy to calculate.
2. It takes into consideration each and every item in the distribution. As a result, a change in the value of any item will have its effect on the magnitude of mean deviation.
3. The values of extreme items have less effect on the value of the mean deviation.
4. As deviations are taken from a central value, it is possible to have meaningful comparisons of the formation of different distributions.

Limitations of Mean Deviation

1. It is not capable of further algebraic treatment.
2. At times it may fail to give accurate results. The mean deviation gives best results when deviations are taken from the median instead of from the mean. But in a series, which has wide variations in the items, median is not a satisfactory measure.
3. Strictly on mathematical considerations, the method is wrong as it ignores the algebraic signs when the deviations are taken from the mean.

In view of these limitations, it is seldom used in business studies. A better measure known as the standard deviation is more frequently used.

1.5.3 Standard Deviation and Coefficient of Variation:

The concept, standard deviation was introduced by Karl Pearson in 1893. It is the most important measure of dispersion and is widely used. It is a measure of the dispersion of a set of data from its mean. The standard deviation is kind of the "mean of the mean," and often can help you find the story behind the data.

The standard deviation is a measure that summarises the amount by which

every value within a dataset varies from the mean. Effectively it indicates how tightly the values in the dataset are bunched around the mean value. It is the most robust and widely used measure of dispersion since, unlike the range and inter-quartile range; it takes into account every variable in the dataset. When the values in a dataset are pretty tightly bunched together the standard deviation is small. When the values are spread apart the standard deviation will be relatively large.

Standard deviation is defined as a statistical measure of dispersion in the value of an asset around mean. The standard deviation calculation tells you how spread out the numbers are in your sample. Standard Deviation is represented using the symbol ? (Greek letter sigma)

For example if you want to measure the performance a mutual fund, SD can be used. It gives an idea of how volatile a fund's performance is likely to be. It is an important measure of a fund's performance. It gives an idea of how much the return on the asset at a given time differs or deviates from the average return. Generally, it gives an idea of a fund's volatility i.e. a higher dispersion (indicated by a higher standard deviation) shows that the value of the asset has fluctuated over a wide range.

The formula for finding SD in a sentence form is: it is the square root of the Variance. So now you ask, 'What is the Variance'? Let us see what variance is.

The Variance is defined as: The average of the squared differences from the Mean.

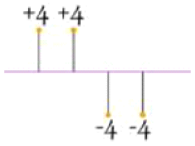
We can calculate the variance follow these steps:

Work out the Mean (the simple average of the numbers)

Then for each number: subtract the Mean and square the result (the squared difference).

Then work out the average of those squared differences.

You may ask Why square the differences. If we just added up the differences from the mean ... the negatives would cancel the positives as shown below. So we take the square



Example

You have figures of the marks obtained by your five bench mates which are as follows:

600, 470, 170, 430 and 300.

Find out the Mean, the Variance, and the Standard Deviation.

Your first step is to find the Mean:

$$\text{Mean} = \frac{600 + 470 + 170 + 430 + 300}{5}$$

$$\text{Mean} = \frac{1970}{5}$$

Mean = 394

So the mean (average) mark is 394. Let us plot this on the chart:

x	x- mean	X^2
600	206	42436
470	76	5776
170	-224	50176
430	36	1296
300	-94	8836
Total		108520

To calculate the Variance, take each difference, square it, find the sum (108520) and find average: $= \frac{108520}{5}$

= 21704

So, the Variance is 21704.

The Standard Deviation is just the square root of Variance, so:

SD = σ = $\sqrt{21704}$ = 147.32 = 147

Now we can see which heights are within one Standard Deviation (147) of the Mean.

Please note that there is a slight difference when we find variance from a population and mean. In the above example we found out variance for data collected from all your bench mates. So it may be considered as population.

Suppose now you collect data only from some of your bench mates. Now it may be considered as a sample. If you are finding variance for a sample data, in the formula to find variance, divide by N-1 instead of N.

For example, if we say that in our problem the marks are of some students in a class, it should be treated as a sample. In that case Variance (or to be precise Sample Variance) = 108,520 / 4 = 27,130. Note that instead of N (i.e.5) we divided by N-1 (5-1=4).

Standard Deviation (Sample Standard Deviation) = ? = ?27130 = 164.31 = 164

Based on the above information, let us build the formula for finding SD. Since we use two different formulae for data which is population and data which is sample, we will have two different formulas for SD also.

The "Population Standard Deviation":

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

The "Sample Standard Deviation":

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Merits of Standard Deviation

- It is rigidly defined and its value is always definite and based on all

observation.

- As it is based on arithmetic mean, it has all the merits of arithmetic mean.
- It is possible for further algebraic treatment.
- It is less affected by sampling fluctuations.

Demerits

- It is not easy to calculate.
- It gives more weight to extreme values, because the values are squared up

Coefficient of Variation

Standard deviation is the absolute measure of dispersion. It is expressed in terms of the units in which the original figures are collected and stated. The relative measure of standard deviation is known as coefficient of variation. Variance: Square of Standard deviation

Symbolically;

Variance = σ^2

$\sigma = \sqrt{\text{varience}}$

Coefficient of standard deviation = $\frac{\sigma}{x}$

Coefficient of variation (CV) calculator - to find the ratio of standard deviation (σ) to mean (μ). The main purpose of finding coefficient of variance (often abbreviated as CV) is used to study of quality assurance by measuring the dispersion of the population data of a probability or frequency distribution, or by determining the content or quality of the sample data of substances. The method of measuring the ratio of standard deviation to mean is also known as relative standard deviation often abbreviated as RSD. It only uses positive numbers in the calculation and expressed in percentage values. Therefore, the resultant value of this formula CV = (**Standard Deviation (σ) / Mean** (μ) will be multiplied by 100. CV is important in the field of probability & statistics to **measure the relative variability** of the data sets on a ratio scale. In probability theory and

statistics, it is also known as unitized risk or the variance coefficient.

The below three formulas are used to find the standard deviation, mean and coefficient of variation to measure the relative variability of data sets having different mean and the unit scale.

Formulas to calculate coefficient of variation:

The formula to find the sample mean

$$\mu = \frac{\sum x}{n}$$

Formula to calculate sample standard deviation

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n - 1}}$$

Formula to calculate **coefficient of variation**

$$CV = \frac{\sigma}{\mu}$$

How to calculate coefficient of variation?

Follow these below step by step calculation using above formulas to find CV of the sample data

1. Calculate the mean of the data set.

2. Calculate the sample SD for the data set.

3. Finding the ratio of sample standard deviation to mean brings the CV of the data set.

Example

Calculate the relative variability (coefficient of variance) for the samples 60.25, 62.38, 65.32, 61.41, and 63.23 of a population

Solution:

Step by step calculation:

Step 1: calculate mean

Mean = (60.25 + 62.38 + 65.32 + 61.41 + 63.23)/5

= 312.59/5

= 62.51

Step 2: calculate standard deviation

$= \sqrt{((1/(5 - 1)) * (60.25 - 62.51799)^2 + (62.38 - 62.51799)^2 + (65.32 - 62.51799)^2 + (61.41 - 62.51799)^2 + (63.23 - 62.51799)^2)}$

$= \sqrt{((1/4) * (-2.26799^2 + -0.13798999^2 + 2.80201^2 + -1.10799^2 + 0.71201^2))}$

$= \sqrt{((1/4) * (5.14377 + 0.01904 + 7.85126 + 1.22764 + 0.50695))}$

$= \sqrt{3.68716}$

$\sigma = 1.92$

Step 3: calculate coefficient of variance

CV = (Standard Deviation (σ) / Mean (μ))

= 1.92 / 62.51

= 0.03071

The relative variability calculation is popularly used in engineering, physics, chemical industries etc. to employ the quality assurance. Therefore the coefficient of variance or relative standard deviation is widely used in various applications across the different types of industry. Any manual calculation can be done by using the above mathematical formulas. However, when it comes to online to measure the relative variability, this coefficient of variation calculator makes your calculation as simple as possible for the given sample data of the population.

1.6 Summary

A measure of central tendency (also referred to as measures of centre or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution. There are three main measures of central tendency: the mode, the median and the mean. Each of these measures describes a different indication of the typical or central

value in the distribution.

We have discussed three sets of summary measures namely Measures of Central Tendency, Variability and Shape. These are called summary measures because they summarise the data. For example, one of summary measure very familiar to you is mean. (Mean comes under measure of central tendency.) If we take mean mark of students in a class for a subject, it gives you a rough idea of what the marks is like. Thus based on just one summary value, we get idea of the entire data.

1.7 Glossary

Measures of central tendency: A measurement of data that indicates where the middle of the information lies.

Mean: The mean, or average, of n numbers is the sum of the numbers divided by n.

Median: The median of n numbers is the middle number when the numbers are written in order. If n is even, the median is the average of the two middle numbers.

Mode: The mode of n numbers is the number that occurs most frequently. If two numbers tie for most frequent occurrence, the collection has two modes and is called bimodal.

Standard deviation: Measure of the unpredictability of a random variable, expressed as the average deviation of a set of data from its arithmetic mean and computed as the positive square root of the variance.

Measures of dispersion: The distribution of values in relation to specific categories. Measurements of dispersion include standard deviations, ranges, and averages.

1.8 Self Assessment Questions

1. Explain Standard deviation as measure of central tendency?

2: Is variance a measure of central tendency?

3: Describe various measures of central tendency?

4: What does u mean by coefficient of variance?

5: What do you mean by dispersion? What are the different measures of dispersion?

6: Why is the standard deviation the most widely used measure of dispersion? Explain

1.9 Lesson End Exercises

1. Is a quartile a measure of central tendency? Discuss.
2. Explain various measures of dispersion.
4. Discuss coefficient of variance
5. Discuss the role of Statistics and its applications in managerial decision making
6. Discuss the following terms

- Mean
- Median
- Quartiles
- Deciles
- Percentiles and Mode
- Range
- Mean deviation

1.10 Suggested Readings

Beri, G.C. Business Statistics, IIIrd Ed. Tata McGraw Hill Pvt. Ltd.; India.

Cooper, Donald R. & Schindler, Pamela S. Business Research Methods, Tata McGraw Hill Compnies; India.

Jhunjhunwala, B. Business Statistics, S Chand & Co. New Delhi.

Sachdeva, J.K. Business Research Methodology, Himalaya Publishing House; New Delhi.

Shajahan, S. Research Methods for Management, Jaico Publishing House, Delhi; India.

Singh, D. & Chaudhary F.S. Theory and Analysis of Sample Survey Designs, New Age International (P) Limited: New Delhi.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. Multivariate data

analysis: A global perspective (7th ed.). Upper Saddle River: Pearson Education.

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. A primer on partial least squares structural equation modeling (PLS-SEM). Thousand Oaks: Sage.

1.11 References

Aaker, D. A., Kumar, V. & Day, G. S Marketing Research, 7th edn, John Wiley, New York.

Baker, M. J. Research for Marketing, Macmillan, London.

Boyd, H., Westfall, R. & Stasch, S. Marketing research: Text and cases. Boston: Irwin.

Bryman, A. Social Research Methods. London: Oxford University Press.

Churchill, G. Marketing research (3rd ed.). Hinsdale, Illinois: Dryden Press.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. Multivariate data analysis (6th ed.). New Jersey: Pearson Education International.

Hair, J. F., Celsi, M., Money, A., Samouel, P., & Page, M. Essentials of business research methods (2nd ed.). Armonk, NY: ME Sharpe.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. Journal of Management, 21(5), 967-988.

Kent, Ray Marketing Reearch: Approaches, methods and applications in Europe. London: Thomson.

Kline, R. B. Principles and practice of structural equation modeling (3rd ed.). New York: The Guilford Press.

Kothari, C.R. Research Methodology Methods and Techniques, New Age International (P) Limited: New Delhi.

Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. Educational and Psychological Measurement, 30, 607-610

Kumar, V., Aaker, D. A., & Day, G. S. Essentials of marketing research New York: John Wiley & Sons, Inc.

Malhotra, N. K. Marketing Research: An Applied Orientation, 3rd edn, Prentice-Hall International, London

Malhotra, N.K. Marketing Research- An Applied Orientation, Pearson Education, Singapore.

Malhotra, N. K. Marketing Research (4th ed.). Harlow: Prentice Hall.

Malhotra, N. Marketing Research (4th ed.). New Jersey: Pearson.

Schmidt, M. J., & Hollensen, S. Marketing research an international approach. Harlow: Pearson Education.

Willis, K. In-depth interviews, in the handbook of international market research techniques, Robin Birn, Ed. London: McGraw-Hill.

LESSON No. 2

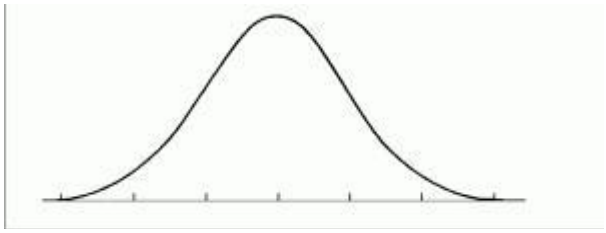
SKEWNESS **UNIT-II**

SKEWNESS AND TIME SERIES

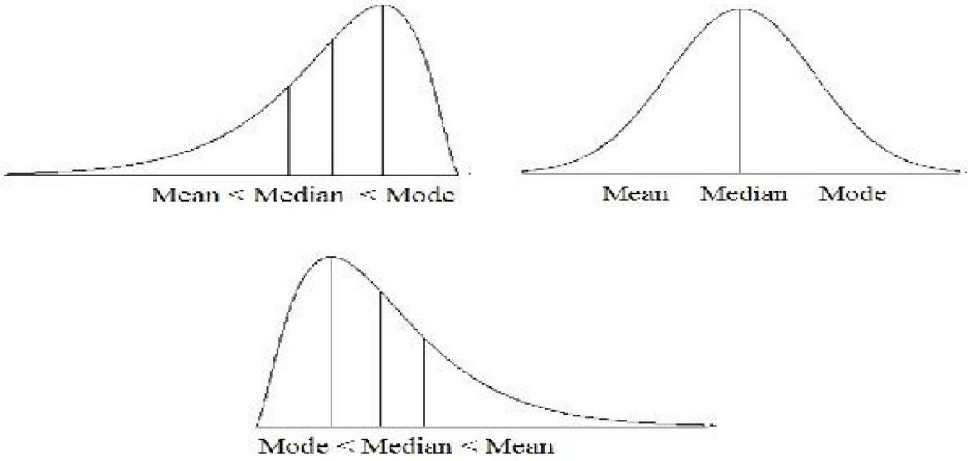
UNIT STRUCTURE

2.1 Introduction

We have discussed earlier techniques to calculate the deviations of a distribution from its measure of central tendency (mean / median, mode). Here we see another measure for that named Skewness. Skewness characterizes the degree of asymmetry of a distribution around its mean. If there is only one mode (peak) in our data (unimodel), and if the other data are distributed evenly to the left and right of this value, if we plot it in a graph, we get a curve like this, which is called a normal curve (See figure below). Here we say that there is no skewness or skewness = 0. If there is zero skewness (i.e., the distribution is symmetric) then the mean = median for this distribution.



However data need not always be like this. Sometimes the bulk of the data is at the left and the right tail is longer, we say that the distribution is skewed right or positively skewed. Positive skewness indicates a distribution with an asymmetric tail extending towards more positive values. On the other hand, sometimes the bulk of the data is at is at the right and the left tail is longer, we say that the distribution is skewed left or negatively skewed. Negative skewness indicates a distribution with an asymmetric tail extending towards more negative values"



In probability theory and statistics, **skewness** is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative, or even undefined.

The qualitative interpretation of the skew is complicated. For a unimodel distribution, negative skew indicates that the tail on the left side of the probability density function is longer or fatter than the right side – it does not distinguish these shapes. Conversely, positive skew indicates that the tail on the right side is longer or fatter than the left side. In cases where one tail is long but the other tail is fat, skewness does not obey a simple rule. For example, a zero value indicates that the tails on both sides of the mean balance out, which is the case for a symmetric distribution, but is also true for an asymmetric distribution where the asymmetries even out, such as one tail being long but thin, and the other being short but fat. Further, in multimodal distributions and discrete distributions, skewness is also difficult to interpret. Importantly, the skewness does not determine the relationship of mean and median.

2.2 Objectives

After studying this chapter students should be able to understand:

- What Skewness is and its meaning and significance.
- What is Karl Pearson's coefficient of skewness?

- What are the different methods of measures of skewness?
- Relationship between moments of mean and moments about origin.

2.3 Meaning of skewness

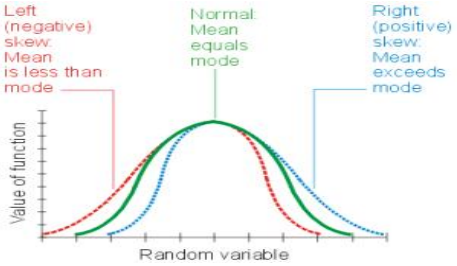
Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the centre point.

A fundamental task in many statistical analyses is to characterise the location and variability of a data set. A further characterisation of the data includes skewness and kurtosis.

Skewness is asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right. Skewness can be quantified to define the extent to which a distribution differs from a normal distribution.

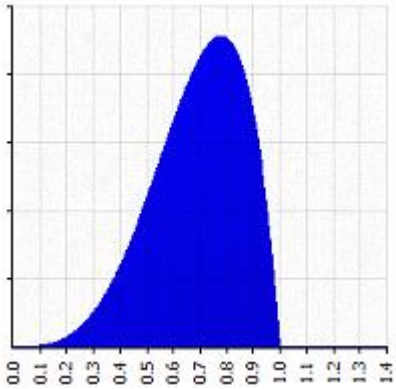
In a normal distribution, the graph appears as a classical, symmetrical "bell-shaped curve." The mean, or average, and the mode, or maximum point on the curve, are equal.

- In a perfect normal distribution (green solid curve in the illustration below), the tails on either side of the curve are exact mirror images of each other.
- When a distribution is skewed to the left (red dashed curve), the tail on the curve's left-hand side is longer than the tail on the right-hand side, and the mean is less than the mode. This situation is also called negative skewness.
- When a distribution is skewed to the right (blue dotted curve), the tail on the curve's right-hand side is longer than the tail on the left-hand side, and the mean is greater than the mode. This situation is also called positive skewness.



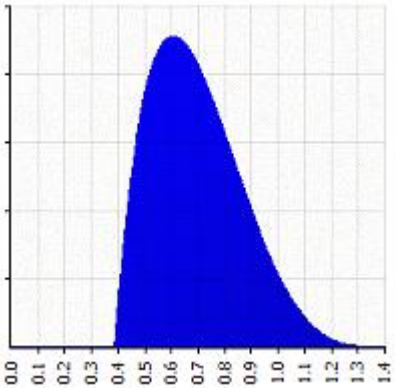
The first thing you usually notice about a distribution's shape is whether it has one mode (peak) or more than one. If it's **unimodal** (has just one peak), like most data sets, the next thing you notice is whether it's **symmetric or skewed** to one side. If the bulk of the data is at the left and the right tail is longer, we say that the distribution is **skewed right or positively skewed**; if the peak is toward the right and the left tail is longer, we say that the distribution is **skewed left or negatively skewed**.

Look at the two graphs below. They both have ? = 0.6923 and ? = 0.1685, but their shapes are different.



Beta (α=4.5, β=2)

Skewness = ?0.5370



1.3846 - Beta (α=4.5, β=2)

Skewness = +0.5370

The first one is moderately skewed left: the left tail is longer and most of the distribution is at the right. By contrast, the second distribution is moderately skewed right: its right tail is longer and most of the distribution is at the left.

The **moment coefficient of skewness** of a data set is

Skewness: $g_1 = m_3 / m_2^{3/2}$ (1)

Where

$m_3 = \Sigma(x?x?)^3 / n$ and $m_2 = \Sigma(x?x?)^2 / n$

x? is the mean and n is the sample size, as usual. m_3 is called the **third moment** of the

data set. m_2 is the **variance**, the square of the standard deviation.

The skewness can also be computed as g_1 = **the average value of z^3**, where z is the familiar z-score, z = (x?x?)/σ. Of course the average value of z is always zero, but what about the average of z^3? Suppose you have a few points far to the left of the mean, and a lot of points less far to the right of the mean. Since cubing the deviations gives the big ones even greater weight, you'll have negative skewness. It works just the opposite if you have big deviations to the right of the mean.

You'll remember that you have to compute the variance and standard deviation slightly differently, depending on whether you have data for the whole population or just a sample. The same is true of skewness. If you have the whole population, then g_1 above is the measure of skewness. But **if you have just a sample**, you need the **sample skewness**:

Sample skewness: $$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1 \qquad (2)$$

(The formula comes from Joanes and Gill 1998.)

Excel doesn't concern itself with whether you have a sample or a population: its measure of skewness is always G_1, the sample skewness.

Computing

The **moment coefficient of skewness** of a data set is

skewness: $g_1 = m_3 / m_2^{3/2}$

(1)where

$m_3 = \Sigma(x?x?)^3 / n$ and $m_2 = \Sigma(x?x?)^2 / n$

x? is the mean and n is the sample size, as usual. m_3 is called the **third moment** of the data set. m_2 is the **variance**, the square of the standard deviation.

The skewness can also be computed as g_1 = **the average value of z^3**, where z is the familiar z-score, z = (x?x?)/σ. Of course the average value of z is always zero, but what about the average of z^3? Suppose you have a few points far to the left of the

mean, and a lot of points less far to the right of the mean. Since cubing the deviations gives the big ones even greater weight, you'll have negative skewness. It works just the opposite if you have big deviations to the right of the mean.

You'll remember that you have to compute the variance and standard deviation slightly differently, depending on whether you have data for the whole population or just a sample. The same is true of skewness. If you have the whole population, then g_1 above is the measure of skewness. But **if you have just a sample**, you need the **sample skewness**:

Sample skewness: $$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1 \qquad (2)$$

(The formula comes from Joanes and Gill 1998.)

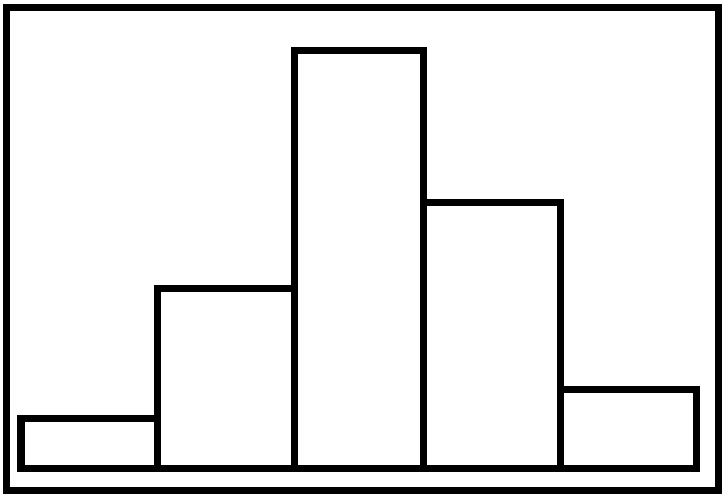
Excel doesn't concern itself with whether you have a sample or a population: its measure of skewness is always G_1, the sample skewness.

Example 1: College Men's Heights

Height (inches)	Class Mark, x	Frequency, f
59.5–62.5	61	5
62.5–65.5	64	18
65.5–68.5	67	42
68.5–71.5	70	27
71.5–74.5	73	8

Here are grouped data for heights of 100 randomly selected male students, adapted from Spiegel and Stephens (1999, 68).

A histogram shows that the data are skewed left, not symmetric.



But **how highly skewed** are they, compared to other data sets? To answer this question, you have to compute the skewness.

Begin with the sample size and sample mean. (The sample size was given, but it never hurts to check.)

n = 5+18+42+27+8 = 100

x? = (61×5 + 64×18 + 67×42 + 70×27 + 73×8) ÷ 100

x? = 9305 + 1152 + 2814 + 1890 + 584) ÷ 100

x? = 6745÷100 = 67.45

Now, with the mean in hand, you can compute the skewness. (Of course in real life you'd probably use Excel or a statistics package, but it's good to know where the numbers come from.)

Class Mark, x	Frequency, f	Xf	$(x-\bar{x})$	$(x-\bar{x})^2 f$	$(x-\bar{x})^3 f$
61	5	305	-6.45	208.01	-1341.68
64	18	1152	-3.45	214.25	-739.15
67	42	2814	-0.45	8.51	-3.83
70	27	1890	2.55	175.57	447.70
73	8	584	5.55	246.42	1367.63
	$\sum$	6745	n/a	852.75	−269.33
	$\bar{x}$, m_2, m_3	67.45	n/a	8.5275	−2.6933

Finally, the skewness is

$g_1 = m_3 / m_2^{3/2}$ = ?2.6933 / $8.5275^{3/2}$ = ?0.1082

But wait, there's more! That would be the skewness if you had data for the whole population. But obviously there are more than 100 male students in the world, or even in almost any school, so what you have here is a sample, not the population. You must compute the **sample skewness**:

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1$$ = [√ (100×99) / 98] [?2.6933 / $8.5275^{3/2}$] = ?0.1098

Interpreting

If skewness is positive, the data are positively skewed or skewed right, meaning that the right tail of the distribution is longer than the left. If skewness is negative, the data are negatively skewed or skewed left, meaning that the left tail is longer.

If skewness = 0, the data are perfectly symmetrical. But a skewness of exactly zero is quite unlikely for real-world data, so **how can you interpret the skewness number**? Bulmer (1979) — a classic — suggests this rule of thumb:

- If skewness is less than ?1 or greater than +1, the distribution is **highly skewed**.
- If skewness is between ?1 and ?½ or between +½ and +1, the distribution is **moderately skewed**.
- If skewness is between -½ and +½, the distribution is **approximately symmetric**.

With a skewness of - 0.1098, the sample data for student heights are approximately symmetric.

Caution: This is an interpretation of the data you actually have. When you have data for the whole population, that's fine. But when you have a sample, the sample skewness doesn't necessarily apply to the whole population. In that case the question is, from the sample skewness, can you conclude anything about the population skewness? To answer that question, see the next section.

Inferring

Your data set is just one sample drawn from a population. Maybe, from ordinary sample variability, your sample is skewed even though the population is symmetric. But if the sample is skewed too much for random chance to be the explanation, then you can conclude that there is skewness in the population.

But what do I mean by "too much for random chance to be the explanation"? To answer that, you need to divide the sample skewness G_1 by the **standard error of skewness (SES)** to get the **test statistic**, which measures how many standard errors separate the sample skewness from zero:

(3) Test statistic: $Z_{g1} = G_1/SES$ where $SES = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}$

This formula is adapted from page 85 of Cramer (1997). (Some authors suggest ?(6/

n), but for small samples that's a poor approximation. And anyway, we've all got calculators, so you may as well do it right.)

The critical value of Z_{g1} is approximately 2. (This is a two-tailed test of skewness ? 0 at roughly the 0.05 significance level.)

- **If Z_{g1} < -2**, the population is very likely skewed negatively (though you don't know by how much).
- **If Z_{g1} is between -2 and +2**, you can't reach any conclusion about the skewness of the population: it might be symmetric, or it might be skewed in either direction.
- **If Z_{g1} > 2**, the population is very likely skewed positively (though you don't know by how much).

Don't mix up the meanings of this test statistic and the amount of skewness. The amount of skewness tells you how highly skewed your sample is: the bigger the number, the bigger the skew. The test statistic tells you whether the whole population is probably skewed, but not by how much: the bigger the number, the higher the probability.

Estimating

GraphPad suggests a **confidence interval for skewness**:

95% confidence interval of population skewness = $G_1 \pm 2$ SES (4)

I'm not so sure about that. Joanes and Gill point out that sample skewness is an unbiased estimator of population skewness for normal distributions, but not others. So I would say, compute that confidence interval, but take it with several grains of salt — and the further the sample skewness is from zero, the more skeptical you should be.

For the college men's heights, recall that the sample skewness was G_1 = ?0.1098. The sample size was n = 100 and therefore the standard error of skewness is

SES = $\sqrt{[(600\times99) / (98\times101\times103)]}$ = 0.2414

The test statistic is

$Z_{g1} = G_1 / SES = ?0.1098 / 0.2414 = ?0.45$

This is quite small, so from this sample **it's impossible to say whether the population is symmetric or skewed.** Since the sample skewness is small, a confidence interval is probably reasonable:

$G_1 \pm 2\ SES = ?.1098 \pm 2 \times .2414 = ?.1098 \pm .4828 = ?0.5926$ to $+0.3730$.

You can give a 95% confidence interval of skewness as about ?0.59 to +0.37, more or less.

Tests of Skewness

There are certain tests to know whether skewness does or does not exist in a frequency distribution.

They are:

1. In a skewed distribution, values of mean, median and mode would not coincide. The values of mean and mode are pulled away and the value of median will be at the centre. In this distribution, mean-Mode = 2/3 (Median - Mode).
2. Quartiles will not be equidistant from median.
3. When the asymmetrical distribution is drawn on the graph paper, it will not give a bell shaped curve.
4. Sum of the positive deviations from the median is not equal to sum of negative deviations.
5. Frequencies are not equal at points of equal deviations from the mode.

Nature of Skewness

Skewness can be positive or negative or zero.

1. When the values of mean, median and mode are equal, there is no skewness.
2. When mean > median > mode, skewness will be positive.
3. When mean < median < mode, skewness will be negative.

Characteristic of a good measure of skewness

- It should be a pure number in the sense that its value should be independent of the unit of the series and also degree of variation in the series.
- It should have zero-value, when the distribution is symmetrical.
- It should have a meaningful scale of measurement so that we could easily interpret the measured value.

2.3.1 Significance of Skewness

Skewness has benefits in many areas. Many models assume normal distribution; i.e., data are symmetric about the mean. The normal distribution has a skewness of zero. But in reality, data points may not be perfectly symmetric. So, an understanding of the skewness of the dataset indicates whether deviations from the mean are going to be positive or negative.

D'Agostino's K-squared test is a goodness-of-fit normality test based on sample skewness and sample kurtosis.

Measures of Skewness

Skewness can be studied graphically and mathematically. When we study Skewness graphically, we can find out whether Skewness is positive or negative or zero. This is what we have shown above.

Mathematically Skewness can be studied as:

(a) Absolute Skewness

(b) Relative or coefficient of skewness

When the skewness is presented in absolute term i.e., in units, it is absolute skewness. If the value of skewness is obtained in ratios or percentages, it is called relative or coefficient of skewness. When skewness is measured in absolute terms, we can compare one distribution with the other if the units of measurement are same. When it is presented in ratios or percentages, comparison become easy. Relative measures of skewness is also called coefficient of skewness.

(a) Absolute measure of Skewness:

Skewness can be measured in absolute terms by taking the difference between mean and mode.

Absolute Skewness = x? - mode

If the value of the mean is greater than mode, the Skewness is positive

If the value of mode is greater than mean, the Skewness is negative.

Greater the amount of Skewness (negative or positive) the more tendency towards asymmetry. The absolute measure of Skewness will be proper measure for comparison, and hence, in each series a relative measure or coefficient of Skeweness has to be computed.

(b) Relative measure of skewness

There are three important measures of relative skewness.

1. Karl Pearson's coefficient of skewness.
2. Bowley's coefficient of skewness.
3. Kelly's coefficient of skewness.

2.3.2 Karl Pearson's coefficient of skewness.

The mean, median and mode are not equal in a skewed distribution. The Karl Pearson's measure of skewness is based upon the divergence of mean from mode in a skewed distribution.

Karl Pearson's measure of skewness is sometimes referred to Skp

$$Skp = \frac{Mean - Mode}{Standard\ Deviation}$$

This method of measuring skewness, also known as Pearsonian coefficient of skewness, was suggested by Karl Pearson, a great British biometrician and statistician. $\mathbf{Sk}_{\mathbf{p}}$ = Karl Pearson's coefficient of skewness.

There is no limit to this measure in theory and this is a slight drawback. But in practice the value given by this formula is rarely very high and usually lies between **± 1**. When a distribution is symmetrical, the values of mean, median and mode coincide and, therefore the coefficient of skewness will be zero. When a distribution is positively skewed, the coefficient of skewness shall have plus sign and when it is negatively skid, the coefficient of skewness shall have minus sign. The degree of skewness shall be obtained by the numeral value. Say, 0.8 or 0.2 etc. thus this formula given both the direction as well as the extent of skinniness.

The above method of measuring skewness cannot be used where mode is ill defined; however, in moderately skewed distribution the averages have the following relationship:

Mode = 3 median – 2 mean and therefore, if this value of mode is substituted in the above formula we arrive at another formula for finding out skewness.

Sk_p = [X – (3 med. - 2X)]/ σ = X – 3 med. / σ = 2 X = 3 (X- med.)/ ?

Theoretically the value of this coefficient varies between **± 3**; however, in practice it is rare that the coefficient of skewness obtained by the above method exceeds **± 1**.

Illustration: calculate Karl Pearson's coefficient of skewness from the following data;

Solution: calculation of Coefficient of skewness by Karl Pearson's method

Profits ($ 0.1 million)	No. of Cos.	Profits ($ 0.1 million)	No. of Cos.
70-80	12	110-120	50
80-90	18	120-130	45
90-100	35	130-140	30
100-110	42	140-150	8

Profits ($ 0.1 million)	m.p.m	F	(m – 115)/10d	fd	fd^2
70-80	75	12	-4	-48	192
80-90	85	18	-3	-54	162
90-100	95	35	-2	-70	140
100-110	105	42	-1	-42	42
110-120	115	50	0	0	0
120-130	125	45	1	45	45
130-140	135	30	2	60	120
	145	8	3	24	72
		N = 240		**Σ = fd = -85**	**Σ = fd^2 = 773**

Coefficient of skewness = Mean – Mode/σ

Mean: ?X = A = ?fd/N × I = 115 – 85/240 × 10 =115 – 3.54 = 111.46

Mode: by inspection mode lies in the class 110 – 120.

Mode = L + Δ1 / Δ_1 + Δ_2 × i

L = 110, Δ_1 = |50 - 42| = 8, Δ_2 = |50 - 45| = 5, I = 10

Therefore, Mode = 110 + 8/8 + 5 × 10 = 110 + 6.15 + 116.15

Standard deviation $\sigma = \sqrt{\Sigma fd^2/N - [\Sigma fd/N]^2} \times I = \sqrt{773/240 - [85/240]^2} \times 10$

Coefficient Of Skp = 111.46 – 116.15/17.795 = - 4.69/17.595 = - 0.266

Properties of Karl Pearson coefficient of Skewness

1. ?1 ? Skp ? 1.
2. Skp = 0 ? distribution is symmetrical about mean.
3. Skp> 0 ? distribution is skewed to the right.

4. Skp< 0 ? distribution is skewed to the left.

Advantage of Karl Pearson coefficient of Skewness

(1) Skp is independent of the scale. Because (mean-mode) and standard deviation have same scale and it will be cancelled out when taking the ratio.

Disadvantage of Karl Pearson coefficient of Skewness

1. Skp depends on the extreme values.

2.4 Bowley's method

Bowley skewness is a way to figure out if you have a positively-skewed or negatively skewed distribution. One of the most popular ways to find skewness is the Pearson Mode Skewness formula. However, in order to use it you must know the mean, mode (or median) and standard deviation for your data. Sometimes you might not have that information; instead you might have information about your quartiles. If that's the case, you can use Bowley Skewness as an alternative to find out more about the asymmetry of your distribution. It's very useful if you have extreme data values (outliers) or if you have an open-ended distribution.

An alternative measure of skewness has been proposed by the late professor Bowley. Bowley's quartiles are based on quartiles. In a symmetrical distribution first and third quartiles are equidistant from the median as can be seen from the following diagram.

In an asymmetrical distribution the third quartile is the same distance over the median as the first quartile is below it i.e.

$$\mathbf{Q_3 - Med. = Med. - Q_1 \text{ or } Q_3 = Q_1 - 2\ Med. = 0}$$

If this distribution is positively skewed the top 25 per cent of the values will tend to be farther from median than the bottom 25 per cent. i.e. $\mathbf{Q_3}$ will be farther from median than $\mathbf{Q_1}$ is form median and the reverse for negative skewness. Hence a possible measure is

$$\mathbf{Sk_b = (Q_3 - Med.) - (Med. - Q_1)/ (Q_3 - Med.) + (Med. - Q_1) \quad or \quad Q_3 + Q_1 - 2/ Q_3 - Q_1 Med.}$$

Sk_b = Bowley's coefficient of skewness.

It must be remembered that the results obtained by these two measures are not to be compared with one another especially. The numerical values are not related to one another since the burley's measure, because of its computational basis, is limited to values between **-1 and +1**, while person's measure has no such limits.

Not only do the numerical values obtained from these two formulae bear no necessary relationship to one another but, on rare occasions, with unusually shaped distributions, it is possible for them to emerge with opposite sings.

Bowley Skewness = $Q_3 + Q_1 - 2Q_2 / (Q_3 - Q1)$

- Skewness = 0 means that the curve is symmetrical.
- Skewness > 0 means the curve is positively skewed.
- Skewness < 0 means the curve is negatively skewed.

Illustration: Find Bowley's Coefficient of Skewness for the following frequency distribution:

No. of children per family	0	1	2	3	4	5	6
No. of families	7	10	16	25	18	11	8

Solution: calculation of Bowley's Coefficient of Skewness

Number of children per family X	**No. of families**	**c.f**
0	7	7
1	10	17
2	16	33
3	25	58
4	18	76
5	11	87
6	8	95

$Sk_B = Q_3 + Q_1 - 2$ Med. $/Q_3 - Q_1$

Q_1 = Size of N + ¼ th item = 95 + 1 = 24 th item, hence Q1 = 2

Q_3 = Size of 3 (N + 1)th item = 3 × 96/4 = 72th item

Size of 72th item is 4, hence Q3 = 4

Med. = Size of N + 1/2th item = 98/2 = 48th item.

Size of 48th item is 3. Hence median = 3

$Sk_B = 4 + 2 - 2\ (3)/4 - 2 = 0/2 = 0.$

Bowley Skewness Example

Q. Find the Bowley Skewness for the following set of data:

# of pets	# of families	cumulative freq
0	60	60
1	60	120
2	50	170
3	20	190
4	25	215
5	10	225
6 or more	5	230

Step 1: Find the Quartiles for the data set. You'll want to look for the "nth" observation using the following formulas:

Q_1 = (total cum freq + 1 / 4)th observation = (230 + 1 / 4) = 57.75

Q_2 = (total cum freq + 1 / 2)th observation = (230 + 1 / 2) = 115.5

Q_3 = 3 (total cum freq + 1 / 4)th observation = 3(230 + 1 / 4) = 173.25

Step 2: Look in your table to find the nth observations you calculated in Step 1:

Q_1 = 57.75th observation = 0

Q_2 = 115.5th observation = 1

Q_3 = 173.25th observation = 3

Step 3: Plug the above values into the formula:

$S_{kq} = Q_3 + Q_1 – 2Q_2 / Q_3 – Q_1$

$S_{kq} = 3 + 0 – 2 / 3 – 0 = 1/3$

S_{kq} = + 1/3, so the distribution is positively skewed.

Why Bowley Skewness works.

In a symmetric distribution, like the normal distribution, the first (Q_1) and third (Q_3) quartiles are at equal distances from the mean (Q_2). In other words, (Q_3 -Q_2) and (Q_2 -Q_1) will be equal. If you have a skewed distribution then there will be a difference between those two values.

Limitations of Bowley Skewness.

Bowley Skewness is an absolute measure of skewness. In other words, it's going to give you a result in the units that your distribution is in. That's compared to the Pearson Mode Skewness, which gives you results in a dimensionless unit — the standard deviation. This means that you cannot compare the skewness of different distributions with different units using Bowley Skewness.

Alternative Bowley Skewness formula.

According to Business Statistics, Bowley recognized that the Bowley Skewness formula could not be used to compare different distributions with different units. For example, you can't compare a distribution measured in heights in cm with one of weights in pounds. He offered an alternative formula. You should use this formula if you want to compare different distributions with different units:

Relative Skewness = ($Q_3 + Q_1$) – (2 * Median / $Q_3 – Q_1$).

2.3.4 Kelley's method

Kelly's Measure of Skewness is one of several ways to measure skewness in a data distribution. Bowley's skewness is based on the middle 50 percent of the observations in a data set. It leaves 25 percent of the observations in each tail of the distribution. Kelly suggested that leaving out fifty percent of data to calculate skewness was too extreme. He created a measure to find skewness with more data. Kelly's measure is based on P_{90} (the 90th percentile) and P_{10} (the 10th percentile). Only twenty percent of observations (ten percent in each tail) are excluded from the measure.

Bowled measure discussed above neglects the two extreme quarters of the data. It would be better for a measure to cover the entire data especially because in measuring scenes we are often interested in the more extreme items bowled measure can be extended by taking any two deciles equidistant from the median or any percentiles equidistant from the median. Kelly has suggested the following formula for measuring skewers upon the 10th and the 90th percentiles (or the first and ninth deciles):

Skk = P_{10} + P^{90} – 2 med. / P_{90} – P_{10} also SK = D_1 + D_9 – 2 med. / D_9 – D_1

Skk = Kelly coefficient of skewness.

This measure of skewness ha one theoretical attraction if skewness is to be based on percentiles however this method is not popular in practice and generally Karl Pearson method is used.

Measure of Skewers based on the third moment

A measure of Skewers may be obtained by making use of the third moment about the mean. This would be discussed under the head moments

Kelly's Measure Formula.

Kelley's measure of skewness is given in terms of percentiles and deciles (D). Kelley's absolute measure of skewness (S_k) is:

$$S_{K} = P_{90} + P_{10} - 2*P_{50} = D_{9} + D_{1} - 2*D_{5}.$$

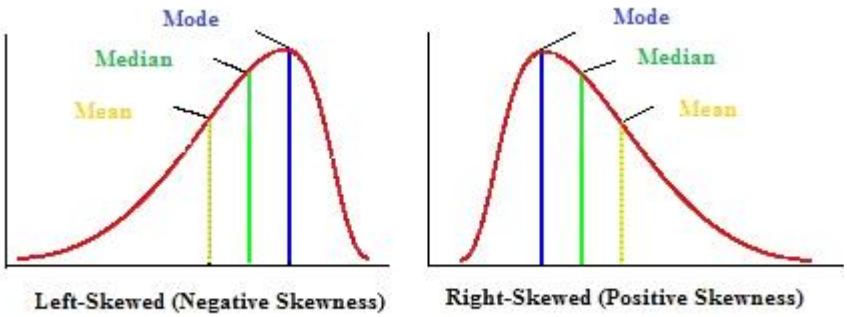
This formula is not practically used. In fact, Kelly's measure of skewness is rarely used at all, even in its more common form, which is measured as coefficient of skewness:

$$S_P = P_{90} - 2*P_{50} + P_{10} / P_{90} - P_{10}$$

What does Kelly's Measure of Skewness tell us?

Kelly's Measure of Skewness gives you the same information about skewness as the other three types of skewness measures (Bowley skewness, Pearson Mode Skewness and Momental Skewness.).

A measure of skewness = 0 means that the distribution is symmetrical.
A measure of skewness > 0 means a positive skewness.
A measure of skewness < means a negative skewness.



Illustration

Calculate Pearson coefficient of Skewers

X	12.5	17.5	22.5	27.5	32.5	37.5	42.5	47.5
f	28	42	54	108	129	61	45	33

Solution

X	***f***	**(X-27-5)/5 d**	***f*d**	***f*d^2**
12.5	28	-3	-84	252
17.5	42	-2	-84	168
22.5	54	-1	-54	54
27.5	108	0	0	0
32.5	129	+1	+129	129
37.5	61	+2	+122	244
42.5	45	+3	+135	405
47.5	33	+4	+132	528
	N = 500		**Σ *f*d = 296**	**Σ *f*d^2 = 1780**

Coefficient of SK = mean – mode/ σ

Mean X = A + σ fd / N x I

A = 27.5 ? fd = 296, N = 500 I = 5.

X = 27.5 + 296 / 500x 5 = 30.46

Mode: Since the maximum frequency is 129, the corresponding value of X, 32.5 is the modal value

S.D σ = $\sqrt{(\Sigma \text{ fd})/ N – (\Sigma \text{ fd} / N)^2}$ xi

Σ fd^2 = 1780 N = 500 Σ fd = 296 I = 5

Σ = $\sqrt{(1780/500) - (296/500)^2}$ x 5 = $\sqrt{(3.56- .35)}$ x 5 = 8.96

Coefficient of SK = 30.46 – 32.5 / 8.96 = - 0.228

Kurtosis

Kurtosis is another measure of the shape of a distribution. Whereas skewness measures the lack of symmetry of the frequency curve of a distribution, kurtosis is a measure of the relative peakedness of its frequency curve. Various frequency curves can be divided into three categories depending upon the shape of their peak. The three shapes are termed as Leptokurtic, Mesokurtic and Platykurhc.

A measure of kurtosis is given by ß2 = 2, a coefficient given by Karl Pearson ß2 .The value of ß2 = 3 for a mesokurtic curve. When ß2 > 3, the curve is more peaked than the mesokurtic curve and is termed as leptokurtic. Similarly, when ß2 < 3, the curve is less peaked than the mesokurtic curve and is called as platykurtic curve

Kurtosis is a statistical measure that's used to describe the distribution, or skewness, of observed data around the mean, sometimes referred to as the volatility of volatility. Kurtosis is used generally in the statistical field to describe trends in charts. Kurtosis can be present in a chart with fat tails and a low, even distribution, as well as be present in a chart with skinny tails and a distribution concentrated toward the mean.

Kurtosis is defined as a normalized form of the fourth central moment of a distribution. There are several flavours of kurtosis, the most commonly encountered variety of which is normally termed simply "the" kurtosis and is denoted or . The kurtosis

pf a theoretical distribution is defined by

$$\beta_2 \equiv \frac{\mu_4}{\mu_2^2}, \quad (1)$$

Where μ_i denotes the  th central moment (and in particular, μ_2 is the variance). This form is implemented in the Wolfram Language as Kurtosis [dist].

The kurtosis "excess" is denoted or , is defined by

$$\gamma_2 \equiv \beta_2 - 3 \quad (2)$$

$$= \frac{\mu_4}{\mu_2^2} - 3, \quad (3)$$

And is implemented in the Wolfram Language as Kurtosis Excess [dist]. Kurtosis excess is commonly used because of a normal distribution is equal to 0, while the kurtosis proper is equal to 3. Unfortunately, few authors confusingly refer to as the "excess or kurtosis."

For many distributions encountered in practice, a positive corresponds to a sharper peak with higher tails than if the distribution were normal. This observation is likely the reason kurtosis was historically (but incorrectly) regarded as a measure of the "peakedness" of a distribution. However, the correspondence between kurtosis and peakedness is not true in general; in fact, a distribution with a perfectly flat top may have infinite kurtosis, while one with infinite peakedness may have negative kurtosis excess. As a result, kurtosis provides a measure of outliers (i.e., the presence of "heavy tails") in a distribution, not its degree of peakedness.

The following table gives terms sometimes applied to different regimes of .

regime term

regime	term
$\gamma_2 < 0$	platykurtic
$\gamma_2 = 0$	mesokurtic
$\gamma_2 > 0$	leptokurtic

An estimator for the kurtosis excess is given by

$$g_2 = \frac{k_4}{k_2^2}, \qquad (4)$$

Where the s are k-statistics. For a normal distribution, the variance of this estimator is

$$\operatorname{var}(\hat{g}_2) \approx \frac{24}{N}. \qquad (5)$$

Breaking down 'kurtosis'

Put simply, kurtosis is a measure of the combined weight of a distribution's tails relative to the rest of the distribution. When a set of data is graphically depicted, it usually has a standard normal distribution, like a bell curve, with a central peak and thin tails. However, when kurtosis is present, the tails of the distribution are different than they would be under a normal bell-curved distribution.

Kurtosis is sometimes confused with a measure of the peakedness of a distribution. However, kurtosis is a measure that describes the shape of a distribution's tails in relation to its overall shape. A data set that shows kurtosis sometimes also displays skewness, or a lack of symmetry. However, kurtosis can be evenly distributed so that both its tails are equal.

Types of Kurtosis

There are three categories of kurtosis that can be displayed by a set of data. All measures of kurtosis are compared against a standard normal distribution, or bell curve.

The first category of kurtosis is a mesokurtic distribution. This type of kurtosis is the most similar to a standard normal distribution in that it also resembles a bell curve. However, a graph that is mesokurtic has fatter tails than a standard normal distribution and has a slightly lower peak. This type of kurtosis is considered normally distributed but is not a standard normal distribution.

The second category is a leptokurtic distribution. Any distribution that is leptokurtic displays greater kurtosis than a mesokurtic distribution. A characteristic of this type of distribution is one with extremely thick tails and a very thin and tall peak. The prefix of "lepto-" means "skinny," making the shape of a leptokurtic distribution easier to re-

member. T-distributions are leptokurtic.

The final type of distribution is a platykurtic distribution. These types of distributions have slender tails and a peak that's smaller than a mesokurtic distribution. The prefix of "platy-means broad and it is meant to describe a short and broad-looking peak. Uniform distributions are platykurtic.

2.3.6 Moments about Mean

In probability and statistics, a moment measure is a mathematical quantity, function or, more precisely, measure that is defined in relation to mathematical objects known as point processes, which are types of stochastic processes often used as mathematical models of physical phenomena represent able as randomly positioned points in time, space or both. Moment measures generalize the idea of (raw) moments of random variables, hence arise often in the study of point processes and related fields.

An example of a moment measure is the **first moment measure** of a point process, often called **mean measure** or **intensity measure,** which gives the expected or average number of points of the point process being located in some region of space. In other words, if the number of points of a point process located in some region of space is a random variable, then the first moment measure corresponds to the first moment of this random variable.

In mathematics, a moment is a specific quantitative measure, used in both mechanics and statistics, of the shape of a set of points. If the points represent mass, then the zeroth moment is the total mass, the first moment divided by the total mass is the centre of mass, and the second moment is the rotational inertia. If the points represent probability density, then the zeroth moment is the total probability (i.e. one), the first moment is the mean, the second central moment is the variance, the third moment is the skewness, and the fourth moment (with normalization and shift) is the kurtosis. The mathematical concept is closely related to the concept of moment in physics.

Suppose that we have a set of data with a total of n discrete points. One important calculation, which is actually several numbers, is called the sth moment. The sth moment of the data set with values x1, x2, x3, . . . , xn is given by the formula:

(x1s + x2s + x3s + . . . + xns)/n

Using this formula requires us to be careful with our order of operations. We need to

do the exponents first, add, then divide this sum by n the total number of data values.

It can also be explained below

For a set of N numbers comprised of X1, X2, ..., XN, the kth moment about the mean is defined as:

$$m_k = \frac{(X_1 - \bar{X})^k + (X_2 - \bar{X})^k + \ldots + (X_N - \bar{X})^k}{N} = \frac{\sum_{j=1}^{N} (X_j - \bar{X})^k}{N}$$

The first moment about the mean, ?1, is zero.

The second moment about the mean, ?2, represents the variance, and is usually denoted ?2, where ? represents the standard deviation.

A Note on the Term Moment

The term moment has been taken from physics. In physics the moment of a system of point masses is calculated with a formula identical to that above, and this formula is used in finding the centre of mass of the points. In statistics the values are no longer masses, but as we will see, moments in statistics still measure something relative to the centre of the values.

For a set of N numbers comprised of X1, X2, ..., XN, the kth moment (also known as kthmoment about zero) is defined as:

$$k^{\text{th}} \text{ moment} = \frac{X_1^k + X_2^k + \ldots + X_N^k}{N} = \frac{\sum_{j=1}^{N} X_j^k}{N}$$

Note: The first moment (i.e., n = 1) equals the arithmetic mean.

Example:

Find the first, second, and third moments for the set of numbers 1, 2, 6, and 7.

Solution:

$$\text{first moment} = \frac{(1 + 2 + 6 + 7)}{4} = \frac{16}{4} = 4$$

$$\text{second moment} = \frac{(1^2 + 2^2 + 6^2 + 7^2)}{4} = \frac{90}{4} = 22.5$$

$$\text{third moment} = \frac{(1^3 + 2^3 + 6^3 + 7^3)}{4} = \frac{568}{4} = 142$$

First Moment

For the first moment we set s = 1. The formula for the first moment is thus:

(x1 x2 + x3 + . . + xn)/n.

This is identical to the formula for the sample mean.

The first moment of the values 1, 3, 6, 10 is (1 + 3 + 6 + 10) / 4 = 20/4 = 5.

Second Moment

For the second moment we set s = 2. The formula for the second moment is:

(x12 + x22 + x32 + . . . + xn2)/n

The second moment of the values 1, 3, 6, 10 is (12 + 32 + 62 + 102) / 4 = (1 + 9 + 36 + 100)/4 = 146/4 = 36.5.

Third Moment

For the third moment we set s = 3. The formula for the third moment is:

(x13 + x23 + x33 + . . . + xn3)/n

The third moment of the values 1, 3, 6, 10 is (13 + 33 + 63 + 103) / 4 = (1 + 27 + 216 + 1000)/4 = 1244/4 = 311.

Higher moments can be calculated in a similar way. Just replace s in the above formula with the number denoting the desired moment

Moments about the Mean

A related idea is that of the sth moment about the mean. In this calculation we perform the following steps:

1 First calculate the mean of the values.

2 Next, subtract this mean from each value.

3 Then raise each of these differences to the sth power.

4 Now add the numbers from step #3 together.

5 Finally, divide this sum by the number of values we started with.

The formula for the sth moment about the mean m of the values values x1, x2, x3, . . . , xn is given by:

ms = ((x1 - m)s + (x2 - m)s + (x3 - m)s + . . . + (xn - m)s)/n

First Moment about the Mean

The first moment about the mean is always equal to zero, no matter what the data set is that we are working with. This can be seen in the following:

m1 = ((x1 - m) + (x2 - m) + (x3 - m) + . . . + (xn - m))/n = ((x1+ x2 + x3 + . . . + xn) **- nm)/n = m - m = 0.**

Second Moment about the Mean

The second moment about the mean is obtained from the above formula by settings = 2:

m2 = ((x1 - m)2 + (x2 - m)2 + (x3 - m)2 + . . . + (xn - m)2)/n

This formula is equivalent to that for the sample variance.

For example, consider the set 1, 3, 6, 10. We have already calculated the mean of this set to be 5. Subtract this from each of the data values to obtain differences of:

1 - 5 = -4

3 - 5 = -2

6 - 5 = 1

10 - 5 = 5

We square each of these values and add them together: (-4)2 + (-2)2 + 12 + 52 = 16 + 4 + 1 + 25 = 46. Finally divide this number by the number of data points: 46/4 = 11.5

Applications of Moments

As mentioned above, the first moment is the mean and the second moment about the mean is the sample variance. Pearson introduced the use of the third moment about the mean in calculating skewness and the fourth moment about the mean in the calculation of kurtosis.

Example:

Find the first, second, and third moments about the mean for the set of numbers 1, 4, 6, and 9.

Solution:

$$\bar{X} = \frac{(1 + 4 + 7 + 12)}{4} = \frac{24}{4} = 6$$

(By definition, $\mu_1 = 0$, however, will include its computation)

$$m_1 = \frac{(1-6) + (4-6) + (7-6) + (12-6)}{4} = \frac{0}{4} = 0$$

$$m_2 = \frac{(1-6)^2 + (4-6)^2 + (7-6)^2 + (12-6)^2}{4}$$

$$m_2 = \frac{(25 + 4 + 1 + 36)}{4}$$

$$m_2 = \frac{66}{4}$$

$$m_2 = 16.5$$

$$m_3 = \frac{(1-6)^3 + (4-6)^3 + (7-6)^3 + (12-6)^3}{4}$$

$$m_3 = \frac{(-125 - 8 + 1 + 216)}{4}$$

$$m_3 = \frac{84}{4}$$

$$m_3 = 21$$

A probability distribution may be characterized by its moments. The rth moment of x about some fixed point x0 is defined as the expectation value of (x - x0)r where r is an integer. An analogy may be drawn here with the moments of a mass distribution in mechanics. In such a case, P(x) plays the role of the mass density.

In practice, only the first two moments are of importance. And, indeed, many problems are solved with only knowledge of these two quantities. The most important is the first moment about zero,

$$\mu = E[x] = \int x\, P(x)\, dx. \qquad (8)$$

This can be recognized as simply the mean or average of x. If the analogy with mass moments is made, the mean thus represents the "centre of mass" of the probability distribution.

It is very important here to distinguish the mean as defined in (Equation 8) from the mean which one calculates from a set of repeated measurements. The first refers to the theoretical mean, as calculated from the theoretical distribution, while the latter is an experimental mean taken from a sample. As we shall see, the sample mean is an estimate of the theoretical mean. Throughout the remainder of this chapter, we shall always use the Greek letter μ to designate the theoretical mean.

The second characteristic quantity is the second moment about the mean (also known as the second central moment),

$$\sigma^2 = E[(x-\mu)^2] = \int (x-\mu)^2\, P(x)\, dx. \qquad (9)$$

This is commonly called the variance and is denoted as 2. The square root of the variance, , is known as the standard deviation. As can be seen from (9), the variance is the average squared deviation of x from the mean. The standard deviation, , thus measures the dispersion or width of the distribution and gives us an idea of how much the random variable x fluctuates about its mean. Like μ, (9) is the theoretical variance and should be distinguished from the sample variance to be discussed.

Further moments, of course, may also be calculated, such as the third moment about the mean. This is known as the skewness and it gives a measure of the distribution's symmetry or asymmetry. It is employed on rare occasions, but very little information

is generally gained from this moment or any of the following ones.

2.3.7 Moments about origin

A **moment** about the origin is sometimes called a raw moment. Note that μ1 = E(X) = μX, the mean of the distribution of X, or simply the mean of X. The rth moment is sometimes written as function of ? where ? is a vector of parameters that characterize the distribution of X.

The **first moment about the origin** is the mean. The phrase "about the origin" merely means that we are looking at differences between the xi and the origin, zero; that is, xi ? 0 is nothing more than xi. We now generalize the idea:

$$m_1 ? \frac{(\sum x1)}{N}$$

$$m_2 ? \frac{(\sum x2)}{N}$$

$$m_3 ? \frac{(\sum x3)}{N}$$

And so on. The symbol m1? is called the first moment (about the origin) and is nothing more than our old friend the mean, ? x; m2? is called the second moment about the origin; and m3? is the third moment about the origin.

2.4 Introduction of Time Series

A time series is a collection of observations of well-defined data items obtained through repeated measurements over time. For example, measuring the value of retail sales each month of the year would comprise a time series. This is because sales revenue is well defined, and consistently measured at equally spaced intervals. Data collected irregularly or only once are not time series. An observed time series can be decomposed into three components: the trend (long term direction), the seasonal (systematic, calendar related movements) and the irregular(unsystematic, short term fluctuations). The simplest form of data is a longish series of continuous measurements at equally spaced time points. That is observations are made at distinct points in time, these time points being equally spaced and, the observations may take values from a

continuous distribution. The above setup could be easily generalized: for example, the times of observation need not be equally spaced in time; the observations may only take values from a discrete distribution. If we repeatedly observe a given system at regular time intervals, it is very likely that the observations we make will be correlated. So we cannot assume that the data constitute a random sample. The time-order in which the observations are made is vital. Time series data often arise when monitoring industrial processes or tracking corporate business metrics. Time series analysis accounts for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for.

Time series are very frequently plotted via line charts. Time series are used in statistics, signal processing, pattern recognition, econometrics, mathematical finance, weather forecasting, intelligent transport, and trajectory forecasting, earthquake prediction, astronomy, communications engineering, control engineering, and largely in any domain of applied science and engineering that involves temporal measurements. Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values. While regression analysis is often employed in such a way as to test theories that the current values of one or more independent time series affect the current value of another time series, this type of analysis of time series is not called "time series analysis", which focuses on comparing values of a single time series or multiple dependent time series at different points in time.

Time series data have a natural temporal ordering. This makes time series analysis distinct from cross-sectional studies, in which there is no natural ordering of the observations (e.g. explaining people's wages by reference to their respective education levels, where the individuals' data could be entered in any order). Time series analysis is also distinct from spatial data analysis where the observations typically relate to geographical locations (e.g. accounting for house prices by the location as well as the intrinsic characteristics of the houses). A stochastic model for a time series will generally reflect the fact that observations close together in time will be more closely related than observations further apart. In addition, time series models will often make

use of the natural one-way ordering of time so that values for a given period will be expressed as deriving in some way from past values, rather than from future values. Time series analysis can be beneficial to see how a given asset, security or economic variable changes over time or how it changes compared to other variables over the same time period. Time series analysis can be also applied to real-valued, continuous data, discrete numeric data, or discrete symbolic data.

Objectives of Time Series Analysis:

- Description - summary statistics, graphs
- Analysis and Interpretation - find a model to describe the time dependence in the data, can we interpret the model
- Forecasting or Prediction - given a sample from the series, forecast the next value, or the next few values
- Control - adjust various control parameters to make the series fit closer to a target
- Adjustment - in a linear model the errors could form a time series of correlated observations, and we might want to adjust estimated variances to allow for this.

2.4.1 Concept of Time Series

In plain English, a time series is simply a sequence of numbers collected at regular intervals over a period of time. In statistics, a time series is a sequence of numerical data points in successive order, usually occurring in uniform intervals. This concerns the analysis of data collected over time, such as weekly values, monthly values, quarterly values, yearly values, etc. Many statistical methods relate to data which are independent, or at least uncorrelated. There are many practical situations where data might be correlated. This is particularly so where repeated observations on a given system are made sequentially in time. Data gathered sequentially in time are called a time series. Moreover, time series is a series of data points listed (or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data. Examples of time series are heights of ocean tides, counts of sunspots, and the daily closing value of the Dow Jones Industrial Average.

Various examples in which time series arise:

- Economics and Finance
- Environmental Modeling
- Meteorology and Hydrology
- Demographics
- Medicine
- Engineering
- Quality Control

Types of time Series

1. Continuous

2. Discrete

Discrete: means that observations are recorded in discrete times - it says nothing about the nature of the observed variable. The time intervals can be annually, quarterly, monthly, weekly, daily, and hourly, etc.

Continuous: means that observations are recorded continuously for example temperature and/or humidity in some laboratory. Again, time series can be continuous regardless of the nature of the observed variable.

Discrete time series can result when continuous time series are sampled. Sometimes quantities that don't have an instantaneous value get aggregated also resulting in a discrete time series e.g. daily rainfall. We will mostly study discrete time series in this course. Note that discrete time series are often the result of discretization of continuous time series (e.g. monthly rainfall).

Importance or Uses of Time Series

There are two main uses of time series analysis: (a) identifying the nature of the phenomenon represented by the sequence of observations, and (b) forecasting (predicting future values of the time series variable). Both of these goals require that the pattern of observed time series data is identified and more or less formally described. Once the pattern is established, we can interpret and integrate it with other data (i.e., use it in our theory of the investigated phenomenon, e.g., seasonal commodity prices).

Regardless of the depth of our understanding and the validity of our interpretation (theory) of the phenomenon, we can extrapolate the identified pattern to predict future events.

The usage of time series models is twofold:

i) Obtain an understanding of the underlying forces and structure that produced the observed data

ii) Fit a model and proceed to forecasting, monitoring or even feedback and feedforward control.

Time Series Analysis is used for many applications such as:

- Economic Forecasting
- Sales Forecasting
- Budgetary Analysis
- Stock Market Analysis
- Yield Projections
- Process and Quality Control
- Inventory Studies
- Workload Projections
- Utility Studies
- Census Analysis

Time series analysis can be useful to see how a given asset, security or economic variable changes over time or how it changes compared to other variables over the same time period. For example, in stock market investments, suppose you wanted to analyze a time series of daily closing stock prices for a given stock over a period of one year. You would obtain a list of all the closing prices for the stock over each day for the past year and list them in chronological order. This would be a one-year, daily closing price time series for the stock. Delving a bit deeper, you might be interested to know if a given stock's time series shows any seasonality, meaning it goes through peaks and valleys at regular times each year. Or you might want to know how a stock's share price changes as an economic variable, such as the unemployment rate,

changes.

The analysis of time series is of great significance not only to the economists and business man but also to the scientist, astronomist, geologist etc. for the reasons given below.

1) It helps in understanding past behavior: It helps to understand what changes have taken place in the past. Time series are very helpful in study of past behavior of business. On this basis, we can invest our money in that type of business. It is duty of businessman to make time series of past sale or profit and see what is the trend of sale or profit in that type of business. Such analysis is helpful in predicting the future behavior.

2) It helps in planning future operations: Statistical techniques have been evolved which enable time series to be analyzed in such a way that the influences which have determined the form of that series may be ascertained. If the regularity of occurrence of any feature over a sufficient long period could be clearly established then. Within limits, prediction of probable future variations would become possible. Forecasting is science of estimation. Today is the day of competition so if you have to win from competition then you must learn this science, this science can be utilized if we make time series and on the basis we can read the history and then we can decide what happen in future. Suppose if we can make the time series of past strategy of our competitor then on this basis we can estimate future strategy of our competitor and on this base we can change our strategy for defeating our competitor.

3) It helps in evaluating current accomplishments: The actual performance can be compared with the expected performance and the cause of variation analysed. Time series is an equipment in your hand on this basis you can evaluate your business achievements if you did good , your performance shows your good face in the time series by up-word trend of your performance. If your business performance is very bad then you can make new policies to stable your business. For example, if expected sale for 2000-01 was 10,000 washing machines and the actual sale was only 9000. One can investigate the cause for the shortfall in achievement.

4) It facilitates comparison: If we can calculate our two or more branches time series then we can compare the performance of our branches. On their performance

we can give those rewards. Different time series are often compared and important conclusions drawn there from.

2.4.2 Components of Time Series

The fluctuations of time series can be classified into four basic types of variations, They are often called components or elements of a time series. They are:

(1) Secular Trend or Long Term Movements (T)

(2) Seasonal Variations (S)

(3) Cyclical Variations (C)

(4) Irregular Variations (I)

The value (y) of a phenomenon observed at any point of time (t) is the net effect of all the above mentioned categories of components of a time series. We will see them in detail here.

(1) Secular Trend

The secular trend is the main component of a time series which results from long term effect of socio-economic and political factors. This trend may show the growth or decline in a time series over a long period. This is the type of tendency which continues to persist for a very long period. Prices, export and imports data, for example, reflect obviously increasing tendencies over time.

(2) Seasonal Variations (Seasonal Trend)

These are short term movements occurring in a data due to seasonal factors. The short term is generally considered as a period in which changes occur in a time series with variations in weather or festivities. For example, it is commonly observed that the consumption of ice-cream during summer us generally high and hence sales of an ice-cream dealer would be higher in some months of the year while relatively lower during winter months. Employment, output, export etc. are subjected to change due to variation in weather. Similarly sales of garments, umbrella, greeting cards and fire-work are subjected to large variation during festivals like Onam, Eid, Christmas, New Year etc. These types of variation in a time series are isolated only when the series is provided biannually, quarterly or monthly.

(3) Cyclical Variations (Cyclical Variations)

These are long term oscillation occurring in a time series. These oscillations are mostly observed in economics data and the periods of such oscillations are generally extended from five to twelve years or more. These oscillations are associated to the well known business cycles. These cyclic movements can be studied provided a long series of measurements, free from irregular fluctuations is available.

The cyclical component of a time series refers to (regular or periodic) fluctuations around the trend, excluding the irregular component, revealing a succession of phases of expansion and contraction. The cyclical component can be viewed as those fluctuations in a time series which are longer than a given threshold, e.g. 1½ years, but shorter than those attributed to the trend.

(4) Irregular Variations (Irregular Fluctuations)

These are sudden changes occurring in a time series which are unlikely to be repeated, it is that component of a time series which cannot be explained by trend, seasonal or cyclic movements. It is because of this fact these variations some-times called residual or random component. These variations though accidental in nature, can cause a continual change in the trend, seasonal and cyclical oscillations during the forthcoming period. Floods, fires, earthquakes, revolutions, epidemics and strikes etc are the root cause of such irregularities.

Irregular variations or random variations constitute one of four components of a time series. They correspond to the movements that appear irregularly and generally during short periods.

Irregular variations do not follow a particular model and are not predictable. In practice, all the components of time series that cannot be attributed to the influence of cyclic fluctuations or seasonal variations or those of the secular tendency are classed as irregular.

2.5 Trend analysis

Trend analysis is the rampant practice of collecting information and attempting to spot a pattern, or trend, in the information. In some fields of study, the term "trend analysis" has more formally defined meanings? In statistics, trend analysis often refers to techniques for extracting an underlying pattern of behaviour in a time series which would otherwise be partly or nearly completely hidden by noise. A simple description of

these techniques is trend estimation, which can be undertaken within a formal regression analysis. Today, trend analysis often refers to the science of studying changes in social patterns, including fashion, technology and consumer behaviour.

A trend analysis is an aspect of technical analysis that tries to predict the future movement of a stock based on past data. Trend analysis is based on the idea that what has happened in the past gives traders an idea of what will happen in the future. There are three main types of trends: short-, intermediate- and long-term. Trend analysis tries to predict a trend such as a bull market run, and ride that trend until data suggests a trend reversal, such as a bull-to-bear market. Trend analysis is helpful because moving with trends, and not against them, will lead to profit for an investor.

A trend is the general direction the market is taking during a specified period of time. Trends can be both upward and downward, relating to bullish and bearish markets, respectively. While there is no specified minimum amount of time required for a direction to be considered a trend, the longer the direction is maintained, the more notable the trend.

Trend analysis is the process of trying to look at current trends in order to predict future ones and is considered a form of comparative analysis. This can include attempting to determine whether a current market trend, such as gains in a particular market sector, is likely to continue, as well as whether a trend in one market area could result in a trend in another. Though an analysis may involve a large amount of data, there is no guarantee that the results will be correct.

Measurement of Trend: Moving Average and The Method of Least Squares:

Mean of time series data (observations equally spaced in time) from several consecutive periods, is called 'moving' because it is continually recomputed as new data becomes available; it progresses by dropping the earliest value and adding the latest value. For example, the moving average of six-month sales may be computed by taking the average of sales from January to June, then the average of sales from February to July, then of March to August, and so on.

Moving averages (1) reduce the effect of temporary variations in data, (2) improve the 'fit' of data to a line (a process called 'smoothing') to show the data's trend more clearly, and (3) highlight any value above or below the trend.

2.5.1 Method of Least Squares

Least Squares Method is a statistical technique to determine the line of best fit for a model. The least squares method is specified by an equation with certain parameters to observed data. This method is extensively used in regression analysis and estimation.

In the most common application - linear or ordinary least squares - a straight line is sought to be fitted through a number of points to minimize the sum of the squares of the distances (hence the name "least squares") from the points to this line of best fit. In contrast to a linear problem, a non-linear least squares problem has no closed solution and is generally solved by iteration. The earliest description of the least squares method was by Carl Freidrich Gauss in 1795.

Field data is often accompanied by noise. Even though all control parameters (independent variables) remain constant, the resultant outcomes (dependent variables) vary. A process of quantitatively estimating the trend of the outcomes, also known as regression or curve fitting, therefore becomes necessary.

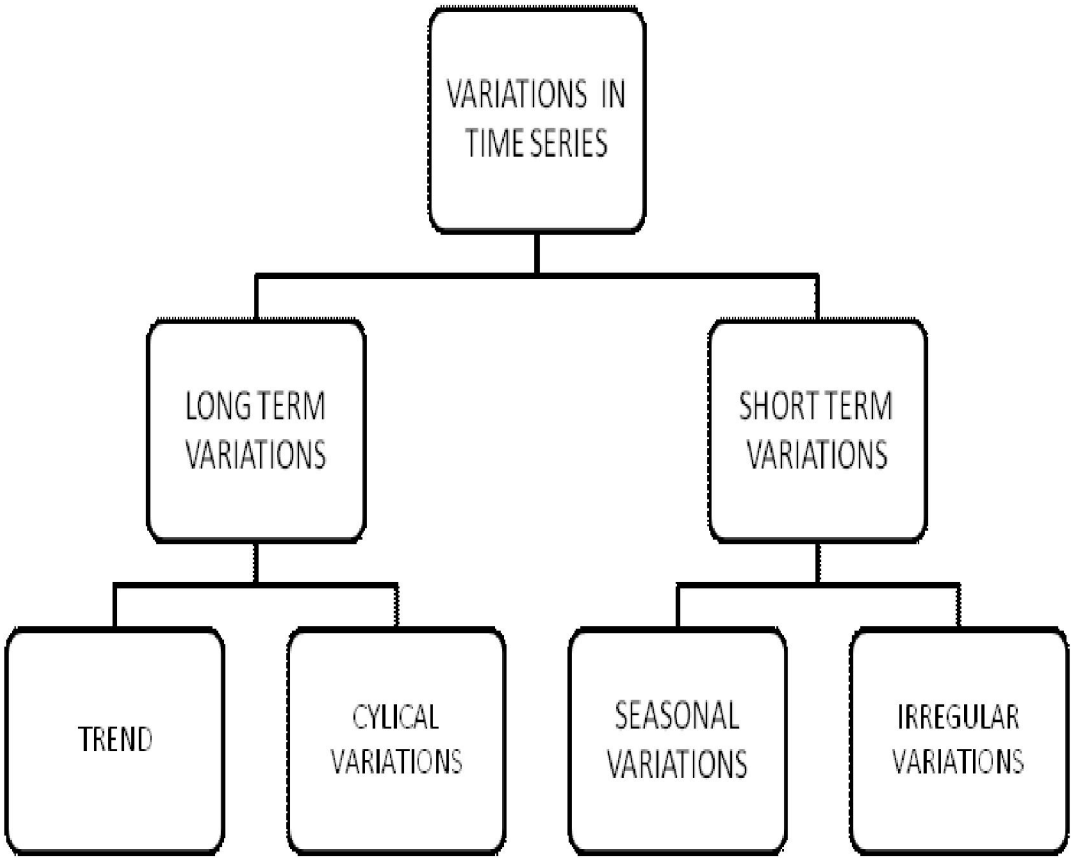
The curve fitting process fits equations of approximating curves to the raw field data. Nevertheless, for a given set of data, the fitting curves of a given type are generally NOT unique. Thus, a curve with a minimal deviation from all data points is desired. This best-fitting curve can be obtained by the method of least squares.

The principle of least squares provides us an analytical or mathematical device to obtain an objective fit to the trend of the given time series. Most of the data relating to economic and business time series conform to definite laws of growth or predictions. This technique can be used to fit linear as well as nonlinear trends.

Methods of time series

Time series is set of data collected and arranged in accordance of time. According to Croxton and Cowdon,"A Time series consists of data arranged chronologically." Such data may be series of temperature of patients, series showing number of suicides in different months of year etc. The analysis of time series means separating out different components which influences values of series. The variations in the time series can be divided into two parts: long term variations and short term variations. Long term variations can be divided into two parts: Trend or Secular Trend and Cyclical variations.

Short term variations can be divided into two parts: Seasonal variations and Irregular Variations.



Methods for time series analysis

In business forecasting, it is important to analyze the characteristic movements of variations in the given time series. The following methods serve as a tool for this analysis:

1 Methods for Measurement of Secular Trend

i) Freehand curve Method (Graphical Method)

ii) Method of selected points

iii) Method of semi-averages

iv) Method of moving averages

v) Method of Least Squares

2 Methods for Measurement of Seasonal Variations

i) Method of Simple Average

ii) Ratio to Trend Method

iii) Ratio to Moving Average Method

iv Method of Link Relatives

3 Methods for Measurement for Cyclical Variations

4 Methods for Measurement for Irregular Variations

Methods for measurement of secular trend

The following are the principal methods of measuring trend from given time series:

I. Graphical or free hand curve method

This is the simple method of studying trend. In this method the given time series data are plotted on graph paper by taking time on X-axis and the other variable on Y-axis. The graph obtained will be irregular as it would include short-run oscillations. We may observe the up and down movement of the curve and if a smooth freehand curve is drawn passing approximately to all points of a curve previously drawn, it would eliminate the short-run oscillations (seasonal, cyclical and irregular variations) and show the long-period general tendency of the data. This is exactly what is meant by Trend. However, It is very difficult to draw a freehand smooth curve and different persons are likely to draw different curves from the same data. The following points must be kept in mind in drawing a freehand smooth curve:

1 That the curve is smooth.

2 That the numbers of points above the line or curve are equal to the points below it.

3 That the sum of vertical deviations of the points above the smoothed line is equal to the sum of the vertical deviations of the points below the line. In this way the positive deviations will cancel the negative deviations. These deviat ions are the effects of seasonal cyclical and irregular variations and by this

process they are eliminated.

4 The sum of the squares of the vertical deviations from the trend line curve is minimum. (This is one of the characteristics of the trend line fitted by the method of lest squares)

The trend values can be read for various time periods by locating them on the trend line against each time period.

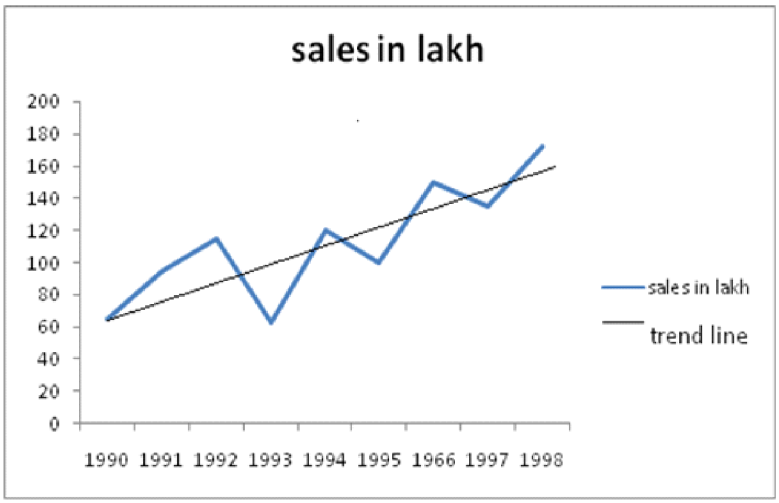
The following example will illustrate the fitting of a freehand curve to set of time series values:

Example:

The table below shows the data of sale of nine years:-

Year	1990	1991	1992	1993	1994	1995	1996	1997	1998
Sales in (lakh units)	65	95	115	63	120	100	150	135	172

If we draw a graph taking year on x-axis and sales on y-axis, it will be irregular as shown below. Now drawing a freehand curve passing approximately through all this points will represent trend line (shown below by black line).



Merits:

1 It is simple method of estimating trend which requires no mathematical calcula tions.

2 It is a flexible method as compared to rigid mathematical trends and, therefore, a better representative of the trend of the data.

3 This method can be used even if trend is not linear.

4 If the observations are relatively stable, the trend can easily be approximated by this method.

5 Being a non mathematical method, it can be applied even by a common man.

Demerits:

1 It is subjective method. The values of trend, obtained by different statisticians would be different and hence, not reliable.

2 Predictions made on the basis of this method are of little value.

II. Method of selected points

In this method, two points considered to be the most representative or normal, are joined by straight line to get secular trend. This, again, is a subjective method since different persons may have different opinions regarding the representative points. Further, only linear trend can be determined by this method.

III. Method of semi-averages

Under this method, as the name itself suggests semi-averages are calculated to find out the trend values. By semi-averages is meant the averages of the two halves of a series. In this method, thus, the given series is divided into two equal parts (halves) and the arithmetic mean of the values of each part (half) is calculated. The computed means are termed as semi-averages. Each semi-average is paired with the centre of time period of its part. The two pairs are then plotted on a graph paper and the points are joined by a straight line to get the trend. It should be noted that if the data is for even number of years, it can be easily divided into two halves. But if it is for odd number of years, we leave the middle year of the time series and two halves constitute the periods on each side of the middle year.

Merits:

1 It is simple method of measuring trend.

2 It is an objective method because anyone applying this to a given data would get identical trend value.

Demerits:

1 This method can give only linear trend of the data irrespective of whether it exists or not.

2 This is only a crude method of measuring trend, since we do not know whether the effects of other components are completely eliminated or not.

IV. Method of moving average

This method is based on the principle that the total effect of periodic variations at different points of time in its cycle gets completely neutralized, i.e. ?St = 0 in one year and ?Ct = 0 in the periods of cyclical variations.

In the method of moving average, successive arithmetic averages are computed from overlapping groups of successive values of a time series. Each group includes all the observations in a given time interval, termed as the period of moving average. The next group is obtained by replacing the oldest value by the next value in the series. The averages of such groups are known as the moving averages. The moving averages of a group are always shown at the centre of its period.

The process of computing moving averages smoothens out the fluctuations in the time series data. It can be shown that if the trend is linear and the oscillatory variations are regular, the moving average with the period equal to the period of oscillatory variations would get minimized because the average of a number of observations must lie between the smallest and the largest observation. It should be noted that the larger is the period of moving average the more would be the reduction in the effect of random components but the more information is lost at the two ends of data .i.e. It reduces the curvature of curvi-linear trends.

When the trend is non linear, the moving averages would give biased rather than the actual trend values.

Suppose that the successive observations are taken at equal intervals of time, say, yearly are Y1, Y2, Y3 , . . .

Moving Average when the period is Odd

Now by a three-yearly moving averages, we shall obtain average of first three consecutive years (beginning with the second year) and place it against time t=2; then

the average of the next three consecutive years (beginning with the second year) and place it against time t=3, and so on. This is illustrated below:

Time	Observations	Moving Total	Moving Average
(t)	Y_t		(3 Years)
1	Y_1		
2	Y_2	$Y_1 + Y_2 + Y_3$	? $(Y_1 + Y_2 + Y_3)$
3	Y_3	$Y_2 + Y_3 + Y_4$	? $(Y_2 + Y_3 + Y_4)$
4	Y_4	$Y_3 + Y_4 + Y_5$	? $(Y_3 + Y_4 + Y_5)$
5	Y_5		

It should be noted that for odd period moving average, it is not possible to get the moving averages for the first and the last periods.

Moving Average when the period is Even

For an even order moving average, two averaging processes are necessary in order to centre the moving average against periods rather than between periods. For example , for a four -yearly moving average we shall first obtain the average Y1=1/4(Y1 + Y2 + Y3 + Y4) of the first four years and place it in between t =2 and t=3 then the average Y2 = 1/4(Y2 + Y3 + Y4 + Y5) of the next four years is and place it in between t=3 and t=4 , and finally obtain the average ½(Y1 + Y2) of the two averages and place it against time t=3. Thus the moving average is brought against time or period rather than between periods. The same procedure is repeated for further results.

This is tabulated below:

Time (t)	Observations Y_t	Moving Average for 4-period	Cantered Value
1	Y_1		
2	Y_2		
		? $¼(Y_1 + Y_2 + Y_3 + Y_4) = A_1$	
3	Y_3		? $½(A_1 + A_2)$
		? $¼(Y_2 + Y_3 + Y_4 + Y_5) = A_2$	
4	Y_4		

It should be noted that when the period of moving average is even, the computed average will correspond to the middle of the two middle most periods.

Merits:

1 This method is easy to understand and easy to use because there are no mathematical complexities involved.

2 It is an objective method in the sense that anybody working on a problem with the method will get the same trend values. It is in this respect better than the free hand curve method.

3 It is a flexible method in the sense that if a few more observations are added, the entire calculations are not changed. This not with the case of semi-average method.

4 When the period of oscillatory movements is equal to the period of moving average, these movements are completely eliminated.

5 By the indirect use of this method, it is also possible to isolate seasonal, cyclical and random components.

Demerits:

1 It is not possible to calculate trend values for all the items of the series. Some information is always lost at its ends.

2 This method can determine accurate values of trend only if the oscillatory and the random fluctuations are uniform in terms of period and amplitude and the trend is, at least, approximately linear. However, these conditions are rarely met in practice. When the trend is not linear, the moving averages will not give correct values of the trend.

3 The trend values obtained by moving averages may not follow any math ematical pattern i.e. fails in setting up a functional relationship between the values of X(time) and Y(values) and thus, cannot be used for forecasting which perhaps is the main task of any time series analysis.

4 The selection of period of moving average is a difficult task and a great deal of care is needed to determine it.

5 Like arithmetic mean, the moving averages are too much affected by extreme values.

V. Method of least squares

This is one of the most popular methods of fitting a mathematical trend. The fitted trend is termed as the best in the sense that the sum of squares of deviations of observations, from it, is minimized. This method of Least squares may be used either to fit linear trend or a non-linear trend (Parabolic and Exponential trend).

A statistical method used to determine a line of best fit by minimizing the sum of squares created by a mathematical function. A "square" is determined by squaring the distance between a data point and the regression line. The least squares approach limits the distance between a function and the data points that a function is trying to explain. It is used in regression analysis, often in nonlinear regression modelling in which a curve is fit into a set of data.

The least squares method is a form of mathematical regression analysis that finds the line of best fit for a dataset, providing a visual demonstration of the relationship between the data points. Each point of data is representative of the relationship be-

tween a known independent variable and an unknown dependent variable.

The method of least squares is a standard approach in regression analysis to the approximate solution of over determined systems, i.e., sets of equations in which there are more equations than unknowns. "Least squares" means that the overall solution minimizes the sum of the squares of the errors made in the results of every single equation.

The most important application is in data fitting. The best fit in the least-squares sense minimizes the sum of squared residuals, a residual being the difference between an observed value and the fitted value provided by a model. When the problem has substantial uncertainties in the independent variable (the x variable), then simple regression and least squares methods have problems; in such cases, the methodology required for fitting errors-in-variables models may be considered instead of that for least squares.

Least squares problems fall into two categories: linear or ordinary least squares and non-linear least squares, depending on whether or not the residuals are linear in all unknowns. The linear least-squares problem occurs in statistical regression analysis; it has a closed-form solution. The non-linear problem is usually solved by iterative refinement; at each iteration the system is approximated by a linear one, and thus the core calculation is similar in both cases.

Polynomial least squares describe the variance in a prediction of the dependent variable as a function of the independent variable and the deviations from the fitted curve.

When the observations come from an exponential family and mild conditions are satisfied, least-squares estimates and maximum-likelihood estimates are identical.The method of least squares can also be derived as a method of moments estimator.

The following discussion is mostly presented in terms of linear functions but the use of least-squares is valid and practical for more general families of functions. Also, by iteratively applying local quadratic approximation to the likelihood (through the Fisher information), the least-squares method may be used to fit a generalized linear model.

For the topic of approximating a function by a sum of others using an objective function based on squared distances, see least squares (function approximation).

The least-squares method is usually credited to Carl Friedrich Gauss (1795), but it

was first published by Adrien-Marie Legendre.

The least squares method provides the overall rationale for the placement of the line of best fit among the data points being studied. The most common application of the least squares method, referred to as linear or ordinary, aims to create a straight line that minimizes the sum of the squares of the errors generated by the results of the associated equations, such as the squared residuals resulting from differences in the observed value and the value anticipated based on the model.

This method of regression analysis begins with a set of data points to be graphed. An analyst using the least squares method will be seeking a line of best fit that explains the potential relationship between an independent variable and a dependent variable. In regression analysis, dependent variables are designated on the vertical Y axis and independent variables are designated on the horizontal X axis. These designations will form the equation for the line of best fit, which is determined from the least squares method.

Example of Least Squares Method

For example, an analyst may want to test the relationship between a company's stock returns and the index returns for which the stock is a component. In this example, the analyst seeks to test the dependence of the stock returns on the index returns. To do this, all of the returns are plotted on a chart. The index returns are then designated as the independent variable, and the stock returns are the dependent variable. The line of best fit provides the analyst, with coefficients explaining the level of dependence.

Fitting of linear trend

Given the data (Yt, t) for n periods, where t denotes time period such as year, month, day, etc. We have to the values of the two constants, 'a' and 'b' of the linear trend equation:

$Yt = a + bt$

Where the value of 'a' is merely the Y-intercept or the height of the line above origin. That is, when X=0, Y= a. The other constant 'b' represents the slope of the trend line. When b is positive, the slope is upwards, and when b is negative, the slope is downward.

This line is termed as the line of best fit because it is so fitted that the total distance of deviations of the given data from the line is minimum. The total of deviations is calculated by squaring the difference in trend value and actual value of variable. Thus, the term "Least Squares" is attached to this method.

Using least square method, the normal equation for obtaining the values of a and b is:

$\sum Yt = na + b\sum t$

$\sum tYt = a\sum X + b \sum t2$

Let X = t - A, such that ?X = 0, where A denotes the year of origin.

The above equations can also be written as

$\sum Y = na + b\sum X$

$\sum XY = a\sum X + b \sum X2$

Since $\sum x = 0$ i.e. deviation from actual mean is zero

We can write

$$a =? Y/n$$

$$b =? XY /? X^2$$

Fitting of parabolic trend

The mathematical form of a parabolic trend is given by:

$Yt = a + bt + ct2$

Here a, b and c are constants to be determined from the given data.

Using the method of least squares, the normal equations for the simultaneous solution of a, b and c are:

$$\sum Y = na + b\sum t + c\sum t^2$$

$$\sum tY = a\sum t + b\sum t^2 + c\sum t^3$$

$$\sum t^2Y = a\sum t^2 + b\sum t^3 + c\sum t^4$$

By selecting a suitable year of origin, i.e. define X = t - origin such that ?X = 0, the computation work can be considerably simplified. Also note that if ?X = 0, then ?X3 will also be equal to zero. Thus, the above equations can be rewritten as:

$\sum Y = na + c\sum X^2$ …….(1)

$\sum XY = b\sum X^2$ …….(2)

$\sum X^2Y = a\sum X^2 + c\sum X^4$…….(3)

From equation (2), we get

$$\mathbf{b = \sum XY / \sum X^2}$$

From equation (1), we get

$$\mathbf{a = \sum Y - c\sum X^2 / n}$$

And from equation (3), we get

$$\mathbf{c = n\sum X^2Y - (\sum X^2)(\sum Y) / n\sum X^4 - (\sum X^2)}$$

Or

$$\mathbf{c = \sum X^2Y - a\sum X^2 / \sum X^4}$$

This are the three equations to find the value of constants a, b and c.

Fitting of exponential trend

The general form of an exponential trend is:

$Y = a.bt$

Where 'a' and 'b' are constants to be determined from the observed data.

Taking logarithms of both side, we gave log Y = log a + log b.

This is linear equation in log Y and t can be fitted in a similar way as done in case of linear trend. Let A=log a and B =log b, then the above equation can be written as:

log Y = A + Bt

The normal equations based on the principle of least squares are:

$$\sum \log Y = n A + B \sum t$$

$$\text{And } \sum t \log Y = n \sum t + B \sum t^2$$

By selecting a suitable origin, i.e. defining X = t - origin such that ?X = 0, the computation work can be simplified. The values of A and B are given by:

$$\mathbf{A = \sum \log Y / n}$$

And

$$\mathbf{B = \sum X \log Y / \sum X^2}$$

Thus, the fitted trend equation can be written as:

log Y = A + BX or

Y = Antilog [A +BX]

= Antilog [log a + X log b]

= Antilog [log a.bx]

= a.bx

Merits:

1 Given the mathematical form of the trend to be fitted, the least squares method is an objective method.

2 Unlike the moving average method, it is possible to compute trend values for all the periods and predict the value for a period lying outside the observed data.

3 The results of the method of least squares are most satisfactory because the fitted trend satisfies the two most important properties, i.e. (1) ?(Y0 - Yt) =

0 and (2) ?(Y0 - Yt)2 is minimum. Here Y0 denotes the observed values and Yt denotes the calculated trend value.

The first property implies that the position of fitted trend equation is such that the sum of deviations of observations above and below this equal to zero. The second property implies that the sums of squares of deviations of observations, about the trend equations, are minimum.

Demerits:

1 As compared with the moving average method, it is cumbersome method.

2 It is not flexible like the moving average method. If some observations are added, then the entire calculations are to be done once again.

3 It can predict or estimate values only in the immediate future or the past.

4 The computation of trend values, on the basis of this method, doesn't take into account the other components of a time series and hence not reliable.

5 Since the choice of a particular trend is arbitrary, the method is not, strictly, objective.

6 This method cannot be used to fit growth curves, the pattern followed by the most of the economic and business time series.

2.5.3 Applications of time series in managerial decision making

The analysis of Time Series is of great significance not only to the economist and businessman but also to the scientist, geologist, biologist, research worker, etc., for the reasons given below:

(1) It helps in understanding past behaviours.

By observing data over a period of time one can easily understanding what changes have taken place in the past, Such analysis will be extremely helpful in producing future behaviour.

(2) It helps in planning future operations.

Plans for the future cannot be made without forecasting events and relationship they will have. Statistical techniques have been evolved which enable time series to be analyzed in such a way that the influences which have determined the form of that

series to be analyzed in such a way that the influences which have determined the form of that series may be ascertained. If the regularity of occurrence of any feature over a sufficient long period could be clearly established then, within limits, prediction of probable future variations would become possible.

(3) It helps in evaluating current accomplishments.

The performance can be compared with the expected performance and the cause of variation analyzed. For example, if expected sale for 1995 was 10,000 refrigerators and the actual sale was only 9,000, one can investigate the cause for the shortfall in achievement. Time series analysis will enable us to apply the scientific procedure of "holding other things constant" as we examine one variable at a time. For example, if we know how much the effect of seasonality on business is we may devise ways and means of ironing out the seasonal influence or decreasing it by producing commodities with complementary seasons.

(4) It facilitates comparison.

Different time series are often compared and important conclusions drawn there from.

However, one should not be led to believe that by time series analysis one can foretell with 100percnet accuracy the course of future events. After all, statisticians are not foretellers. This could be possible only if the influence of the various forces which affect these series such as climate, customs and traditions, growth and decline factors and the complex forces which proclimate, customs and traditions, growth and decline factors and the complex forces which produce business cycles would have been regular in their operation. However, the facts of life reveal that this type of regularity does not exist. But this then does not mean that time series analysis is of value. When such analysis is couples with a careful examination of current business indicators once can undoubtedly improve substantially upon guest mates i.e., estimates based upon pure guesswork) in forecasting future business conditions.

2.6 Summary

The skewness statistics, like all the descriptive statistics, are designed to help us think about the distributions of scores that our tests create. Skewness measures the symmetry of the distribution - most importantly it compares relative frequency of

extreme low (left tail) and extreme high (right tail) values. Normal distribution, which is perfectly symmetric, has skewness of zero. Positive skewness means that extremely high values are relatively more common (right tail is fat), while negative skewness means that extremely low values are more common (left tail is fat). In finance and investing (and even more so in options pricing and trading), knowing skewness of return distributions is very useful, as it may indicate frequency or probability of huge gains and (more importantly) huge losses.

The calculation of skewness may look complicated at first, but as soon as you get the underlying logic, it is quite straightforward. It is not unlike calculating variance and standard deviation. Skewness too has slightly different formula for population and sample. Here you can see a detailed explanation and derivation of skewness formula. The coefficient of skewness measures the skewness of a distribution. It is based on the notion of the moment of the distribution. This coefficient is one of the measures of skewness. Between the end of the nineteenth century and the beginning of the twentieth century, Pearson, Karl studied large sets of data which sometimes deviated significantly from normality and exhibited considerable skewness.

Time series is a compilation of observations of well-defined data items acquired by repeated measurements over time. Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time Series Analysis is used for many applications such as, economic forecasting, sales forecasting, budgetary analysis, stock market analysis, yield projections, process and quality control, inventory studies, workload projections, utility studies, census analysis. Time series can also be used in statistics, signal processing, pattern recognition, econometrics, mathematical finance, weather forecasting, intelligent transport, and trajectory forecasting, earthquake prediction, astronomy, communications engineering, control engineering, and largely in any domain of applied science and engineering that involves temporal measurements. There are various components of time series analysis. The fluctuations of time series can be classified into four basic types of variations. They are often called components or elements of a time series, these are: secular trend or long term movements, seasonal variations, cyclical variations, irregular variations. This chapter also discussed the measurement of trend: moving average and the method of least squares and this unit finally disused

the utility of utility of time series in managerial decision making.

2.7 Glossary

Skewness: In probability theory and statistics, skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative, or even undefined. The qualitative interpretation of the skew is complicated.

Pearson's first coefficient of skewness: also called Pearson mode skewness is a way to figure out the skewness of a distribution. The mean, mode and median can be used to figure out if you have a positively or negatively skewed distribution.

Standard deviation (SD): also represented by the Greek letter sigma **Σ** or the Latin letter s) is a measure that is used to quantify the amount of variation or dispersion of a set of data values

Moment: In mathematics, a moment is a specific quantitative measure, used in both mechanics and statistics, of the shape of a set of points. If the points represent mass, then the zeroth moment is the total mass, the first moment divided by the total mass is the centre of mass, and the second moment is the rotational inertia.

Time series: A time series is a sequence of numerical data points in successive order.

Seasonal component: The seasonal component is that part of the variations in a time series representing intra-year fluctuations that are more or less stable year after year with respect to timing, direction and magnitude. The seasonal component is also referred to as the seasonality of a time series.

Trend: A pattern of gradual change in a condition, output, or process, or an average or general tendency of a series of data points to move in a certain direction over time, represented by a line or curve on a graph.

Cyclical component: The cyclical component of a time series refers to (regular or periodic) fluctuations around the trend, excluding the irregular component, revealing a succession of phases of expansion and contraction.

Irregular component: Irregular component is the residual variation remaining after

the trend-cycle and seasonality have been extracted from original time series.

2.8 Self Assessment Questions

Q1: What are skewness and its significance?

Q2: Describe Karl Pearson's coefficient of skewness in statistics.

Q3: Explain Bowley's method and Kelly's method.

Q4: What is meant by moments of mean and moments about origin and their relationship?

Q5. Discuss the concepts of time series analysis and its importance in business decision making.

Q6. Explain the components of time series with examples?

Q7. Explain the measurement of trend and its methods?

Q8. Discuss the method of least squares?

2.9 Lesson End Exercises

- What is skewness of a distribution?
- **Calculation of Karl-Pearson's coefficient of skewness by using the following formula:**

  Coefficient of skewness = ?

  For the given data X = 12, 18, 18, 22, 35

  Mean = 21, Median = 18, ? = 7.7

- Calculate the Karl-Pearson's coefficient of skewness from the following data :

Marks (above) : 0 10 20 30 40 50 60 70 80

No. of Students: 150 140 100 80 80 70 30 14 0

Explain the utility of time series in managerial decision making

Comprehend the seasonal components?

List some examples of simple average method?

List some examples of cyclic component?

Discuss the irregular component with few examples in business decision making?

2.10 Suggested Readings

Beri, G.C. Business Statistics, IIIrd Ed. Tata McGraw Hill Pvt. Ltd.; India.

Cooper, Donald R. & Schindler, Pamela S. Business Research Methods, Tata McGraw Hill Compnies; India.

Jhunjhunwala, B. Business Statistics, S Chand & Co. New Delhi.

Sachdeva, J.K. Business Research Methodology, Himalaya Publishing House; New Delhi.

Shajahan, S. Research Methods for Management, Jaico Publishing House, Delhi; India.

Singh, D. & Chaudhary F.S. Theory and Analysis of Sample Survey Designs, New Age International (P) Limited: New Delhi.

2.11 References

Beri, G.C. Business Statistics, IIIrd Ed. Tata McGraw Hill Pvt. Ltd.; India.

Bloomfield, P. Fourier analysis of time series: An introduction. New York: Wiley.

Cooper, Donald R. & Schindler, Pamela S. Business Research Methods, Tata McGraw Hill Compnies; India.

Imdadullah. "Time Series Analysis" Basic Statistics and Data Analysis. itfeature.com.

Jhunjhunwala, B. Business Statistics, S Chand & Co. New Delhi.

Johnson, NL, Kotz, S, Balakrishnan N (1994) Continuous Univariate Distributions,

Vol 1, 2nd Edition Wiley ISBN 0-471-58495-9

Lawson, Charles L.; Hanson, Richard J. Solving Least Squares Problems. Philadelphia: Society for Industrial and Applied Mathematics.

MacGillivray, HL (1992). "Shape properties of the g- and h- and Johnson

families". Comm. Statistics - Theory and Methods.

Sachdeva, J.K. Business Research Methodology, Himalaya Publishing House; New Delhi.

Shajahan, S. Research Methods for Management, Jaico Publishing House, Delhi; India.

Shumway, R. H. Applied statistical time series analysis. Englewood Cliffs, NJ: Prentice Hall.

Singh, D. & Chaudhary F.S. Theory and Analysis of Sample Survey Designs, New Age International (P) Limited: New Delhi.

Structure

3.10 Suggested Readings

3.11 References

3.1 Introduction

In many business research situations, the key to decision making lies in understanding the relationships between two or more variables. For example, in an effort to predict the behaviour of the bond market, a broker might find it useful to know whether the interest rate of bonds is related to the prime interest rate. While studying the effect of advertising on sales, an account executive may find it useful to know whether there is a strong relationship between advertising dollars and sales dollars for a company.

In Social Study as well as Psychology there are times where it is needed to know whether there exists any relationship between the different abilities of the individual or they are independent of each other. There are many types of correlation like simple, curvilinear, and partial or multiple correlations that are computed in statistics. As we, in this text, aim to have an knowledge of the statistical methods like simple correlation, regression, Spearman's coefficient of rank correlation, Karl Pearson's coefficient of simple correlation.

Correlation is a statistical technique which tells us if two variables are related. For example, consider the variables family income and family expenditure. It is well known that income and expenditure increase or decrease together. Thus they are related in the sense that change in any one variable is accompanied by change in the other variable. Again price and demand of a commodity are related variables; when price increases demand will tend to decreases and vice versa. If the change in one variable is accompanied by a change in the other, then the variables are said to be correlated. We can therefore say that family income and family expenditure, price and demand are correlated. Correlation can tell us something about the relationship between variables. It is used to understand a. whether the relationship is positive or negative b. the strength of relationship. Correlation is a powerful tool that provides these vital pieces of information. In the case of family income and family expenditure, it is easy to see that they both rise or fall together in the same direction. This is called positive correlation. In

case of price and demand, change occurs in the opposite direction so that increase in one is accompanied by decrease in the other. This is called negative correlation.

In business, several times it becomes necessary to have some forecast so that the management can take a decision regarding a product or a particular course of action. In order to make a forecast, one has to ascertain some relationship between two or more variables relevant to a particular situation. For example, a company is interested to know how far the demand for television sets will increase in the next five years, keeping in mind the growth of population in a certain town. Here, it clearly assumes that the increase in population will lead to an increased demand for television sets. Thus, to determine the nature and extent of relationship between these two variables becomes important for the company. In general sense, regression analysis means the estimation or prediction of the unknown value of one variable from the known value(s) of the other variable(s). It is one of the most important and widely used statistical techniques in almost all sciences - natural, social or physical.

The two most popular correlation coefficients are: Spearman's correlation coefficient rho and Pearson's product-moment correlation coefficient. When calculating a correlation coefficient for ordinal data, select Spearman's technique. For interval or ratio-type data, use Pearson's technique. The value of a correlation coefficient can vary from minus one to plus one. A minus one indicates a perfect negative correlation, while a plus one indicates a perfect positive correlation. A correlation of zero means there is no relationship between the two variables. When there is a negative correlation between two variables, as the value of one variable increases, the value of the other variable decreases, and vise versa. In other words, for a negative correlation, the variables work opposite each other. When there is a positive correlation between two variables, as the value of one variable increases, the value of the other variable also increases. The variables move together.

3.2 Objectives

After reading this chapter, the students should be able,

1. Understand the importance and significance of correlation

2. Understand the various types of correlation
3. Understand the Karl Pearson's coefficient of simple correlation
4. Understand the importance of Spearman's coefficient of rank correlation
5. Understand the importance of regression
6. Comprehend the types of regression
6. Differentiate between correlation and regression analysis
7. Understand Scatter diagrams and regression coefficients

3.3 Meaning and Significance of Correlation

Correlation is a measure of association between two or more variables. When two or more variables very in sympathy so that movement in one tends to be accompanied by corresponding movements in the other variable(s), they are said to be correlated.

"The correlation between variables is a measure of the nature and degree of association between the variables".

Simple correlation or Bivariate correlation or Pearson correlation coefficient is denoted by 'r', summarizing the strength of association between two metric variables say X and Y. It is an index used to determine whether a linear or straight-line relationship exists between x and y. Correlation indicates the degree to which the variation of one variable X is related to variation in another variable Y. Correlation means the average relationship between two or more variables. When changes in the values of a variable affect the values of another variable, we say that there is a correlation between the two variables. The two variables may move in the same direction or in opposite directions. Simply because of the presence of correlation between two variables only, we cannot jump to the conclusion that there is a cause-effect relationship between them. Sometimes, it may be due to chance also.

As a measure of the degree of relatedness of two variables, correlation is widely used in exploratory research when the objective is to locate

variables that might be related in some way to the variable of interest.

Significance or Importance of Correlation:

i). Most of the variables show some kind of relationship. For instance, there is relationship between price and supply, income and expenditure etc. With the help of correlation analysis we can measure in one figure the degree of relationship.

ii). Once we know that two variables are closely related, we can estimate the value of one variable given the value of another. This is known with the help of regression.

iii). Correlation analysis contributes to the understanding of economic behaviour, aids in locating the critically important variables on which others depend.

iv). Progressive development in the methods of science and philosophy has been characterized by increase in the knowledge of relationship. In nature also one finds multiplicity of interrelated forces.

v). The effect of correlation is to reduce the range of uncertainty. The prediction based on correlation analysis is likely to be more variable and near to reality.

Advantages of Correlation

i). Show the amount (strength) of relationship present.

ii). Can be used to make predictions about the variables under study.

iii). Can be used in many places, including natural settings, libraries, etc.

iv). Easier to collect co-relational data

3.3.1 Types of Correlation

Positive correlation

If two variables x and y move in the same direction, we say that there is a positive correlation between them. In this case, when the value of one variable increases, the value of the other variable also increases and when the value of one variable

decreases, the value of the other variable also decreases. Eg. The age and height of a child.

Negative correlation

If two variables x and y move in opposite directions, we say that there is a negative correlation between them. i.e., when the value of one variable increases, the value of the other variable decreases and vice versa. Eg. The price and demand of a normal good.

Perfect Positive Correlation

If changes in two variables are in the same direction and the changes are in equal proportion, we say that there is a perfect positive correlation between them.

Perfect Negative Correlation

If changes in two variables are in opposite directions and the absolute values of changes are in equal proportion, we say that there is a perfect negative correlation between them.

Zero Correlation

If there is no relationship between the two variables, then the variables are said to be independent. In this case the correlation between the two variables is zero.

Pearson's product-moment coefficient is the measurement of correlation and ranges (depending on the correlation) between +1 and -1. +1 indicates the strongest positive correlation possible, and -1 indicates the strongest negative correlation possible. Therefore the closer the coefficient to either of these numbers the stronger the correlation of the data it represents. On this scale 0 indicates no correlation, hence values closer to zero highlight weaker/poorer correlation than those closer to +1/-1.

Linear and Non-linear (Curvilinear) Correlation

If the change in one variable is accompanied by change in another variable in a constant ratio,

it is a case of linear correlation. Observe the following data:

X : 10 20 30 40 50

Y : 25 50 75 100 125

The ratio of change in the above example is the same. It is, thus, a case of linear correlation. If we plot these variables on graph paper, all the points will fall on the same straight line. On the other hand, if the amount of change in one variable does not follow a constant ratio with the change in another variable, it is a case of non-linear or curvilinear correlation. If a couple of figures in either series X or series Y are changed, it would give a non-linear correlation.

Simple, Partial and Multiple Correlations

Simple Correlation:

In simple correlation, we study the relationship between two variables. Of these two variables one is principal and the other is secondary? For instance, income and expenditure, price and demand etc. Here income and price are principal variables while expenditure and demand are secondary variables.

Partial Correlation:

If in a given problem, more than two variables are involved and of these variables we study the relationship between only two variables keeping the other variables constant, correlation is said to be partial. It is so because the effect of other variables is assumed" to be constant

Multiple Correlations:

Under multiple correlations, the relationship between two and more variables is studied jointly. For instance, relationship between rainfall, use of fertilizer, manure on per hectare productivity of maize crop.

Example of Simple Correlation

A company wanted to know if there is a significant relationship between the total number of sales people and the total number of sales. They collect data for five months.

Variable 1	Variable 2
207	6907
180	5991
220	6810
205	6553
190	6190

Standard error of the coefficient = .068

T-test for the significance of the coefficient = 4.100

Degrees of freedom = 3

Two-tailed probability = .0263

Correlation coefficient = .921

Therefore we can say there is a significant relationship between the total number of sales people and the total number of sales as the correlation coefficient is equal to .921.

Example 2

Respondents to a survey were asked to judge the quality of a product on a four-point Likert scale (excellent, good, fair, poor). They were also asked to judge the reputation of the company that made the product on a three-point scale (good, fair, poor). Is there a significant relationship between respondent's perceptions of the company and their perceptions of quality of the product?

Since both variables are ordinal, Spearman's method is chosen. The first variable is the rating for the quality the product. Responses are coded as 4=excellent, 3=good, 2=fair, and 1=poor. The second variable is the perceived reputation of the company and is coded 3=good, 2=fair, and 1= poor.

Variable 1	Variable 2
4	3
2	2
1	2
3	3
4	3
1	1
2	1

Correlation coefficient rho = .830

T-test for the significance of the coefficient = 3.332

Number of data pairs = 7

Hence, there is a significant relationship between respondent's perceptions of the company and their perceptions of quality of the product as correlation coefficient rho = .830 and, t-test for the significance of the coefficient = 3.332.

3.3.2 Scatter Diagram

This method is also known as Dotogram or Dot diagram. Scatter diagram is one of the simplest methods of diagrammatic representation of a bivariate distribution. Under this method, both the variables are plotted on the graph paper by putting dots. The diagram so obtained is called "Scatter Diagram". By studying diagram, we can have rough idea about the nature and degree of relationship between two variables. The term scatter refers to the spreading of dots on the graph. A scatter diagram is a graphic representation of the relationship between two variables. Scatter diagrams help teams identify and understand cause-effect relationships.

When to Use a Scatter Diagram

- When you have paired numerical data.
- When your dependent variable may have multiple values for each value of

your independent variable.

- When trying to determine whether the two variables are related, such as…
- When trying to identify potential root causes of problems.

After brainstorming causes and effects using a fishbone diagram, to determine objectively whether a particular cause and effect are related.

- When determining whether two effects that appear to be related both occur with the same cause.
- When testing for autocorrelation before constructing a control chart.

Scatter Diagram Procedure

1. Collect pairs of data where a relationship is suspected.

2 Draw a graph with the independent variable on the horizontal axis and the dependent variable on the vertical axis. For each pair of data, put a dot or a symbol where the x-axis value intersects the y-axis value. (If two dots fall together, put them side by side, touching, so that you can see both.)

3 Look at the pattern of points to see if a relationship is obvious. If the data clearly form a line or a curve, you may stop. The variables are correlated. You may wish to use regression or correlation analysis now. Otherwise, complete steps 4 through 7.

4 Divide points on the graph into four quadrants. If there are X points on the graph,

Count X/2 points from top to bottom and draw a horizontal line.

Count X/2 points from left to right and draw a vertical line.

If number of points is odd, draw the line through the middle point.

5 Count the points in each quadrant. Do not count points on a line.

6 Add the diagonally opposite quadrants. Find the smaller sum and the total of points in all quadrants.

A = points in upper left + points in lower right

B = points in upper right + points in lower left

Q = the smaller of A and B

N = A + B

7 Look up the limit for N on the trend test table.

If Q is less than the limit, the two variables are related.

If Q is greater than or equal to the limit, the pattern could have occurred from random chance.

Table 5.18 Trend test table.

N	Limit	*N*	Limit
1–8	0	51–53	18
9–11	1	54–55	19
12–14	2	56–57	20
15–16	3	58–60	21
17–19	4	61–62	22
20–22	5	63–64	23
23–24	6	65–66	24
25–27	7	67–69	25
28–29	8	70–71	26
30–32	9	72–73	27
33–34	10	74–76	28
35–36	11	77–78	29
37–39	12	79–80	30
40–41	13	81–82	31
42–43	14	83–85	32
44–46	15	86–87	33
47–48	16	88–89	34
49–50	17	90	35

Scatter Diagram Example

The ZZ-400 manufacturing team suspects a relationship between product purity (percent purity) and the amount of iron (measured in parts per million or ppm). Purity and iron are plotted against each other as a scatter diagram, as shown in the figure below.

There are 24 data points. Median lines are drawn so that 12 points fall on each

side for both percent purity and ppm iron.

To test for a relationship, they calculate:

A = points in upper left + points in lower right = 9 + 9 = 18

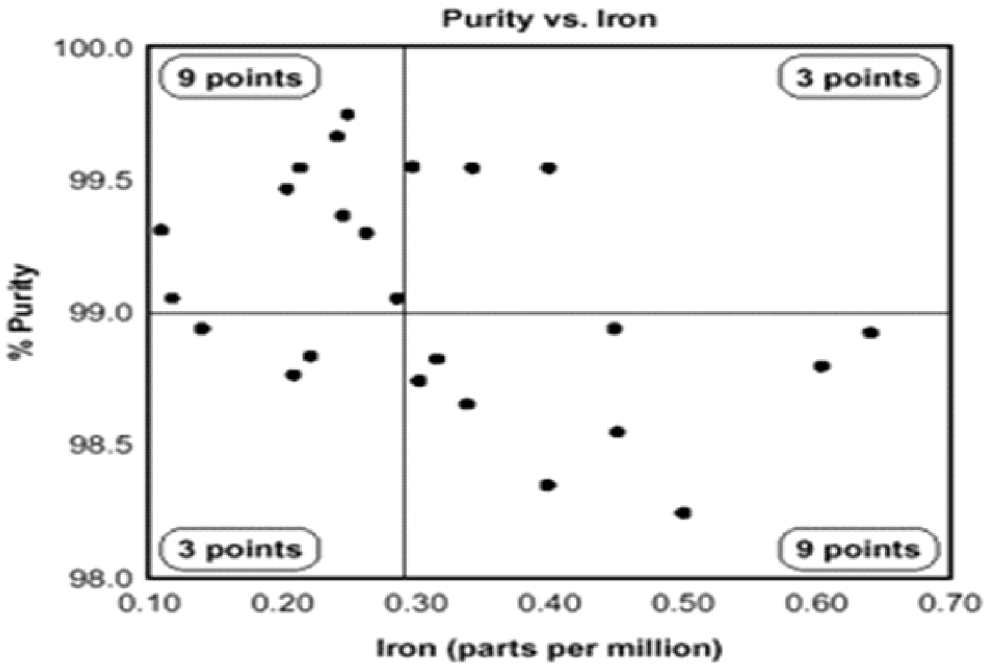
B = points in upper right + points in lower left = 3 + 3 = 6

Q = the smaller of A and B = the smaller of 18 and 6 = 6

N = A + B = 18 + 6 = 24

Then they look up the limit for N on the trend test table. For N = 24, the limit is 6.

Q is equal to the limit. Therefore, the pattern could have occurred from random chance, and no relationship is demonstrated.



Scatter Diagram Example

Scatter Diagram Considerations

Here are some examples of situations in which might you use a scatter

diagram:

- Variable A is the temperature of a reaction after 15 minutes. Variable B measures the color of the product. You suspect higher temperature makes the product darker. Plot temperature and color on a scatter diagram.
- Variable A is the number of employees trained on new software, and variable B is the number of calls to the computer help line. You suspect that more training reduces the number of calls. Plot number of people trained versus number of calls.
- To test for autocorrelation of a measurement being monitored on a control chart, plot this pair of variables: Variable A is the measurement at a given time. Variable B is the same measurement, but at the previous time. If the scatter diagram shows correlation, do another diagram where variable B is the measurement two times previously. Keep increasing the separation between the two times until the scatter diagram shows no correlation.

Even if the scatter diagram shows a relationship, do not assume that one variable caused the other. Both may be influenced by a third variable.

When the data are plotted, the more the diagram resembles a straight line, the stronger the relationship.

If a line is not clear, statistics (N and Q) determine whether there is reasonable certainty that a relationship exists. If the statistics say that no relationship exists, the pattern could have occurred by random chance.

If the scatter diagram shows no relationship between the variables, consider whether the data might be stratified.

If the diagram shows no relationship, consider whether the independent (x-axis) variable has been varied widely. Sometimes a relationship is not apparent because the data don't cover a wide enough range.

Think creatively about how to use scatter diagrams to discover a root cause.

Drawing a scatter diagram is the first step in looking for a relationship between variables.

3.3. 3 Karl Pearson's Coefficient of Simple Correlation

Coefficient of Correlation

Correlation is measured by what is called coefficient of correlation (r). A correlation coefficient is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another. Correlation coefficients are expressed as values between +1 and -1. Its numerical value gives us an indication of the strength of relationship. In general, r > 0 indicates positive relationship, r < 0 indicates negative relationship while r = 0 indicates no relationship (or that the variables are independent and not related). Here r = +1.0 describes a perfect positive correlation and r = ?1.0 describes a perfect negative correlation. Closer the coefficients are to +1.0 and ?1.0, greater is the strength of the relationship between the variables. As a rule of thumb, the following guidelines on strength of relationship are often useful (though many experts would somewhat disagree on the choice of boundaries).

The coefficient of correlation between two variables X, Y is a measure of the degree of association (i.e., strength of relationship) between them. The coefficient of correlation is usually denoted by 'r'.

Value of r	**Strength of relationship**
?1.0 to ?0.5 or 1.0 to 0.5	Strong
?0.5 to ?0.3 or 0.3 to 0.5	Moderate
?0.3 to ?0.1 or 0.1 to 0.3	Weak
?0.1 to 0.1	none or very weak
1	A perfect positive correlation
0	No Correlation (No relation between two variables)
1	A perfect negative correlation

Karl Pearson's Coefficient of Simple Correlation: Karl Pearson's Product-Moment Correlation Coefficient or simply Pearson's Correlation Coefficient for short, is one of the important methods used in Statistics to measure Correlation between two variables. A mathematical method for measuring the intensity or the magnitude of linear relationship between two variables was suggested by Karl Pearson (1867-1936). Karl Pearson was a British mathematician, statistician, lawyer and a eugenicist. He established the discipline of mathematical statistics. He founded the world's first statistics department In the University of London in the year 1911. He along with his colleagues Weldon and Galton founded the journal 'Biometrika' whose object was the development of statistical theory. The Pearson product-moment correlation coefficient (r) is a common measure of the correlation between two variables X and Y. When measured in a population the Pearson Product Moment correlation is designated by the Greek letter rho. When computed in a sample, it is designated by the letter "r" and is sometimes called "Pearson's r." Pearson's correlation reflects the degree of linear relationship between two variables. Mathematical Formula:-

The quantity r, called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables. (The linear correlation coefficient is a measure of the strength of linear relation between two quantitative variables. We use the Greek letter ? (rho) to represent the population correlation coefficient and r to represent the sample correlation coefficient.)

Let N denote the number of pairs of observations of two variables X and Y. The correlation coefficient r between X and Y is defined by

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

This formula is suitable for solving problems with hand calculators. To apply this formula, we have to calculate ? X, ?Y, ?XY, ?X2, ?Y2.

Properties of Correlation Coefficient

Let r denote the correlation coefficient between two variables. r? is interpreted using the following properties:

1. The value of r ranges from - 1.0 to 0.0 or from 0.0 to 1.0
2. A value of r = 1.0 indicates that there exists perfect positive correlation between the two variables.
3. A value of r = - 1.0 indicates that there exists perfect negative correlation between the two variables.
4. A value r = 0.0 indicates zero correlation i.e., it shows that there is no correlation at all between the two variables.
5. A positive value of r shows a positive correlation between the two variables.
6. A negative value of r shows a negative correlation between the two variables.
7. A value of r = 0.9 and above indicates a very high degree of positive correlation between the two variables.
8. A value of - 0.9 ? r > - 1.0 shows a very high degree of negative correlation between the two variables.
9. For a reasonably high degree of positive correlation, we require r to be from 0.75 to 1.0.
10. A value of r from 0.6 to 0.75 may be taken as a moderate degree of positive correlation.

Assumptions of Pearson's Correlation Coefficient

1. There is linear relationship between two variables, i.e. when the two variables are plotted on a scatter diagram a straight line will be formed by the points.
2. Cause and effect relation exists between different forces operating on the item of the two variable series.

Advantages of Pearson's Coefficient

1. It summarizes in one value, the degree of correlation & direction of

correlation also.

Disadvantages

While 'r' (correlation coefficient) is a powerful tool, it has to be handled with care.

1. The most used correlation coefficients only measure linear relationship. It is therefore perfectly possible that while there is strong non-linear relationship between the variables, r is close to 0 or even 0. In such a case, a scatter diagram can roughly indicate the existence or otherwise of a non-linear relationship.
2. One has to be careful in interpreting the value of 'r'. For example, one could compute 'r' between the size of shoe and intelligence of individuals, heights and income. Irrespective of the value of 'r', it makes no sense and is hence termed chance or non-sense correlation.
3. 'r' should not be used to say anything about cause and effect relationship. Put differently, by examining the value of 'r', we could conclude that variables X and Y are related.

However the same value of 'r' does not tell us if X influences Y or the other way round. Statistical correlation should not be the primary tool used to study causation, because of the problem with third variables.

Problem 1

The following are data on Advertising Expenditure (in Rupees Thousand) and Sales (Rupees in lakhs) in a company.

Advertising Expenditure	:	18	19	20	21	22	23
Sales	:	17	17	18	19	19	19

Determine the correlation coefficient between them and interpret the result.

Solution:

We have N = 6. Calculate ? X, ?Y, ?XY, ?Y2, ?Y2 as follows:

X	Y	XY	X2	Y2
18	17	306	324	289
19	17	323	361	289
20	18	360	400	324
21	19	399	441	361
22	19	418	484	361
23	19	437	529	361
Total:	123	109	2243	2539	1985

The correlation coefficient r between the two variables is calculated as follows:

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

$$r = \frac{6 \times 2243 - 123 \times 109}{\sqrt{6 \times 2539 - (123)^2}\sqrt{6 \times 1985 - (109)^2}}$$

= (13458 – 13407) / {√(15234- 15129) √(11910- 11881)}

=51/{√105 √29} = 51/ (10.247 x 5.365)

= 51/ 54.975

= 0.9277

Interpretation

The value of r is 0.92. It shows that there is a high, positive correlation between the two variables 'Advertising Expenditure' and 'Sales'. This provides a basis to consider some functional relationship between them.

Problem 2

Consider the following data on supply and price. Determine the correlation

Coefficient between the two variables and interpret the result.

Supply	:	11	13	17	18	22	24	26	28
Price	:	25	32	26	25	20	17	11	10

Determine the correlation coefficient between the two variables and interpret the result.

Solution:

We have N = 8. Take X = Supply and Y = Price.

Calculate ? X, ?Y, ?XY, ?X2, ?Y2 as follows:

X	Y	XY	X2	Y2
11	25	275	121	625
13	32	416	169	1024
17	26	442	289	676
18	25	450	324	625
22	20	440	484	400
24	17	408	576	289
26	11	286	676	121
28	10	280	784	100
Total: 159	166	2997	3423	3860

The correlation coefficient between the two variables is r =

{8 x 2997 - (159 x 166)} / { ?(8 x 3423 - 1592) ?(8 x 3860 - 1662) }

= (23976 - 26394) / {?(27384- 25281) ?(30880- 27566)}

= - 2418 / {?2103 ?3314}

= - 2418 / (45.86 x 57.57)

= - 2418 / 2640.16

= - 0.9159

Interpretation

The value of r is - 0.92. The negative sign in r shows that the two variables move in opposite directions. The absolute value of r is 0.92 which is very high. Therefore we conclude that there is high negative correlation between the two variables 'Supply' and 'Price'.

3.3.4 Spearman Rank Correlation

Sometimes we come across statistical series in which the variables under consideration are not capable of quantitative measurement but can be arranged in serial order. This happens when we are dealing with qualitative characteristics (attributes) such as honesty, beauty, character, morality, etc., which cannot be measured quantitatively but can be arranged serially. In such situations Karl Pearson's coefficient of correlation cannot be used as such. Charles Edward Spearman, a British Psychologist, developed a formula in 1904, which consists in obtaining the correlation coefficient between the ranks of N individuals in the two attributes under study.

Suppose we want to find if two characteristics A, say, intelligence and B, say, beauty are related or not. Both the characteristics are incapable of quantitative measurements but we can arrange a group of N individuals in order of merit (ranks) w.r.t. proficiency in the two characteristics. Let the random variables X and Y denote the ranks of the individuals in the characteristics A and B respectively. If we assume that there is no tie, i.e., if no two individuals get the same rank in a characteristic then, obviously, X and Y assume numerical values ranging from 1 to N

If ranks can be allocated to pairs of observations for two variables X and Y, then the correlation between the ranks is called the rank correlation coefficient. It is usually denoted by the symbol ? (rho).

Correlation coefficients

Some of the more popular rank correlation statistics include

1. Spearman's ?
2 .Kendall's ?
3. Goodman and Kruskal's ?
4. Somers' D

An increasing rank correlation coefficient implies increasing agreement between rankings. The coefficient is inside the interval (?1, 1) and assumes the value:

- if the agreement between the two rankings is perfect; the two rankings are the same.
- 0 if the rankings are completely independent.
- ? 1 if the disagreement between the two rankings is perfect; one ranking is the reverse of the other.

Following Diaconis (1988), a ranking can be seen as a permutation of a set of objects. Thus we can look at observed rankings as data obtained when the sample space is (identified with) a symmetric group. We can then introduce a metric, making the symmetric group into a metric space. Different metrics will correspond to different rank correlations.

Meaning

In statistics, a rank correlation is any of several statistics that measure an ordinal association - the relationship between rankings of different ordinal variables or different rankings of the same variable, where a "ranking" is the assignment of the labels "first", "second", "third", etc. to different observations of a particular variable. A rank correlation coefficient measures the degree of similarity between two rankings, and can be used to assess the significance of the relation between them. For example, two common nonparametric methods of significance that use rank correlation are the Mann-Whitney U test and the Wilcoxon signed-rank test.

Importance of Spearman's Rank Correlation

In statistics, Spearman's rank correlation coefficient or Spearman's rho, named after Charles Spearman and often denoted by the Greek letter (rho) or as , is a nonparametric measure of rank correlation (statistical dependence between the ranking of two variables). The Spearman's rank-order correlation is the nonparametric version of the Pearson product moment correlation. Spearman's correlation coefficient, measures the strength of association between two ranked variables. The Spearma's rank-order correlation is used when there is a monotonic relationship between our variables. A monotonic relationship is a relationship that does one of the following: (1) as the value of one variable increases, so does the value of the other variable; or (2) as the value of one variable increases, the other variable value decreases. A monotonic relationship is an important underlying assumption of the Spearman rank-order correlation. It is also important to recognize the assumption of a monotonic relationship is less restrictive than a linear relationship (an assumption that has to be met by the Pearson product-moment correlation). The middle image above illustrates this point well: A non-linear relationship exists, but the relationship is monotonic and is suitable for analysis by Spearman's correlation, but not by Pearson's correlation.

The rank correlation method seeks to assess the consumer's preferences for individual attributes.

Let us make the relevance of use of Spearman Rank Correlation Coefficient with the aid of an example.

If ranks can be assigned to pairs of observations for two variables X and Y, then the correlation between the ranks is called the rank correlation coefficient. It is usually denoted by the symbol ? (rho). It is given by the formula

$$\rho = 1 - \frac{6\sum D^2}{N^{3-N}}$$

where

D = difference between the corresponding ranks of X and Y

$$= RX - RY$$

and N is the total number of pairs of observations of X and Y.

Advantages/Importance of Spearman's Rank Correlation

1. This method is simpler to understand and easier to apply compared to Karl Pearson's correlation method.
2. This method is useful where we can give the ranks and not the actual data. (qualitative term)
3. This method is to use where the initial data in the form of ranks.
4. It is a powerful tool of determining co variability among attributes such as honesty, beauty etc.

Disadvantages Spearman's Rank Correlation

1. It cannot be used for finding out correlation in a grouped frequency distribution.
2. This method should be applied where N exceeds 30.
3. As Spearman's rank only uses rank, it is not affected by significant variations in readings. As long as the order remains the same, the coefficient will stay the same. As with any comparison, the possibility of chance will have to be evaluated to ensure that the two quantities are actually connected. 4. A significant correlation does not necessarily mean cause and effect.
4. It is not accurate as the Karl Pearson's coefficient of correlation.

Problem 3

Alpha Recruiting Agency short listed 10 candidates for final selection. They were examined in written and oral communication skills. They were ranked as follows:

Candidates serial no.	1	2	3	4	5	6	7	8	9	10
Rank in written communication	8	7	2	10	3	5	1	9	6	4
Rank in oral communication	10	7	2	6	5	4	1	9	8	3

Find out whether there is any correlation between the written and oral communication skills of the short listed candidates.

Solution:

Take X = Written Communication Skill and Y = Oral Communication Skill.

Rank of X: R1	Rank of Y: R2	D= R1-R2	D2
8	10	- 2	4
7	7	0	0
2	2	0	0
10	6	4	16
3	5	- 2	4
5	4	1	1
1	1	0	0
9	9	0	0
6	8	- 2	4
4	3	1	1
		Total:	30

We have N = 10. The rank correlation coefficient is

? = 1 - {6 ? D2 / (N3 - N)}

= 1 - {6 x 30 / (1000 - 10)}

= 1 - (180 / 990)

= 1 - 0.18

= 0.82

Inference:

From the value of r, it is inferred that there is a high, positive rank correlation

between the written and oral communication skills of the short listed candidates.

Problem 4: Resolving ties in ranks

The following are the details of ratings scored by two popular insurance schemes. Determine the rank correlation coefficient between them.

Scheme 1	80	80	83	84	87	87	89	90
Scheme 11	55	56	57	57	57	58	59	60

Solution:

From the given values, we have to determine the ranks.

Step 1.

Arrange the scores for Insurance Scheme I in descending order and rank them as 1, 2, 3,…,8.

Scheme 1 Score	90	89	87	87	84	83	80	80
Scheme 11 Rank	1	2	3	4	5	6	7	8

The score 87 appears twice. The corresponding ranks are 3, 4. Their average is (3 + 4) / 2 = 3.5. Assign this rank to the two equal scores in Scheme I.

The score 80 appears twice. The corresponding ranks are 7, 8. Their average is (7 + 8) / 2 = 7.5. Assign this rank to the two equal scores in Scheme I.

The revised ranks for Insurance Scheme I are as follows:

Scheme 1 Score	90	89	87	87	84	83	80	80
Scheme 11 Rank	1	2	3.5	3.5	5	6	7.5	7.5

Step 2.

Arrange the scores for Insurance Scheme II in descending order and rank them as 1,2,3,…,8.

Scheme 11	60	59	58	57	57	57	56	55
Rank	1	2	3	4	5	6	7	8

The score 57 appears thrice. The corresponding ranks are 4, 5, 6.

Their average is (4 + 5 + 6) / 3 = 15 / 3 = 5. Assign this rank to the three equal scores in Scheme II.

The revised ranks for Insurance Scheme II are as follows:

Scheme 11 Score	60	59	58	57	57	57	56	55
Rank	1	2	3	4	5	6	7	8

Step 3.

Calculation of D2: Assign the revised ranks to the given pairs of values and calculate D2 as follows:

Scheme 1 Score	Scheme 11 Score	Scheme 1 Rank: R1	Scheme 11 Rank:R2	D= R1 - R2	D2
80	55	7.5	8	- 0.5	0.25
80	56	7.5	7	0.5	0.25
83	57	6	5	1	1
84	57	5	5	0	0
87	57	3.5	5	-1.5	2.25
87	58	3.5	3	0.5	0.25
89	59	2	2	0	0
90	60	1	1	0	0
				Total:	4

Step 4.

Calculation of ?:

We have N = 8.

Since there are 2 ties with 2 items each and another tie with 3 items, the correction term is 0.5 + 0.5 + 2.

The rank correlation coefficient is

? = 1 - [{ 6 ? D2 + (1/2) + (1/2) +2 }/ (N3 - N)}]

= 1 - { 6 (4.+0.5+0.5+2) / (512 - 8) } = 1 - (6 x 7 / 504) = 1 - (42/504)

= 1 - 0.083 = 0.917

Inference:

It is inferred that the two insurance schemes are highly, positively correlated.

3.4 Meaning and Importance of Regression

Regression analysis is a statistical tool, which helps understand the relationship between variables and predicts the unknown values of the dependent variable from known values of the independent variable. Regression analysis is one of the most extensively utilized methods between the analytical models of association employed in business research.

Regression analysis tries to analyze the connection between a dependent variable and a group of independent variables (one or more). One example is, in demand analysis, demand is versely linked to price for normal commodities. We can write D = A - BP, where D is, the demand which is the dependent variable, P is the unit price of the commodity, an independent variable. It is an example of a simple linear regression equation. The multiple linear regressions model is the prototype of single criterion multiple predictor association model where we wish to research the combined impact of several independent variables upon one dependent variable. In the above example if P is the consumer price index, and Q is the index of industrial production, we might manage to research demand as a function of 2 independent variables P and Q and write D = A - BP + C Q as a multiple linear regression model.

Regression analysis is commonly employed for prediction and forecasting, where its use has considerable overlap with the field of machine learning. It is also utilized to understand which among the independent variables are related to the

dependent variable, and to take a look at the types of these relationships. In restricted situations, regression analysis enables you to infer causal relationships between the independent and dependent variables. However this can result in illusions or false relationships, so caution is advisable; for instance, correlation doesn't mean causation.

Let us assume that the number of books in circulation, in a library is related to the number of users. For example, it can be postulated that as the number of users' increases, the number of books in circulation also increases. Here, the number of users is the independent variable and the number of books in circulation is the dependent variable.

Let us denote the dependent variable as Y and the independent variable as X. In regression analysis, we gather data over a period of time or across units at a point of time. Let us assume that 'n' pairs of observations in X and Y is collected. The next step is to find out the relationship X and Y.

The relationship between X and Y can take many forms. The general practice is to express the relationship in terms of some mathematical equations. The simplest of these equations is the linear equation. This means that the relationship between X and Y is in the form of a straight line. When the equation represents curves (not a straight line) the regression is called non-linear or curvilinear.

Now the question arises, "how do we identify the equational form?" There is no hard and fast rule as such. The form of equation depends upon the reasoning and assumptions made by the researcher. However, the researcher may plot the X and Y variables on a graph paper to prepare a scatter diagram. From the scatter diagram, the location of the points on the graph paper helps in identifying the equational from. If the points are more or less in a straight line then linear equation is assumed. If the points are not in a straight line and are in the form of a curve, a suitable non-linear equation, which resembles the scatter, is assumed.

The researcher has to make another assumption: viz. identification of independent and dependent variables. The again depends upon the logic of the researcher and purpose of analysis: whether Y depends upon X or X depends upon Y. Thus, there may be two regression lines from the same data (a) when Y is assumed to

be dependent upon X, this is termed 'Y on X' line, and (b) when X is assumed to be dependent upon Y, this is termed 'X on Y' line.

Let us take an example of a linear equation with Y as the dependent variable and X as the independent variable. Y=3+2X By taking in different values of X, we can determine the values of Y, e.g., when X=1, Y=5; when X=2, Y=7 and so on. If we plot these pairs of points (1,5) (2,7), etc. on a graph paper we get a straight line.

Generalising the above relationship, it can be said that a linear equation of Y on X take3s the form Y=a+bX, where a and b are constants. Similarly, non-linear equations can be specified in many forms. A simple example is Y=a+bX+cX2

Meaning of regression

Regression is a statistical technique to determine the linear relationship between two or more variables. Regression is primarily used for prediction and causal inference. In its simplest (bivariate) form, regression shows the relationship between one independent variable (X) and a dependent variable (Y), as in the formula below:

Y = ß0 + ß1 X + u

The magnitude and direction of that relation are given by the slope parameter (ß1), and the status of the dependent variable when the independent variable is absent is given by the intercept parameter (ß0). An error term (u) captures the amount of variation not predicted by the slope and intercept terms. The regression coefficient (R2) shows how well the values fit the data. Regression thus shows us how variation in one variable co-occurs with variation in another. What regression cannot show is causation; causation is only demonstrated analytically, through substantive theory. For example, a regression with shoe size as an independent variable and foot size as a dependent variable would show a very high regression coefficient and highly significant parameter estimates, but we should not conclude that higher shoe size causes higher foot size. All that the mathematics can tell us is whether or not they are correlated, and if so, by how much. It is important to recognize that regression analysis is fundamentally different

from ascertaining the correlations among different variables. Correlation determines the strength of the relationship between variables, while regression attempts to describe that relationship between these variables in more detail.

Importance of regression

Regression analysis is one of the most important statistical techniques for business applications. It's a statistical methodology that helps estimate the strength and direction of the relationship between two or more variables. The analyst may use regression analysis to determine the actual relationship between these variables by looking at a corporation's sales and profits over the past several years. The regression results show whether this relationship is valid.

In addition to sales, other factors may also determine the corporation's profits, or it may turn out that sales don't explain profits at all. In particular, researchers, analysts, portfolio managers, and traders can use regression analysis to estimate historical relationships among different financial assets. They can then use this information to develop trading strategies and measure the risk contained in a portfolio. As mentioned above, regression analysis estimates the relationship between two or more variables. Let's understand this with an easy example:

Let's say, you want to estimate growth in sales of a company based on current economic conditions. You have the recent company data which indicates that the growth in sales is around two and a half times the growth in the economy. Using this insight, we can predict future sales of the company based on current & past information.

There are multiple benefits of using regression analysis. They are as follows:

1 It indicates the **significant relationships** between dependent variable and independent variable.

2 It indicates the **strength of impact** of multiple independent variables on a dependent variable.

Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of

promotional activities. These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models. Regression analysis is an indispensable tool for analyzing relationships between financial variables. For example, it can:

- Identify the factors that are most responsible for a corporation's profits
- Determine how much a change in interest rates will impact a portfolio of bonds
- Develop a forecast of the future value of the Dow Jones Industrial Average

4 .1 Types of Regression

There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics (number of independent variables, type of dependent variables and shape of regression line).

In this section, we are going to study in detail **Simple and multiple regression analysis, Linear and non-linear regression**

Simple regression analysis

Regression analysis is most often used for prediction. The goal in regression analysis is to create a mathematical model that can be used to predict the values of a dependent variable based upon the values of an independent variable. In other words, we use the model to predict the value of Y when we know the value of X. (The dependent variable is the one to be predicted). Correlation analysis is often used with regression analysis because correlation analysis is used to measure the strength of association between the two variables X and Y.

In regression analysis involving one independent variable and one dependent variable the values are frequently plotted in two dimensions as a scatter plot. The scatter plot allows us to visually inspect the data prior to running a regression analysis. Often this step allows us to see if the relationship between the two variables is increasing or decreasing and gives only a rough idea of the relationship. The simplest relationship between two variables is a straight-line or linear relationship. Of course the data may well be curvilinear and in that case

we would have to use a different model to describe the relationship. Simple linear regression analysis finds the straight line that best fits the data.

Regression Analysis is a very powerful and flexible tool in the field of statistical analysis in predicting the value of one variable, given the value of another variable, when those variables are related to each other. Regression goes beyond correlation by adding prediction capabilities. If two variables are significantly correlated, and if there is some theoretical basis for doing so, it is possible to predict values of one variable from the other. This observation leads to a very important concept known as 'Regression Analysis'. Regression analysis, in general sense, means the estimation or prediction of the unknown value of one variable from the known value of the other variable.

Regression analysis was explained by M. M. Blair as follows:

"Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data." Simple regression is used to examine the relationship between one dependent and one independent variable. The variable which is used to predict the variable of interest is called the 'independent variable' or 'explanatory variable' and the variable we are trying to predict is called as 'dependent variable' or 'explained variable'. Independent variable is denoted by x and dependent variable is denoted by y. The analysis is used is called simple linear regression. Simple because there is only one predictor or independent variable and linear because of the assumed relationship between the dependent and independent variables. In other words, regression analysis is mathematical measure of average relationship between two or more variables. After performing an analysis, the regression statistics can be used to predict the dependent variable when the independent variable is known.

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:

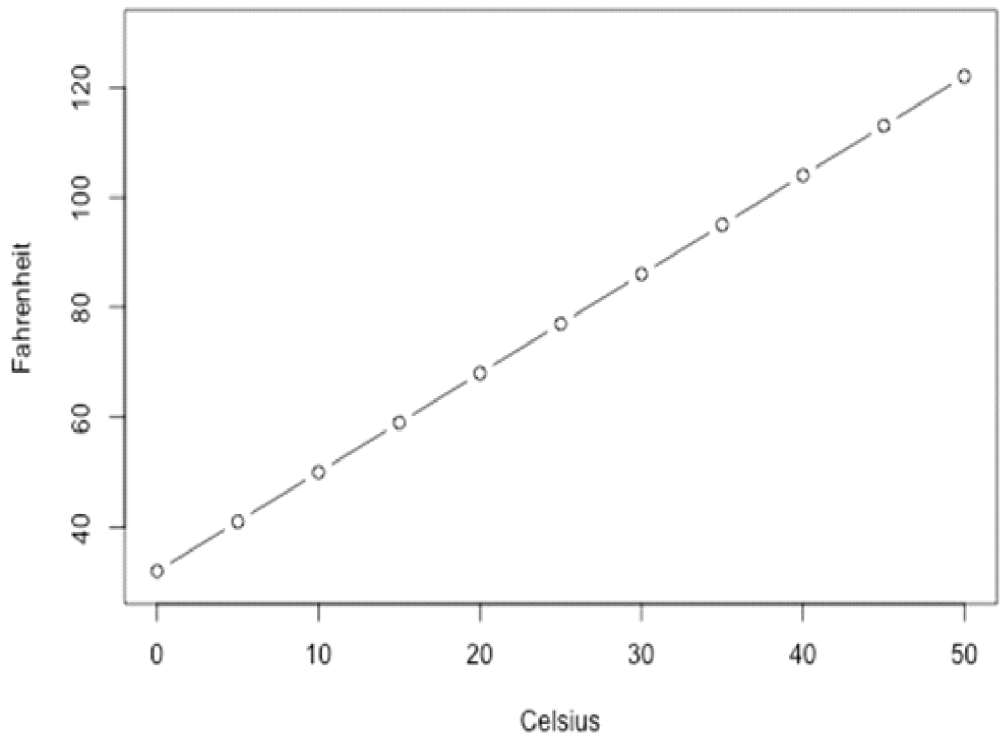
- One variable, denoted x, is regarded as the predictor, explanatory, or independent variable.
- The other variable, denoted y, is regarded as the response, outcome, or

dependent variable.

Because the other terms are used less frequently today, we'll use the "predictor" and "response" terms to refer to the variables encountered in this course. The other terms are mentioned only to make you aware of them should you encounter them in other arenas. Simple linear regression gets its adjective "simple," because it concerns the study of only one predictor variable. In contrast, multiple linear regression, which we study later in this course, gets its adjective "multiple," because it concerns the study of two or more predictor variables.

Types of relationships

Before proceeding, we must clarify what types of relationships we won't study in this course, namely, deterministic (or functional) relationships. Here is an example of a deterministic relationship. Here is an example of a deterministic relationship.



Note that the observed (x, y) data points fall directly on a line. As you may remember, the relationship between degrees Fahrenheit and degrees Celsius is known to be:

Fahr =95Cels+32Fahr =95Cels+32

That is, if you know the temperature in degrees Celsius, you can use this equation to determine the temperature in degrees Fahrenheit exactly.

Here are some examples of other deterministic relationships that students from previous semesters have shared:

Circumference = ? × diameter

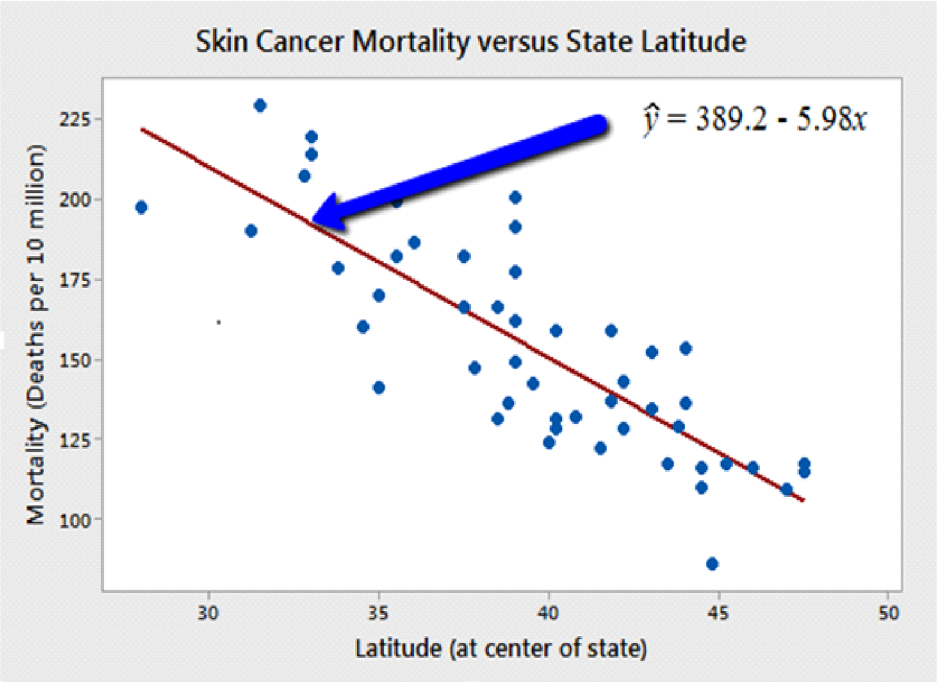
Hooke's Law: Y = ? + ?X, where Y = amount of stretch in a spring, and X = applied weight.

Ohm's Law: I = V/r, where V = voltage applied, r = resistance, and I = current.

Boyle's Law: For a constant temperature, P = ?/V, where P = pressure, ? = constant for each gas, and V = volume of gas.

For each of these deterministic relationships, the equation exactly describes the relationship between the two variables. This course does not examine deterministic relationships. Instead, we are interested in statistical relationships, in which the relationship between the variables is not perfect.

Here is an example of a statistical relationship. The response variable y is the mortality due to skin cancer (number of deaths per 10 million people) and the predictor variable x is the latitude (degrees North) at the center of each of 49 states in the U.S. (skincancer.txt) (The data were compiled in the 1950s, so Alaska and Hawaii were not yet states. And, Washington, D.C. is included in the data set even though it is not technically a state.)



You might anticipate that if you lived in the higher latitudes of the northern U.S., the less exposed you'd be to the harmful rays of the sun, and therefore, the less risk you'd have of death due to skin cancer. The scatter plot supports such a hypothesis. There appears to be a negative linear relationship between latitude and mortality due to skin cancer, but the relationship is not perfect. Indeed, the plot exhibits some "trend," but it also exhibits some "scatter." Therefore, it is a statistical relationship, not a deterministic one.

Some other examples of statistical relationships might include:

- Height and weight - as height increases, you'd expect weight to increase, but not perfectly.
- Alcohol consumed and blood alcohol content - as alcohol consumption increases, you'd expect one's blood alcohol content to increase, but not perfectly.

- Vital lung capacity and pack-years of smoking - as amount of smoking increases (as quantified by the number of pack-years of smoking), you'd expect lung function (as quantified by vital lung capacity) to decrease, but not perfectly.
- Driving speed and gas mileage - as driving speed increases, you'd expect gas mileage to decrease, but not perfectly.

Importance:

1. It is one of the most important statistical tools which is extensively used in almost all sciences - Natural, Social and Physical. It is specially used in business and economics to study the relationship between two or more variables that are related causally and for the estimation of demand and supply graphs, cost functions, production and consumption functions and so on.
2. Prediction or estimation is one of the major problems in almost all the spheres of human activity. The estimation or prediction of future production, consumption, prices, investments, sales, profits, income etc. are of very great importance to business professionals. The most common use of regression in business is to predict events that have yet to occur.
3. Similarly, population estimates and population projections, GNP, Revenue and Expenditure etc. are indispensable for economists and efficient planning of an economy.
4. Insurance companies heavily rely on regression analysis to estimate, for example, how many policy holders will be involved in accidents or be victims of theft.
5. Optimization: Another key use of regression models is the optimization of business processes. A factory manager might, for example, build a model to understand the relationship between oven temperature and the shelf life of the cookies baked in those ovens. A company operating a call centre may wish to know the relationship between wait times of callers and number of complaints.

6. A fundamental driver of enhanced productivity in business and rapid economic advancement around the globe during the 20th century was the frequent use of statistical tools in manufacturing as well as service industries. Today, a manager considers regression an indispensable tool.

7 People use regression on an intuitive level every day. In business, a well-dressed man is thought to be financially successful. A mother knows that more sugar in her children's diet results in higher energy levels. The ease of waking up in the morning often depends on how late you went to bed the night before. Quantitative regression adds precision by developing a mathematical formula that can be used for predictive purposes.

8. For example, a medical researcher might want to use body weight (independent variable) to predict the most appropriate dose for a new drug (dependent variable). The purpose of running the regression is to find a formula that fits the relationship between the two variables. Then you can use that formula to predict values for the dependent variable when only the independent variable is known. A doctor could prescribe the proper dose based on a person's body weight.

Multiple Regression

Multiple regression is an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable). The variables we are using to predict the value of the dependent variable are called the independent variables (or sometimes, the predictor, explanatory or regressor variables).

For example, you could use multiple regressions to understand whether exam performance can be predicted based on revision time, test anxiety, lecture attendance and gender. Alternately, you could use multiple regressions to understand whether daily cigarette consumption can be predicted based on smoking duration, age when smoking, smoker type, income and gender started.

Multiple regressions also allow you to determine the overall fit (variance explained)

of the model and the relative contribution of each of the predictors to the total variance explained. For example, you might want to know how much of the variation in exam performance can be explained by revision time, test anxiety, lecture attendance and gender "as a whole", but also the "relative contribution" of each independent variable in explaining the variance.

General Purpose

The general purpose of multiple regression (the term was first used by Pearson, 1908) is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable. For example, a real estate agent might record for each listing the size of the house (in square feet), the number of bedrooms, the average income in the respective neighborhood according to census data, and a subjective rating of appeal of the house. Once this information has been compiled for various houses it would be interesting to see whether and how these measures relate to the price for which a house is sold. For example, you might learn that the number of bedrooms is a better predictor of the price for which a house sells in a particular neighborhood than how "pretty" the house is (subjective rating). You may also detect "outliers," that is, houses that should really sell for more, given their location and characteristics.

Personnel professionals customarily use multiple regression procedures to determine equitable compensation. You can determine a number of factors or dimensions such as "amount of responsibility" (Resp) or "number of people to supervise" (No_Super) that you believe to contribute to the value of a job. The personnel analyst then usually conducts a salary survey among comparable companies in the market, recording the salaries and respective characteristics (i.e., values on dimensions) for different positions. This information can be used in a multiple regression analysis to build a regression equation of the form:

Salary = .5*Resp + .8*No_Super

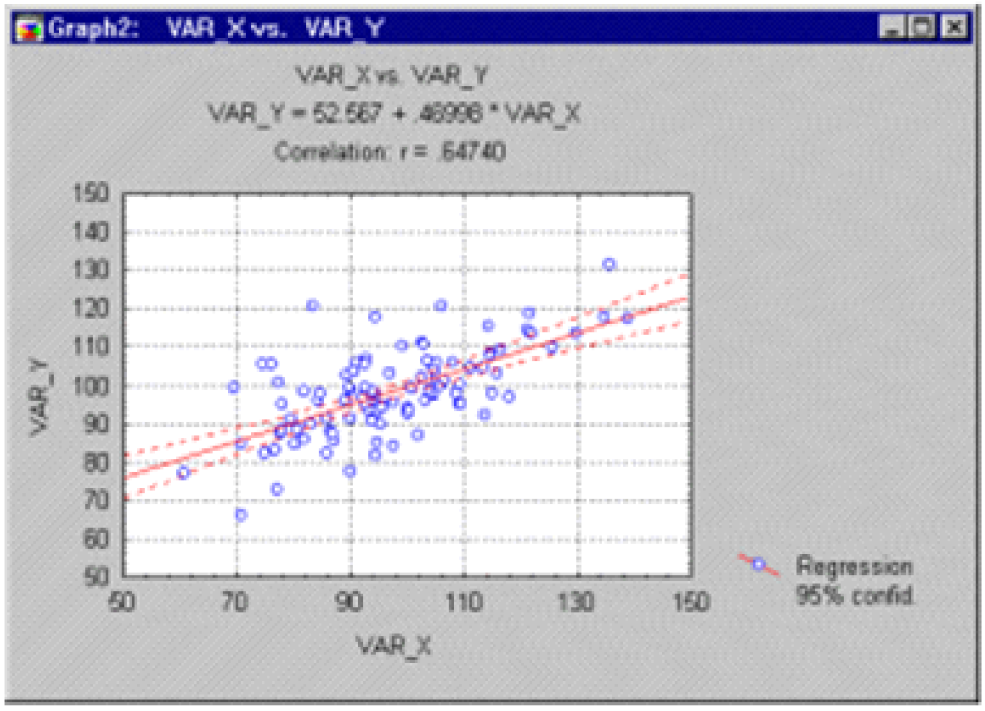
Once this so-called regression line has been determined, the analyst can now easily construct a graph of the expected (predicted) salaries and the actual salaries of job incumbents in his or her company. Thus, the analyst is able to determine

which position is underpaid (below the regression line) or overpaid (above the regression line), or paid equitably.

In the social and natural sciences multiple regression procedures are very widely used in research. In general, multiple regression allows the researcher to ask (and hopefully answer) the general question "what is the best predictor of ...". For example, educational researchers might want to learn what are the best predictors of success in high-school. Psychologists may want to determine which personality variable best predicts social adjustment. Sociologists may want to find out which of the multiple social indicators best predict whether or not a new immigrant group will adapt and be absorbed into society.

Computational Approach

The general computational problem that needs to be solved in multiple regression analysis is to fit a straight line to a number of points.



In the simplest case - one dependent and one independent variable - you can visualize this in a scatterplot.

- Least Squares
- The Regression Equation
- Unique Prediction and Partial Correlation
- Predicted and Residual Scores
- Residual Variance and R-square
- Interpreting the Correlation Coefficient R

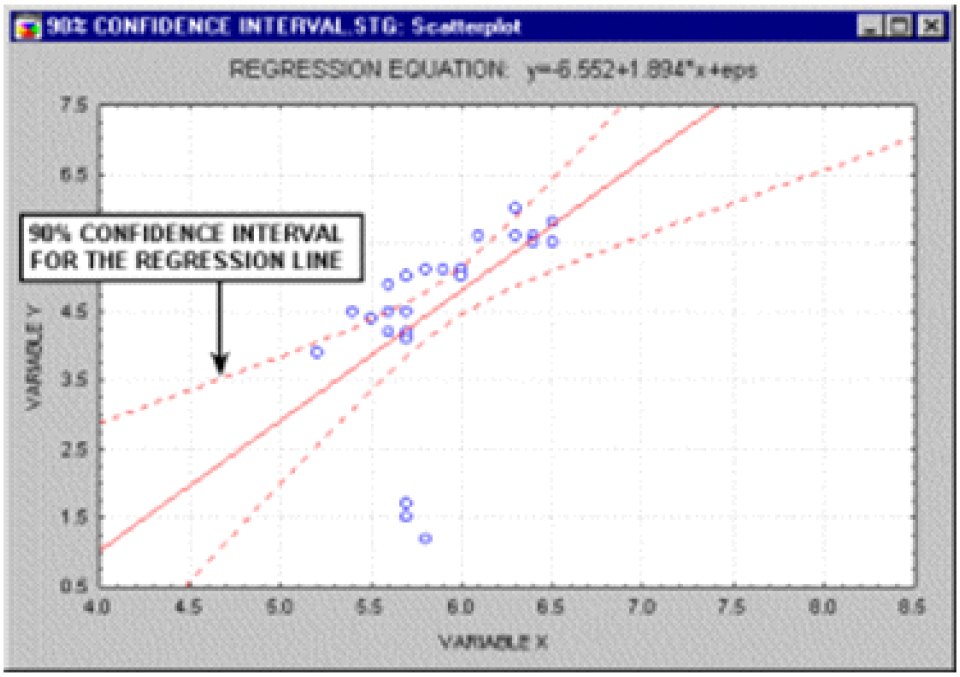
Least Squares

In the scatterplot, we have an independent or X variable, and a dependent or Y variable. These variables may, for example, represent IQ (intelligence as measured by a test) and school achievement (grade point average; GPA), respectively. Each point in the plot represents one student, that is, the respective student's IQ and GPA. The goal of linear regression procedures is to fit a line through the points. Specifically, the program will compute a line so that the squared deviations of the observed points from that line are minimized. Thus, this general procedure is sometimes also referred to as least squares estimation.

The Regression Equation

A line in a two dimensional or two-variable space is defined by the equation Y=a+b*X; in full text: the Y variable can be expressed in terms of a constant (a) and a slope (b) times the X variable. The constant is also referred to as the intercept, and the slope as the regression coefficient or B coefficient. For example, GPA may best be predicted as 1+.02*IQ. Thus, knowing that a student has an IQ of 130 would lead us to predict that her GPA would be 3.6 (since, 1+.02*130=3.6).

For example, the animation below shows a two dimensional regression equation plotted with three different confidence intervals (90%, 95% and 99%).



In the multivariate case, when there is more than one independent variable, the regression line cannot be visualized in the two dimensional space, but can be computed just as easily. For example, if in addition to IQ we had additional predictors of achievement (e.g., Motivation, Self- discipline) we could construct a linear equation containing all those variables. In general then, multiple regression procedures will estimate a linear equation of the form:

Y = a + b1*X1 + b2*X2 + … + bp*Xp

Unique prediction and partial correlation

Note that in this equation, the regression coefficients (or B coefficients) represent the independent contributions of each independent variable to the prediction of the dependent variable. Another way to express this fact is to say that, for example, variable X1 is correlated with the Y variable, after controlling for all other independent variables. This type of correlation is also referred to as a partial correlation (this term was first used by Yule, 1907). Perhaps the following example will clarify this issue. You would probably find a significant negative correlation

between hair length and height in the population (i.e., short people have longer hair). At first this may seem odd; however, if we were to add the variable Gender into the multiple regression equation, this correlation would probably disappear. This is because women, on the average, have longer hair than men; they also are shorter on the average than men. Thus, after we remove this gender difference by entering Gender into the equation, the relationship between hair length and height disappears because hair length does not make any unique contribution to the prediction of height, above and beyond what it shares in the prediction with variable Gender. Put another way, after controlling for the variable Gender, the partial correlation between hair length and height is zero.

Predicted and residual scores

The regression line expresses the best prediction of the dependent variable (Y), given the independent variables (X). However, nature is rarely (if ever) perfectly predictable, and usually there is substantial variation of the observed points around the fitted regression line (as in the scatterplot shown earlier). The deviation of a particular point from the regression line (its predicted value) is called the residual value.

Residual variance and R-square

R-Square, also known as the Coefficient of determination is a commonly used statistic to evaluate model fit. R-square is 1 minus the ratio of residual variability. When the variability of the residual values around the regression line relative to the overall variability is small, the predictions from the regression equation are good. For example, if there is no relationship between the X and Y variables, then the ratio of the residual variability of the Y variable to the original variance is equal to 1.0. Then R-square would be 0. If X and Y are perfectly related then there is no residual variance and the ratio of variance would be 0.0, making R-square = 1. In most cases, the ratio and R-square will fall somewhere between these extremes, that is, between 0.0 and 1.0. This ratio value is immediately interpretable in the following manner. If we have an R-square of 0.4 then we know that the variability of the Y values around the regression line is 1-0.4 times the original variance; in other words we have explained 40% of the original

variability, and are left with 60% residual variability. Ideally, we would like to explain most if not all of the original variability. The R-square value is an indicator of how well the model fits the data (e.g., an R-square close to 1.0 indicates that we have accounted for almost all of the variability with the variables specified in the model).

Interpreting the correlation coefficient R

Customarily, the degree to which two or more predictors (independent or X variables) are related to the dependent (Y) variable is expressed in the correlation coefficient R, which is the square root of R-square. In multiple regression, R can assume values between 0 and 1. To interpret the direction of the relationship between variables, look at the signs (plus or minus) of the regression or B coefficients. If a B coefficient is positive, then the relationship of this variable with the dependent variable is positive (e.g., the greater the IQ the better the grade point average); if the B coefficient is negative then the relationship is negative (e.g., the lower the class size the better the average test scores). Of course, if the B coefficient is equal to 0 then there is no relationship between the variables.

Assumption of linearity

First of all, as is evident in the name multiple linear regression, it is assumed that the relationship between variables is linear. In practice this assumption can virtually never be confirmed; fortunately, multiple regression procedures are not greatly affected by minor deviations from this assumption. However, as a rule it is prudent to always look at bivariate scatterplot of the variables of interest. If curvature in the relationships is evident, you may consider either transforming the variables, or explicitly allowing for nonlinear components.

Normality assumption

It is assumed in multiple regression that the residuals (predicted minus observed values) are distributed normally (i.e., follow the normal distribution). Again, even though most tests (specifically the F-test) are quite robust with regard to violations of this assumption, it is always a good idea, before drawing final conclusions, to review the distributions of the major variables of interest. You can produce

histograms for the residuals as well as normal probability plots, in order to inspect the distribution of the residual values.

Limitations

The major conceptual limitation of all regression techniques is that you can only ascertain relationships, but never be sure about underlying causal mechanism. For example, you would find a strong positive relationship (correlation) between the damage that a fire does and the number of firemen involved in fighting the blaze. Do we conclude that the firemen cause the damage? Of course, the most likely explanation of this correlation is that the size of the fire (an external variable that we forgot to include in our study) caused the damage as well as the involvement of a certain number of firemen (i.e., the bigger the fire, the more firemen are called to fight the blaze). Even though this example is fairly obvious, in real correlation research, alternative causal explanations are often not considered.

Choice of the number of variables

Multiple regression is a seductive technique: "plug in" as many predictor variables as you can think of and usually at least a few of them will come out significant. This is because you are capitalizing on chance when simply including as many variables as you can think of as predictors of some other variable of interest. This problem is compounded when, in addition, the number of observations is relatively low. Intuitively, it is clear that you can hardly draw conclusions from an analysis of 100 questionnaire items based on 10 respondents. Most authors recommend that you should have at least 10 to 20 times as many observations (cases, respondents) as you have variables; otherwise the estimates of the regression line are probably very unstable and unlikely to replicate if you were to conduct the study again.

Multicollinearity and matrix ill-conditioning

This is a common problem in many correlation analyses. Imagine that you have

two predictors (X variables) of a person's height: (1) weight in pounds and (2) weight in ounces. Obviously, our two predictors are completely redundant; weight is one and the same variable, regardless of whether it is measured in pounds or ounces. Trying to decide which one of the two measures is a better predictor of height would be rather silly; however, this is exactly what you would try to do if you were to perform a multiple regression analysis with height as the dependent (Y) variable and the two measures of weight as the independent (X) variables. When there are very many variables involved, it is often not immediately apparent that this problem exists, and it may only manifest itself after several variables have already been entered into the regression equation. Nevertheless, when this problem occurs it means that at least one of the predictor variables is (practically) completely redundant with other predictors. There are many statistical indicators of this type of redundancy (tolerances, semi-partial R, etc., as well as some remedies (e.g., Ridge regression).

Fitting cantered polynomial models

The fitting of higher-order polynomials of an independent variable with a mean not equal to zero can create difficult multicollinearity problems. Specifically, the polynomials will be highly correlated due to the mean of the primary independent variable. With large numbers (e.g., Julian dates), this problem is very serious, and if proper protections are not put in place, can cause wrong results. The solution is to "center" the independent variable (sometimes, this procedures is referred to as "centered polynomials"), i.e., to subtract the mean, and then to compute the polynomials. See, for example, the classic text by Neter, Wasserman, & Kutner (1985, Chapter 9), for a detailed discussion of this issue (and analyses with polynomial models in general).

3.4.2 Linear and Non-linear Regression

Linear Regression

Let us consider the following data. The number of visitors and the number of books issued during weekdays in a week are given.

No. of visitors to a library (X)	6	2	1	0	4	8

No. of books issued (Y) 8 4 1 0 7 8

If we plot the data on a graph paper, the scatter diagram looks something like

J

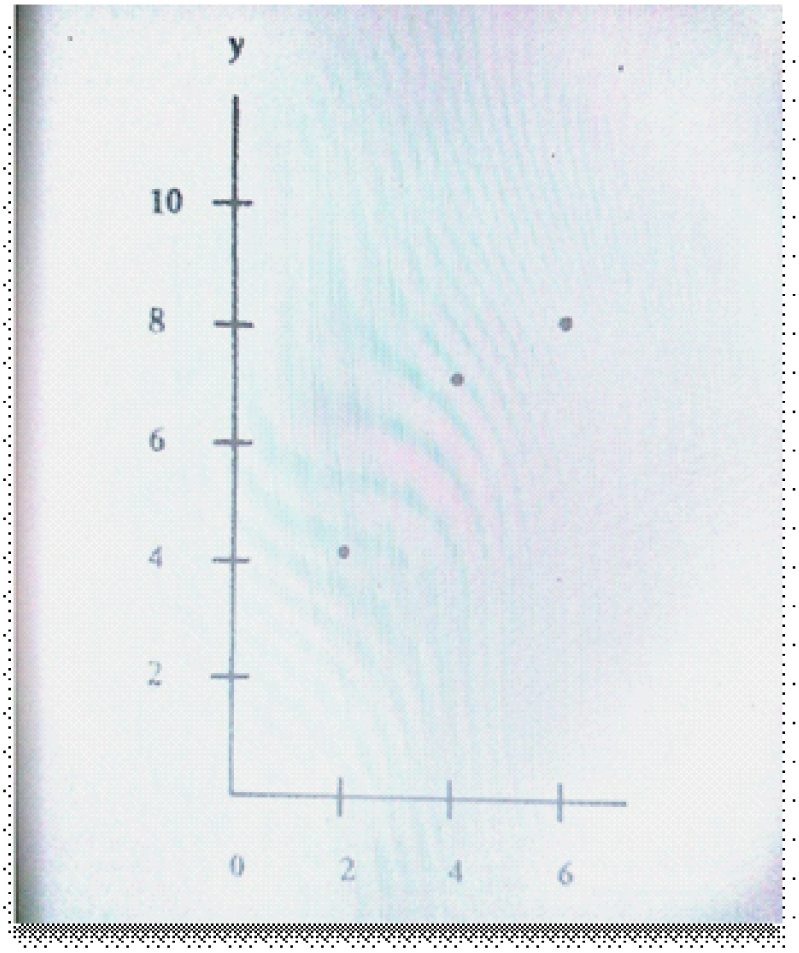
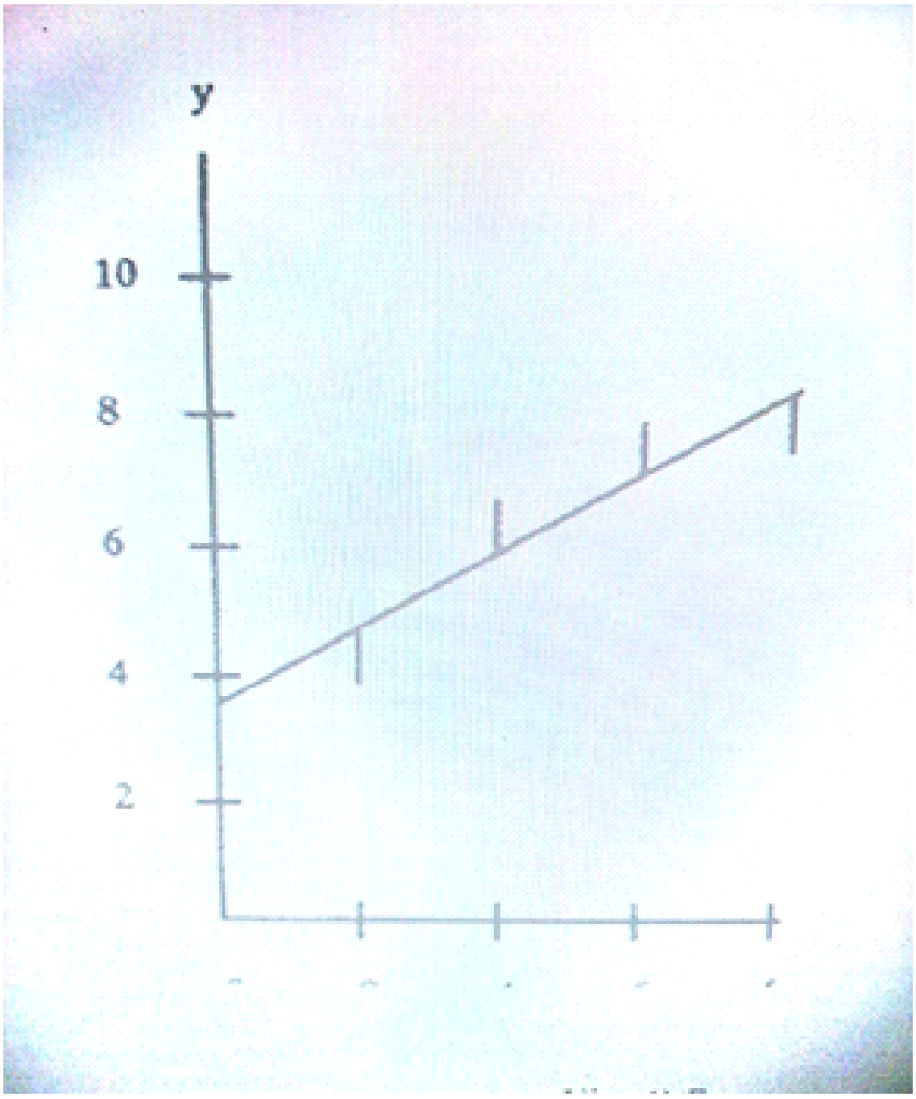


Fig 1: Scatter Diagram for No. of visitors to a library and Books issued

As is obvious from the graph, the points do not strictly lie in a straight line. But they show an upward rising tendency where a straight line can be fitted. If we plot the straight line along with the scattered points, the diagram looks like Figure 2. The difference between the regression line and the observations is the 'error'. For example, against X value of 2, the Y value is 4. This is called the observed value.



But the regression line shows Y value of 4.8 against X value of 2. This value, which is calculated from the regression line, is the expected value. The difference between the observed value and the expected value is termed as the error value. So we see that observed value is the sum of expected value and error and value.

Our objective in fitting a regression line is to minimise the error values. This is usually done by the method of 'least squares'. The method of least squares

minimises the value of E, where e is the difference between observed value and expected value. We will not go into the details of the method here. Instead, two equations derived on the basis of least squares method and known as normal equations are given.

These are:

Y na b X

As a rule of thumb we can say that these normal equations are derived by multiplying the coefficients of 'a' and 'b' to the linear equation and summing over all observations.

Here the linear equation is Y = a + bX. The first normal equation is simply the linear equation Y = a +bX summed over all observations.

SY = a +SbX or SY = na + bSX

The second normal equation is the linear equation multiplied by X and summed over all observations.

SXY = SaX + SbX2 or SXY = aSX +bSX 2

It is evident that all the terms in these equations are given numbers, calculated from the data, except a and b.

The values of a and b need to be calculated for getting the estimated value of the dependent variable. This is done in the following.

It can be seen from Table 8.1 that data on X and Y variables are given in the first two columns. The succeeding two columns in the table give the calculations necessary to solve two normal equations given above. The expected value of Y and error value are given in the last two columns.

Table 8.1: Computations of Data for Regression Analysis

(1) X	(2) Y	(3) X2	(4) XY	(5) Expected value of Y	(6) Error
6	8	36	48	7.4	0.6
2	4	4	8	4.8	-0.8
10	10	100	100	10.0	0.0
4	7	16	28	6.1	0.9
8	8	64	64	8.7	-0.7
Total 30	37	220	248	37	0.0

As the normal equations are:

$Y = na + b X$ ..... (8, 8)

$XY = a X + b X2$ ..... (8.9)

We substitute the respective values from the Table 1.

Thus,

$37 = 5a + 30b$ .....(8.10)

$248 = 30a + 220b$ .....(8.11)

If we multiply equation (3) by 6 and subtract the product from equation (4) we get :

$248 = 30a + 220b$

$222 = 30a + 180b$

\- - -

$40b = 26$

Or $b = 0.65$

On substituting the value of b in equation (8.10) we get:

$37 = 5a + 30 \times 0.65$ or $5a = 17.5$ or $a = 3.5$

So the regression line is

$Y = 3.5 + 0.65\ X$ ……(8.12)

If we substitute the values of X in the regression line that is equation (8.12), we get the expected values of Y. For example, when X = 2

$Y = 3.5 + 0.65 \times 2 = 4.8$

But our observed value of Y against X = 2 is 4. This difference between the observed value and the expected value (4 - 4.8 = -0.8) is the error 'e'.

The expected values of Y and e are given in the Table 8.1 above in columns (5) and (6). Notice that the sum of errors for the sample is zero, i.e. Se = 0

For computational purposes we may use the following formulae to find out the value of a and b.

Non-linear Regression

In the previous sub-section we discussed the simple linear regression involving two variables: one dependent and the other independent. Regression can involve one dependent variable and more than one independent variable. Such cases are called multiple regressions.

The equation fitted in regression can be non-linear or curvilinear. It can take numerous forms. A simpler form involving two variables is the quadratic form. The equation is

$Y = a + bX + CX2$

There are three parameters here, viz., a, b and c, and the normal equations are :

$Y = na + b\ X + c\ X2$

$XY = a\ X + b\ X2 + c\ X3$

X2 Y = a X2 + b X3 + c X4

Notice again that the normal equations are the regression equation multiplied by the coefficients of a, b and c and summed over all observations.

Certain non-linear equations can be transformed into linear equations by taking logarithms. Finding out the optimum values of the parameters from the transformed linear equations is the same as the process discussed in the previous sections. We give below some of the frequently used non-linear equations and the respective transformed linear equations.

1) $Y = a\ e^{bx}$

By taking natural log (that is, ln), it can be written as

$\ln Y = \ln a + b X$

Or $Y = a + ß X$

Where, $Y = \ln Y$, $a = \ln a$, $X = X$ and $ß = b$

2) $Y = a\ X^b$

By taking log, the equation can be transformed into

$\log Y = \log a + b \log X$

Or $Y = a + ß X$

Where, $Y = \log Y$, $= \log a$, $ß = b$ and $X = \log X$

3)

$Y = a + b X$

If we take $X = X$ then

$Y = a + b X$

Once the non-linear equation is transformed, the fitting of a regression line is as per the method discussed in the beginning of this section. We derive the normal equations and substitute the values calculated from the observed data. From the transformed parameters, the actual parameters can be obtained by making the

reverse transformation.

Objectives of Regression Analysis

- To research a pattern linking the dependent variable and independent variables by establishing a functional relationship between the two. In this equation the level of relationship comes from which is a matter of interest to the researcher in his study.
- To make use of the well-established regression equation for problems concerning forecasting.
- To analyze how much of the variation in the dependent variable is described by the group of independent variables. This would allow him to get rid of particular unwanted variables from the system.

Advantages of Regression Analysis

1. Regression analysis provides estimates of values of the dependent variables from the values of independent variables.
2. Regression analysis also helps to obtain a measure of the error involved in using the regression line as a basis for estimations.
3. Regression analysis helps in obtaining a measure of the degree of association or correlation that exists between the two variables.

Assumptions in Regression Analysis

1. Existence of actual linear relationship.
2. The regression analysis is used to estimate the values within the range for which it is valid.
3. The relationship between the dependent and independent variables remains the same till the regression equation is calculated.
4. The dependent variable takes any random value but the values of the independent variables are fixed.
5. In regression, we have only one dependant variable in our estimating

equation. However, we can use more than one independent variable.

3.4.3 Statement of Regression Lines

A regression line is a straight line that describes how a response variable y changes as an explanatory variable x changes. We often use a regression line to predict the value of y for a given value of x.

In the pairs of observations, if there is a cause and effect relationship between the variables X and Y, then the average relationship between these two variables is called regression, which means "stepping back" or "return to the average". The linear relationship giving the best mean value of a variable corresponding to the other variable is called a regression line or line of the best fit. The regression line (known as the least squares line) is a plot of the expected value of the dependent variable for all values of the independent variable. Technically, it is the line that "minimizes the squared residuals". The regression line is the one that best fits the data on a scatterplot.

Using the regression equation, the dependent variable may be predicted from the independent variable. The slope of the regression line (b) is defined as the rise divided by the run. The y intercept (a) is the point on the y axis where the regression line would intercept the y axis. The slope and y intercept are incorporated into the regression equation. The intercept is usually called the constant, and the slope is referred to as the coefficient. Since the regression model is usually not a perfect predictor, there is also an error term in the equation.

In the regression equation, y is always the dependent variable and x is always the independent variable. Here are three equivalent ways to mathematically describe a linear regression model.

y = intercept + (slope ± x) + error

y = constant + (coefficient ± x) + error

$y = a + bx + e$

The significance of the slope of the regression line is determined from the t-statistic. It is the probability that the observed correlation coefficient occurred

by chance if the true correlation is zero. Some researchers prefer to report the F-ratio instead of the t-statistic. The F-ratio is equal to the t-statistic squared.

The t-statistic for the significance of the slope is essentially a test to determine if the regression model (equation) is usable. If the slope is significantly different than zero, then we can use the regression model to predict the dependent variable for any value of the independent variable.

On the other hand, take an example where the slope is zero. It has no prediction ability because for every value of the independent variable, the prediction for the dependent variable would be the same. Knowing the value of the independent variable would not improve our ability to predict the dependent variable. Thus, if the slope is not significantly different than zero, don't use the model to make predictions.

The coefficient of determination (r-squared) is the square of the correlation coefficient. Its value may vary from zero to one. It has the advantage over the correlation coefficient in that it may be interpreted directly as the proportion of variance in the dependent variable that can be accounted for by the regression equation. For example, an r-squared value of .49 means that 49% of the variance in the dependent variable can be explained by the regression equation. The other 51% is unexplained.

The standard error of the estimate for regression measures the amount of variability in the points around the regression line. It is the standard deviation of the data points as they are distributed around the regression line. The standard error of the estimate can be used to develop confidence intervals around a prediction.

3.4.4 Regression Coefficients

Properties of Regression Coefficient

Coefficient 'b' is called the regression coefficient. Notice that we can draw two regression lines from the data on X and Y.

(a) Y on X line, $Y = a + bX$

(b) X on Y line, $X = + Y$

The two coefficients, b and ß, demonstrate some interesting properties. First, the product of both regression coefficients is equal to the square of r (correlation coefficient), i.e., b ß = r 2

So once we know both regression coefficients we can find out the value of r2. By taking the square root of r2 we get r. Second if the regression coefficients are negative in sign, then the correlation coefficients also is negative. If the regression coefficients are positive then correlation is positive. Third, you know that

, i.e., r lies between - 1 and +1

Therefore r2 lies between zero and +1. Regression coefficient can take finite value. But is one regression coefficient is more than 1 the other regression coefficient is less than 1. Both regression coefficients cannot exceed unity. Also it follows that the product of both, which is r2, cannot exceed unity. The square of correlation coefficient is called the coefficient of determination and implies important characteristics. If r2, the coefficient of determination, is closer to one we can infer that the independent variable explains the movements in the dependent variable. If the coefficient of determination is closer to zero, the independent variable does not explain the variation in the dependent variable.

Limitations of Regression Analysis

There are three main limitations:

1. Parameter Instability - This is the tendency for relationships between variables to change over time due to changes in the economy or the markets, among other uncertainties. If a mutual fund produced a return history in a market where technology was a leadership sector, the model may not work when foreign and small-cap markets are leaders.

2. Public Dissemination of the Relationship - In an efficient market, this can limit the effectiveness of that relationship in future periods. For example, the discovery that low price to book value stocks outperform high price-to-book value means that these stocks can be bid higher, and value-based investment approaches will not retain the same relationship as in the past.

3. Violation of Regression Relationships - Earlier we summarized the six classic

assumptions of a linear regression. In the real world these assumptions are often unrealistic - e.g. assuming the independent variable X is not random.

Example of Simple Regression

A company wants to know if there is a significant relationship between its advertising expenditures and its sales volume. The independent variable is advertising budget and the dependent variable is sales volume. A lag time of one month will be used because sales are expected to lag behind actual advertising expenditures. Data was collected for a six month period. All figures are in thousands of dollars. Is there a significant relationship between advertising budget and sales volume?

Independent variable	Dependent variable
4.2	27.1
6.1	30.4
3.9	25.0
5.7	29.7
7.3	40.1
5.9	28.8

Model: $y = 9.873 + (3.682 \pm x) + \text{error}$

Standard error of the estimate = 2.637

T-test for the significance of the slope = 3.961

Degrees of freedom = 4

Two-tailed probability = .0149

r-squared = .807

You might make a statement in a report like this: A simple linear regression was performed on six months of data to determine if there was a significant relationship between advertising expenditures and sales volume. The t-statistic

for the slope was significant at the .05 critical alpha level, t(4)=3.96, p=.015. Thus, we reject the null hypothesis and conclude that there was a positive significant relationship between advertising expenditures and sales volume. Furthermore, 80.7% of the variability in sales volume could be explained by advertising expenditures.

3. 5 Difference between Regression and Correlation

Correlation and regression analysis are related in the sense that both deal with relationships among variables. The difference between them as given below:-

i) A statistical measure which determines the co-relationship or association of two quantities is known as Correlation. Regression describes how an independent variable is numerically related to the dependent variable.

ii) The objective of correlation is to find a numerical value expressing the relationship between variables while as in regression; the objective is to estimate values of random variable on the basis of the values of fixed variable.

iii) In regression the emphasis is on predicting one variable from the other, in correlation the emphasis is on the degree to which a linear model may describe the relationship between two variables.

iv) In regression the interest is directional, one variable is predicted and the other is the predictor; in correlation the interest is non-directional, the relationship is the critical aspect.

v) Correlation is used to represent the linear relationship between two variables. On the contrary, regression is used to fit the best line and estimate one variable on the basis of another variable.

vi) In correlation, there is no difference between dependent and independent variables i.e. correlation between x and y is similar to y and x. Conversely, the regression of y on x is different from x on y.

vii) Correlation indicates the strength of association between variables. As opposed to, regression reflects the impact of unit change in the independent

variable on the dependent variable.

viii) Correlation aims at finding a numerical value that expresses the relationship between variables. Unlike regression whose goal is to predict values of the random variable on the basis of the values of fixed variable.

ix) Correlation makes no a priori assumption as to whether one variable is dependent on the other(s) and is not concerned with the relationship between variables; instead it gives an estimate as to the degree of association between the variables. In fact, correlation analysis tests for interdependence of the variables. As regression attempts to describe the dependence of a variable on one (or more) explanatory variables; it implicitly assumes that there is a one-way causal effect from the explanatory variable(s) to the response variable, regardless of whether the path of effect is direct or indirect.

With the above discussion it is evident, that there is a big difference between these two concepts, although these two are studied together. Correlation is used when the researcher wants to know that whether the variables under study are correlated or not, if yes then what is the strength of their association. Pearson's correlation coefficient is regarded as a best measure of correlation. In regression analysis, a functional relationship between two variables is established so as to make future projections on events.

3.6 Summary

This chapter describes simple correlation and regression analysis which are widely used for explaining variations in social sciences, psychology and marketing research such as market share, brand preferences and other marketing results in terms of marketing management variables such as advertising, price, distribution and product quality.

In introducing the simple correlation, we also discussed its types, importance and advantages. Further we also discussed the karl pearson's coefficient of simple correlation, its importance, assumptions, properties, advantages and disadvantages. After that spearman's rank correlation, its importance, advantages and disadvantages are also discussed. Problems in both

the cases such as karl pearson's coefficient of simple correlation and spearman's rank correlation of are also discussed. Simple regression analysis, its importance, advantages and disadvantages, regression line are explained. Difference between correlation and regression are discussed finally in this chapter.

3.7 Glossary

Correlation: means the average relationship or association between two or more variables.

Correlation Coefficient: is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another.

Karl Pearson's Product-Moment Correlation Coefficient or simply Pearson's Correlation Coefficient for short is one of the important methods used in Statistics to measure correlation between two variables. It is a common measure of the correlation between two variables X and Y.

Simple Correlation: is the strength of association between two variables. Simply, because of the presence of correlation between two variables only.

Spearman's Rank-Order Correlation: is the nonparametric version of the pearson product moment correlation. Spearman's correlation coefficient, measures the strength of association between two ranked variables.

Rank Correlation: If ranks can be allocated to pairs of observations for two variables X and Y, then the correlation between the ranks is called the rank correlation coefficient.

Regression Analysis: The goal in regression analysis is to create a mathematical model that can be used to predict the values of a dependent variable based upon the values of an independent variable.

3.8 Self Assessment Questions

1. Explain the aim of correlation analysis? Discuss its types.
2. What is Rank difference method of computing coefficient of correlation? Discuss its procedure with the help of hypothetical example.

3. Mention and explain the formula for computing rank correlation coefficient.

5. Mention and explain the formula for computing spearman's coefficient of rank correlation method.

6. Explain the concept of regression.

7. Describe the types of regression analysis

3.9 Lesson End Exercises

1. Discuss and explain the formula for computing karl pearson's coefficient of simple correlation method.

2. Explain how to resolve ties while calculating ranks.

3. How can u differentiate correlation and regression analysis? Mention some examples?

4. Explain the managerial uses of correlation analysis and regression analysis.

3.10 Suggested Readings

Aaker, D. A., Kumar, V. & Day, G. S. Marketing Research, 7th edn, John Wiley, New York.

Sachdeva, J.K. Business Research Methodology, Himalaya Publishing House; New Delhi.

Shajahan, S. Research Methods for Management, Jaico Publishing House, Delhi; India.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. Multivariate data analysis: A global perspective (7th ed.). Upper Saddle River: Pearson Education.

Kent, R. Marketing Research: Approaches, methods and applications in Europe. London: Thomson.

Kumar, V., Aaker, D. A., & Day, G. S. Essentials of marketing research New York: John Wiley & Sons, Inc.

3.11 References

Bryman, A. Social Research Methods. London: Oxford University Press.

Hair, J. F., Celsi, M., Money, A., Samouel, P., & Page, M. Essentials of business research methods (2nd ed.). Armonk, NY: ME Sharpe

Johnson, R.A. & Wichern, D.W., Applied Multivariate Statistical Analysis, Pearson Education, Delhi.

Panneerselvam, R., Research Methodology, Prentice hall of India, New Delhi.

Kanji, G. K., 100 Statistical Tests, Sage Publications, New Delhi.

Kothari C. R. Research Methodology-Methods and Techniques, New Wiley Eastern ltd., Delhi.

Malhotra, N. K. Marketing Research (4th ed.). Harlow: Prentice Hall.

Malhotra, N.K. Marketing Research (4th ed.). New Jersey: Pearson.

Rencher, A.V., Methods of Multivariate Analysis, Wiley Inter-Science, Second Edition, New Jersey.

RESEARCH

Structure

4.1 Introduction of research:

All societies from the primitive to the most modern sophisticated societies have progressed only on the acquisition of knowledge and its application, depending upon their capability to understand their environments and control them through concerted efforts. Initially knowledge acquisition was more on the basis of observation, experience, learning by trial and error, simple logics of deduction and inference, etc. But with the increasing ability to conduct research and getting positive results and the ability to apply them in solving problems, although confined to a few individuals, human societies were slowly advancing materially.

With science and technology opening up new directions of growth and development from the 15th century in Western Europe and its influence in other parts of the world, methods of research have become a mode of acquiring knowledge through scientific methods. It was largely an individual flair that pushed up the frontiers of knowledge albeit with very limited facilities for research. With the advent of universities, research became one of their important functions, besides their teaching, training, and publications functions. Increasing pursuit of research has resulted in the growth of a body of literature over the years on research methodology, which has now developed into a subject in its own right.

In the course of time, institutions, associations and cognate bodies, have been established to deal with various development problems through research, with financial aids from governments and industry. Today there are research institutions, which have been set up to deal exclusively with research in different subjects, including library and information science.

In this Unit, we are trying, in a general way, to study the subject of research methodology in all its dimensions. Formal definitions of research need to pursue research to expand the horizons of knowledge, contours of research processes

with an understanding of the conceptual framework model of research methodology, characteristics of research, scientific research, research design and other related aspects are discussed in this Unit. Another important point to be noted in a study of research methodology by students of library and information science is not only to get the necessary skills in doing research in their own field but also to be of assistance and help to the research community offering high quality information service. This aspect is also elaborated in this Unit.

There are likely to be some overlapping of ideas in discussing these aspects in the different sections of this unit. They are reiterations and should be understood in the contexts in which each of these ideas is discussed.

4.1.1 Concept of research

The general meaning of a concept is that it is an idea complex of something, formed up of its characteristics. It is a construct, putting together all parts of a thing. We can perceive concepts of research in two sets. One, as their attributes for ensuring high quality of research, and another from the functional or operational point of view.

Generally a research topic is identified for study and research in a discipline, either to add further to the existing knowledge by creating new ideas or modifying existing knowledge by new findings. A research topic is always selected on the basis of the theoretical knowledge of a subject possibly to bridge up a gap or reinterprets a known finding or finds new ideas to solve a new problem. To do this, a hypothesis is formulated, identifying variables to test the hypothesis. The process gets to the next step to find evidences by collecting appropriate data or facts by a suitable method, analyse the collected data systematically, interpret the results to arrive at conclusions and generalise the findings, possibly to be applied to an appropriate or a typical situation to test the result and its validity and finally add the validated findings to the already existing body of knowledge after peer review.

4.1.2 Meaning of research

The task of business research is to generate accurate information for use in decision

making as we say above, the emphasis of business research is on shifting decision makers from intuitive information gathering to systematic and objective investigation. Business research is defined as the systematic and objective process of gathering, recording, and analyzing data for aid in making business decisions.

This definition suggests, first, that research information is neither intuitive nor haphazardly gathered. Literally, research (re-search) means to "search again". It connotes patient study and scientific investigation wherein the researcher takes another, more careful look at data to discover all that can be known about the subject of study.

Second, if the information generated or data collected and analyzed are to be accurate, the business researcher must be objective. The need for objectivity was cleverly stated by the nineteenth-century American humorist Artemus Ward, who said, "It ain't the things we don't know that gets us in trouble. It's the things we know that ain't so". Thus the role of the researcher is to be detached and impersonal, rather than biased in an attempt to prove preconceived ideas. If bias enters the research process, the value of the data is considerably reduced.

A developer who owned a large area of land on which he wished to build a high-prestige shopping centre wanted a research report to demonstrate to prospective retailers that there was a large market potential for such a centre. Because he conducted his survey exclusively in an elite neighbourhood, not surprisingly his findings showed that a large percentage of respondents wanted a "high-prestige" shopping centre. Results of this kind are misleading, of course, and should be disregarded. If the user of such findings discovers how they were obtained, the developer loses credibility. If the user is ignorant of the bias

4.1.3 Significance of research

The scope of business research is limited by one's definition of "business". Certainly research in the production, finance, marketing, or management areas of a for-profit corporation is within the scope of business research. A broader definition of business, however, includes not-for-profit organizations, such as the American Heart Association, the Sac Diego Zoo, and the Boston Pops Orchestra, Each of these organizations exists to satisfy social needs, and they

require business skills to produce and distribute the services that people want. Business research may be conducted by organizations that are not business organizations. The reserve bank of India, for example, performs many functions that are similar, if not identical, to those of business organizations. Reserve bank economists may use research techniques for evaluative purposes much the same way as managers at Reliance or Ford. The term business research is utilized because all its techniques are applicable to business settings.

Business research covers a wide range of phenomena. For managers the purpose of research is to fulfil their need for knowledge of the organization, the market, the economy, or another area of uncertainty. A financial manager may ask, "Will the environment for long-term financing be better two years from now?" A personnel manager may ask, "What kind of training is necessary for production employees?" or "What is the reason for the company's high turnover?" A marketing manager may ask, "How can I monitor my sales in retail trade activities?"

4.2 Objectives

After reading this lesson you should be able to-

- ✓ Describe what research is and how is it defined;
- ✓ Distinguish between descriptive and exploratory research;
- ✓ Explain why managers should know about research;
- ✓ Discuss what managers should and should not do in order to interact most effectively with researchers.
- ✓ Explain research design and process.

4.3 Application of research in various functions of management

We have argued that research facilitates effective management. At the Ford Motor Company a marketing manager stated, "Research is fundamental to everything we do, so much so that we hardly make any significant decision without the benefit of some kind of market research. The risks are too big." Managers in other functional areas have similar beliefs about research in their specialties.

The prime managerial value of business research is that it reduces uncertainty by

providing information that improves the decision-making process. The decision making process associated with the development and implementation of a strategy involves three interrelated stages.

1. Identifying problems or opportunities
2. Selecting and implementing a course of action
3. Evaluating the course of action

Business research, by supplying managers with pertinent information, may play an important role by reducing managerial uncertainty in each of these stages.

Identifying Problems or Opportunities

Before any strategy can be developed, an organization must determine where it wants to go and how it will get there. Business research can help managers plan strategies by determining the nature of situations by identifying the existence of problems or opportunities present in the organization.

Business research may be used as a diagnostic activity to provide information about what is occurring within an organization or in its environment. The mere description of some social or economic activity may familiarize managers with organizational and environmental occurrences and help them understand a situation. For example, the description of the dividend history of stocks in an industry may point to an attractive investment opportunity.

Information supplied by business research may also indicate problems. For example, employee interviews undertaken to delineate the dimensions of an airline reservation clerk's job may reveal that reservation clerks emphasize competence in issuing tickets over courtesy and friendliness in customer contact. Once business research indicates a problem, managers may feel that the alternatives are clear enough to make a decision based on experience or intuition, or they may decide that more business research is needed to generate additional information for a better understanding of the situation.

Whether an organization recognizes a problem or gains insight into a potential opportunity, an important aspect of business research is its provision of information

that identifies or clarifies alternative courses of action.

Selecting and implementing a course of action

After the alternative courses of action have been identified, business research is often conducted to obtain specific information that will aid in evaluating the alternatives and in selecting the best course of action. For example, suppose a facsimile (fax) machine manufacturer must decide to build a factory either in Japan or in Sough Korea. In such a case, business research can be designed to supply the exact information necessary to determine which course of action is best of the organization.

Opportunities may be evaluated through the use of various performance criteria. For example, estimates of market potential allow managers to evaluate the revenue that will be generated by each of the possible opportunities. A good forecast supplied by business researchers is among the most useful pieces of planning information a manager can have. Of course, complete accuracy in forecasting the future is not possible because change is constantly occurring in the business environment. Nevertheless, objective information generated by business research to forecast environmental occurrences may be the foundation for selecting a particular course of action.

Clearly, the best plan is likely to result in failure if it is not properly implemented. Business research may be conducted with the people who will be affected by a pending decision to indicate the specific tactics required to implement that course of action.

Evaluating course of action

After a course of action has been implemented, business research may serve as a tool to inform managers whether planned activities were properly executed and whether they accomplished what they were expected to accomplish. In other words, business research may be conducted to provide feedback for evaluation and control of strategies and tactics.

Evaluation research is the formal, objective measurement and appraisal of the extent to which a given action, activity, or program has achieved its objectives.

In addition to measuring the extent to which completed programs achieved their objectives or to which continuing programs are presently performing as projected, evaluation research may provide information about the major factor influencing the observed performance levels.

In addition to business organization, non-profit organization, such as agencies of the federal government, frequently conduct evaluation research. It is estimated that every year more than, 1,000 federal evaluation studies are undertaken to systematically assess the effects of public programs. For example, the General Accounting Office has been responsible for measuring outcomes of the Employment Opportunity Act, the Head Start program, and the Job Corps program.

Performance-monitoring research is a term used to describe a specific type of evaluation research that regularly, perhaps, routinely, provides feedback for the evaluation and control of recurring business activity. For example, most firms continuously monitor wholesale and retail activity to ensure early detection of sales declines and other anomalies. In the grocery and retail drug industries, sales research may use the Universal Product Code (UPC) for packages, together with computerized cash registers and electronic scanners at checkout counters, to provide valuable market share information to store and brand managers interested in the retail sales volume of specific product.

United Airlines' Omnibus in-flight surveys provide a good example of performance monitoring research. United routinely selects sample flights and administers questionnaire about in-flight service, food and other aspects of air travel. The Omnibus survey is conducted quarterly to determine who is flying and for what reasons. It enables United to track demographic changes and to monitor customer ratings of its services on a continuing basis, allowing the airline to gather vast amounts of information at low cost. The information relating to customer reaction to services can be compared over time. For example, suppose United decided to change its menu for in-flight meals. The results of the Omnibus survey might indicate that shortly after the menu changed, the customers' rating of the airline's food declined. Such information would be extremely valuable, as it would allow management to quickly spot similar trends among passengers in other aspects of

air travel, such as airport lounges, gate-line waits, or cabin cleanliness, thus managerial action to remedy problems could be rapidly taken.

When analysis of performance indicated that all is not going as planned, business research may be required to explain why something "went wrong." Detailed information about specific mistakes or failures is frequently sought. If a general problem area is identified, breaking down industry sales volume and a firm's sales volume into different geographic areas may provide an explanation of specific problems, and exploring these problems in greater depth may indicate which managerial judgments were erroneous.

Research Applications in Various Functions of Management

While many business ideas blossom into successful businesses, there are many others that did not move beyond the business plan or offering memorandum. And among those that get funded and started, many fail eventually. While there can be many reasons for business failures, often these were due to flaws in the business concept or the business model that rendered them vulnerable to the difficulties encountered.

Recognizing the relevance of research in business planning has become even more crucial in the current economic scenario with funding getting difficult as banks, other financial institutions and venture capitalists are bound to put business propositions through a most rigorous assessment process. Research indicates that many ventures fail on account of market and industry factors. Yet, very often we come across ambitious entrepreneurs starting ventures without having researched the market and industry to determine the viability and sustainability of the business concept. The significance of research stems from the fact the success of a business does not depend only on the entrepreneur's perspective on the service or product offered. It also depends greatly on what, the customers want or need. Here arises the need of research. The research carried out, in different areas, is called management research, production research, personnel research, financial management research, accounting research, Marketing research etc.

Management research includes various functions of management such as

planning, organizing, staffing, communicating, coordinating, motivating, controlling. Various motivational theories are the result of research. Production (also called manufacturing) research focuses more on materials and equipment rather than on human aspects. It covers various aspects such as new and better ways of producing goods, inventing new technologies, reducing costs, improving product quality.

Research in personnel management may range from very simple problems to highly complex problems of all types. It is primarily concerned with the human aspects of the business such as personnel policies, job requirements, job evaluation, recruitment, selection, placement, training and development, promotion and transfer, morale and attitudes, wage and salary administration, industrial relations. Basic research in this field would be valuable as human behaviour affects organizational behaviour and productivity.

Research in Financial Management includes financial institutions, financing instruments (for example shares, debentures), financial markets (capital market, money market, primary market, secondary market), financial services (for example merchant banking, discounting, factoring), financial analysis (e.g. investment analysis, ratio analysis, funds flow/cash flow analysis) etc.

Accounting research though narrow in its scope, but is a highly significant area of business management. Accounting information is used as a basis for reports to the management, shareholders, investors, tax authorities, regulatory bodies and other interested parties. Areas for accounting research include inventory valuation, depreciation accounting, generally accepted accounting principles, accounting standards, corporate reporting etc.

Marketing research deals with product development and distribution problems, marketing institutions, marketing policies and practices, consumer behaviour, advertising and sales promotion, sales management and after sales service etc. **Marketing research** is one of the very popular areas and also a well established one. Marketing research includes market potentials, sales forecasting, product testing, sales analysis, market surveys, test marketing, consumer behaviour studies, marketing information system etc.

Business policy research is basically the research with policy implications. The results of such studies are used as indices for policy formulation and implementation. It helps in developing the standards, objectives, long-term goals, and growth strategies.

Production Management: The research plays a significant role in product development, diversification, launching a new product, product improvement, process technologies, selecting a site, new investment etc.

Materials Management: It is used in selecting the supplier, taking the decisions pertaining to make or buy as well as in deciding negotiation strategies.

When is business research needed?

A manager faced with two or more possible courses of action faces the initial decision of whether or not research should be conducted. The determination of the need for research centres on (1) time constraints, (2) the availability of data, (3) the nature of the decision that must be made, and (4) the value of the business research information in relation to its costs.

Time constraints

Systematically conducting research takes time. In many instances management concludes that because a decision must be made immediately, there will be no time for research. As a consequence, decisions are sometimes made without adequate information or thorough understanding of the situation. Although not ideal, sometimes the urgency of a situation precludes the use of research.

Availability of data

Frequently managers already possess enough information to make a sound decision without business research. When there is an absence of adequate information, however, research must be considered. Managers must ask themselves, "Will the research provide the information needed to answer the basic questions about this decision?" If the data cannot be made available, research cannot be conducted. For example, prior to 1980 the people's republic of China had never conducted a population census. Organizations engaged in international business often find that data about business activity or population characteristics,

found in abundance when investigating the United States, are nonexistent or sparse when the geographic area of interest is an underdeveloped country. Further, if a potential source of data exists, managers will want to know how much it costs to obtain those data.

Nature of the decision

The value of business research will depend on the nature of the managerial decision to be made. A routine tactical decision that does not require a substantial investment may not seem to warrant a substantial expenditure for business research. For example, a computer software company must update its operator's instruction manual when minor product modifications are made. The cost of determining the proper wording for the updated manual is likely to be too high for such a minor decision. The nature of such a decision is not totally independent from the next issue to be considered: the benefits versus the costs of the research. However, in general the more strategically or tactically important the decision, the more likely that research will be conducted.

Benefits versus costs

Some of the managerial benefits of business research have already been discussed. Of course, conducting research activities to obtain these benefits requires expenditure; thus there are both costs and benefits in conducting business research. In any decision-making situation, managers must identify alternative courses of action, and then weigh the value of each alternative against its cost. It is useful to think of business research as an investment alternative. When deciding whether to make a decision without research or to postpone the decision in order to conduct research, managers should ask: (1) Will the payoff or rate of return be worth the investment? (2) Will the information gained by business research improve the quality of the decision to an extent sufficient to warrant the expenditure? And (3) Is the proposed research expenditure the best use of the available funds?

For example, TV Cable Week was not test-marketed before its launch. While the magazine had articles and stories about television personalities and events, its main feature was a channel-by-channel program listing showing the exact

programs that a particular subscriber could receive. To produce a "custom" magazine for each individual cable television system in the country required developing a costly computer system. Because development required a substantial expenditure, one that could not be scaled down for research, the conducting of research was judged to be an improper investment. The value of the research information was not positive, because the cost of the information exceeded its benefits. Unfortunately, pricing and distribution problems became so compelling after the magazine was launched that it was a business failure. Nevertheless, the publication's managers, without the luxury of hindsight, made a reasonable decision not to conduct research. They analyzed the cost of the information (i.e. the cost of business research) relative to the potential benefits.

4.4 Types of research

The names of the three types of research design describe their purpose very well. The goal of exploratory research is to discover ideas and insights. Descriptive research is usually concerned with describing a population with respect to important variables. Causal research is used to establish cause-and-effect relationships between variables. Experiments are commonly used in causal research designs because they are best suited to determine cause and effect.

Almost all marketing research projects include exploratory and descriptive research. How much of each is necessary depends mostly on how much managers already know about the issue to be studied. When a decision problem has arisen from unplanned changes in the environment, there is usually a need for exploratory research to better understand what is happening and why it is happening. Sometimes, however, managers know a lot about the situation-they understand the key issues and know what questions need to be asked-and the focus quickly shifts to descriptive research that is geared more toward providing answers than generating initial insights. Unlike crime investigations, however, in business situations managers are often perfectly happy with a "most likely" result produced by descriptive research. Only occasionally do they choose to establish cause-and-effect relationships through causal research. As you can probably tell, the three basic research designs can be viewed as stages in a continuous process.

4.4.1 Exploratory research

Exploratory research is conducted to provide a better understanding of a situation. It isn't designed to come up with final answers or decisions. Through exploratory research, researchers hope to produce hypotheses about what is going on in a situation.

A hypothesis is a statement that describes how two or more variables are related. For example, if sales for a particular line of vehicles dropped during the latest quarter, as a researcher you might use exploratory research to provide insights about what caused the decrease in revenue. Suppose that you conducted interviews with potential car buyers and noticed that they seemed to be more excited about the new styles of other car brands than they were about the brand in question. This might lead to the hypothesis that style preferences had changed, resulting in lower sales. You can't really confirm or reject the hypothesis with exploratory research, though. That job is left for descriptive and/or causal research (these are often called quantitative research).

This is an important point, so we'll stress it again: Exploratory research (sometimes referred to as qualitative research) shouldn't be expected to provide answers to the decision problem that you are attempting to solve for a client. It can provide very rich, meaningful information-or even definitive explanations-for particular individuals ("I hate the old-fashioned styling of that car; that's why I won't buy one"), but exploratory research doesn't provide definitive answers for the overall population. There are two reasons for this: (1) Exploratory research usually involves only a relatively small group of people, and (2) these people are almost never randomly selected to participate. In addition to developing one or more hypotheses about actual (or potential) causes of changes in the marketing environment for an organization, exploratory research is also useful for other purposes.

Sometimes it's necessary for helping define the problem, in particular, the research problems that might be addressed. There are often several possible hypotheses about a given marketing phenomenon, and exploratory research can help you identify which research problem(s) ought to be pursued. Exploratory research is

also used to increase a researcher's familiarity with a problem, especially when the researcher doesn't know much about the organization and/or problem to be studied.

Regardless of the particular methods employed, exploratory studies should almost always be relatively small in size. You simply can't afford to devote the bulk of the research budget to exploratory research. And because we often don't know a lot at the beginning of a project, exploratory studies are very flexible with regard to the methods used for gaining insights and developing hypotheses. Basically, anything goes! Although there are a number of common types of exploratory research, be creative and follow your intuition.

Types of Exploratory Research

Some of the more popular methods of exploratory research include literature searches, depth interviews, focus groups, and case analyses.

Literature Search: One of the quickest and least costly ways to discover hypotheses is to conduct a literature search. Almost all marketing research projects should start here. There is an incredible amount of information available in libraries, through online sources, in commercial data bases, and so on. The literature search may involve popular press (newspapers, magazines, etc.), trade literature, academic literature, or published statistics from research firms or governmental agencies such as the U.S. Census Bureau.

Depth Interviews: It's important to start with a good literature search, but at some point you'll probably want to talk to people and ask questions. Depth interviews are used to tap the knowledge and experience of those with information relevant to the problem or opportunity at hand. Anyone with relevant information is a potential candidate for a depth interview, including current customers, members of the target market, executives and managers of the client organization, sales representatives, wholesalers, retailers, and so on. A series of depth interviews can be very expensive. Well-trained interviewers command high salaries; data are collected from one respondent at a time; and, if recorded, audio/video recordings must be transcribed, coded, and analyzed. This technique, however, can yield important insights and more often than not is well worth the

effort.

Focus Groups: Focus group interviews are among the most often used techniques in marketing research. Some would argue that they are among the most overused and misused techniques as well, a point we'll return to later. In a focus group, a small number of individuals (e.g., 8-12) are brought together to talk about some topic of interest to the focus group sponsor. The discussion is directed by a moderator who is in the room with the focus group participants; managers, ad agency representatives, and/or others often watch the session from outside the room via a two-way mirror or video link. The moderator attempts to follow a rough outline of issues while simultaneously having the comments made by each person considered in group discussion. Participants are thus exposed to the ideas of others and can respond to those ideas with their own.

Group interaction is the key aspect that distinguishes focus group interviews from depth interviews, which are conducted with one respondent at a time. It is also the primary advantage of the focus group over most other exploratory techniques. Because of their interactive nature, ideas sometimes drop "out of the blue" during a focus group discussion. In addition, there is a snowballing effect: A comment by one individual can trigger a chain of responses from others. As a result, responses are often more spontaneous and less conventional than they might be in a depth interview.

In general, focus groups are less expensive to conduct than are individual depth interviews, mostly because multiple respondents are handled simultaneously. That's not to say that they are inexpensive, however. By the time the facility has been rented, an experienced moderator has been hired to conduct the session and write the report, and incentives paid to participants, a focus group has become costly. And that's just one focus group; add a series of focus groups and the costs can really rise.

Case Analyses: Often, researchers can learn a lot about a situation by studying carefully selected examples or cases of the phenomenon. This is the essence of case analysis, another form of exploratory research. As a researcher, you might examine existing records, observe the phenomenon as it occurs, conduct

unstructured interviews, or use any one of a variety of other approaches to analyze what is happening in a given situation.

Case analyses can be performed in lots of different ways. Sometimes internal records are reviewed, sometimes individuals are interviewed, and sometimes situations or people are observed carefully. Several years ago, a company decided to improve the productivity of its sales force. A researcher carefully observed several of the company's best salespeople in the field and compared them to several of the worst. It turned out that the best salespeople were checking the stock of retailers and pointing out items on which they were low; the low performers were not taking the time to do this. Without being in the field with the sales force, this insight probably wouldn't have been uncovered.

4.4.2 Descriptive research designs

Descriptive research is very common in business and other aspects of life. In fact, most of the marketing research you've heard about or participated in can be categorized as descriptive research. With a descriptive research design we are usually trying to describe some group of people or other entities. We use descriptive research for the following purposes:

1. **To describe the characteristics of certain groups.** For example, a research group gathered information from individuals who had eaten at a particular barbecue restaurant chain in a Midwestern U.S. city to help managers develop a profile of the "average user" with respect to income, sex, age, and so on. The managers were surprised to learn that about half of their customers were women; they had started with the mistaken belief that a clear majority of their customers were men.

2. To determine the proportion of people who behave in a certain way. We might be interested, for example, in estimating the proportion of people within a specified radius of a proposed shopping complex who currently shop or intend to shop at the centre. And most behavioural data are collected via descriptive research. For example, when a shopper makes a purchase at most retailers, the purchase behaviour is recorded as part of scanner data.

3. To make specific predictions. We might want to predict the level of sales for each of the next five years so that we could plan for the hiring and training of new sales representatives.

4. To determine relationships between variables. It's very common to use descriptive research to examine differences between groups ("awareness levels for our product are higher for men than for women in the target market") or other relationships between variables ("as satisfaction increases, the intention to switch to another service provider decreases").

Descriptive research can be used to accomplish a wide variety of research objectives. However, descriptive data become useful for solving problems only when the process is guided by one or more specific research problems, much thought and effort, and quite often exploratory research to clarify the problem and develop hypotheses. A descriptive study design is very different from an exploratory study design. Exploratory studies are flexible in nature; descriptive studies are not.

Two Types of Descriptive Studies

The basic distinction is between cross- sectional designs, which traditionally have been the most common, and longitudinal designs. Typically, a cross-sectional study involves drawing a sample of elements from the population of interest. Characteristics of the elements, or sample members, are measured only once.

A longitudinal study, on the other hand, involves a panel, which is a fixed sample of elements. The elements may be stores, dealers, individuals, or other entities. The panel, or sample, remains relatively constant through time, although members may be added to replace dropouts or to keep it representative. The sample members in a panel are measured repeatedly over time, in contrast with the one-time measurement in a cross-sectional study.

4.4.3 Causal research designs

Sometimes managers need stronger evidence that a particular action is likely to produce a particular outcome. For example, if you were considering a change in product packaging, you might want to test this hypothesis: "A redesign of the

cereal package so that it is shorter and less likely to tip over will improve consumer attitudes toward the product." For really important decisions, sometimes we need stronger evidence than we can get with descriptive research. (Using descriptive research, we might have learned that there was a negative correlation between consumer ratings of likelihood of tipping over and attitude toward the product, but not a lot more.) Descriptive research is fine for testing hypotheses about relationships between variables, but we need causal designs for testing cause-and-effect relationships.

Concept of Causality

Everyone is familiar with the general notion of causality, the idea that one thing leads to the occurrence of another. The scientific notion of causality is quite complex, however; scientists tell us that it is impossible to prove that one thing causes another. Establishing that variable X causes variable Y requires meeting a number of conditions, one of which (the elimination of all other possible causes of Y) we can never know for certain no matter how carefully we have planned and conducted our research.

Does this mean that researchers shouldn't bother trying to establish causal relationships? Not at all! Although we can't prove with certainty that a change in one variable produces a change in another, we can conduct research that helps us narrow down the likely causal relationship between two variables by eliminating the other possible causes that we are aware of. Causal research designs work toward establishing possible causal relationships through the use of experiments.

Experiments as Causal Research

An experiment can provide more convincing evidence of causal relationships because of the control it gives investigators. In an experiment, a researcher manipulates, or sets the levels of, one or more causal variables (independent variables) to examine the effect on one or more outcome variables (dependent variables) while attempting to account for the effects of all other possible causal variables, usually by holding them constant.

Laboratory experiments allow us to be almost certain that the variables we

manipulate produce the outcomes we observe because we can hold all other factors constant.

A field experiment is a research study conducted in a realistic or natural situation. Just like lab experiments, one or more variables are manipulated to see their effect on an outcome variable. Because it's conducted in the field, you won't have the same degree of control as with a lab study, but you'll attempt to control as much as possible.

Market testing involves the use of a controlled experiment done in a limited but carefully selected section of the marketplace. Market testing is often used to predict the sales or profit outcomes of one or more proposed marketing actions. Very often, the action in question is the marketing of a new product or service. For example, in response to the rising popularity of specialty coffee drinks, McDonald's used test markets to determine that a market existed for McDonald's own higher-end coffee drink before beginning the commercialization process on a larger scale

Test marketing is not restricted to testing the sales potential of new products; it has been used to examine the effectiveness of almost every element of the marketing mix. Market tests have been used to measure the effectiveness of new displays, the responsiveness of sales to shelf-space changes, the impact of changes in retail prices on market shares, the price elasticity of demand for products, the effect of different commercials on sales of products, and the differential effects of price and advertising on demand.

4.5 Process of Research

Before embarking on the details of research methodology and techniques, it seems appropriate to present a brief overview of the research process. Research process consists of series of actions or steps necessary to effectively carry out research and the desired sequencing of these steps. In other words scientific research involves a systematic process that focuses on being objective and gathering a multitude of information for analysis so that the researcher can come to a conclusion. This process is used in all research and evaluation projects, regardless

of the research method (scientific method of inquiry, evaluation research, or action research). The process focuses on testing hunches or ideas in a park and recreation setting through a systematic process. In this process, the study is documented in such a way that another individual can conduct the same study again. This is referred to as replicating the study. Any research done without documenting the study so that others can review the process and results is not an investigation using the scientific research process. The scientific research process is a multiple-step process where the steps are interlinked with the other steps in the process. If changes are made in one step of the process, the researcher must review all the other steps to ensure that the changes are reflected throughout the process.

4.5.1 Steps involved in research process

The following order concerning various steps provides a useful procedural guideline regarding the research process:

1. Formulating the research problem
2. Extensive literature survey
3. Developing the hypothesis;
4. Preparing the research design;
5. Determining sample design;
6. Collecting the data;
7. Execution of the project;
8. Analysis of data;
9. Hypothesis testing;
10. Generalisations and interpretation,
11. Preparation of the report or presentation of the results, i.e., formal write-up of conclusions reached.

A brief description of the above stated steps will be helpful.

1. Formulating the research problem: There are two types of research

problems, viz., those which relate to states of nature and those which relate to relationships between variables. At the very outset the researcher must single out the problem he wants to study, i.e., he must decide the general area of interest or aspect of a subject-matter that he would like to inquire into. Initially the problem may be stated in a broad general way and then the ambiguities, if any, relating to the problem be resolved. Then, the feasibility of a particular solution has to be considered before a working formulation of the problem can be set up. The formulation of a general topic into a specific research problem, thus, constitutes the first step in a scientific enquiry. Essentially two steps are involved in formulating the research problem, viz., understanding the problem thoroughly, and rephrasing the same into meaningful terms from an analytical point of view. The best way of understanding the problem is to discuss it with one's own colleagues or with those having some expertise in the matter. In an academic institution the researcher can seek the help from a guide who is usually an experienced man and has several research problems in mind. Often, the guide puts forth the problem in general terms and it is up to the researcher to narrow it down and phrase the problem in operational terms. In private business units or in governmental organisations, the problem is usually earmarked by the administrative agencies with which the researcher can discuss as to how the problem originally came about and what considerations are involved in its possible solutions. The researcher must at the same time examine all available literature to get himself acquainted with the selected problem. He may review two types of literature-the conceptual literature concerning the concepts and theories, and the empirical literature consisting of studies made earlier which are similar to the one proposed. The basic outcome of this review will be the knowledge as to what data and other materials are available for operational purposes which will enable the researcher to specify his own research problem in a meaningful context. After this the researcher rephrases the problem into analytical or operational terms i.e., to put the problem in as specific terms as possible. This task of formulating, or defining, a research problem is a step of greatest importance in the entire research process. The problem to be investigated must be defined unambiguously for that will help discriminating relevant data from irrelevant ones. Care must; however, be taken to verify the objectivity and validity of the

background facts concerning the problem. Professor W.A. Neiswanger correctly states that the statement of the objective is of basic importance because it determines the data which are to be collected, the characteristics of the data which are relevant, relations which are to be explored, the choice of techniques to be used in these explorations and the form of the final report. If there are certain pertinent terms, the same should be clearly defined along with the task of formulating the problem. In fact, formulation of the problem often follows a sequential pattern where a number of formulations are set up, each formulation more specific than the preceding one, each one phrased in more analytical terms, and each more realistic in terms of the available data and resources.

2. Extensive literature survey: Once the problem is formulated, a brief summary of it should be written down. It is compulsory for a research worker writing a thesis for a Ph.D. degree to write a synopsis of the topic and submit it to the necessary Committee or the Research Board for approval. At this juncture the researcher should undertake extensive literature survey connected with the problem. For this purpose, the abstracting and indexing journals and published or unpublished bibliographies are the first place to go to. Academic journals, conference proceedings, government reports, books etc., must be tapped depending on the nature of the problem. In this process, it should be remembered that one source will lead to another. The earlier studies, if any, which are similar to the study in hand should be carefully studied. A good library will be a great help to the researcher at this stage.

3. Development of working hypotheses: After extensive literature survey, researcher should state in clear terms the working hypothesis or hypotheses. Working hypothesis is tentative assumption made in order to draw out and test its logical or empirical consequences. As such the manner in which research hypotheses are developed is particularly important since they provide the focal point for research. They also affect the manner in which tests must be conducted in the analysis of data and indirectly the quality of data which is required for the analysis. In most types of research, the development of working hypothesis plays an important role. Hypothesis should be very specific and limited to the piece of research in hand because it has to be tested. The role of the hypothesis is to

guide the researcher by delimiting the area of research and to keep him on the right track. It sharpens his thinking and focuses attention on the more important facets of the problem. It also indicates the type of data required and the type of methods of data analysis to be used.

How does one go about developing working hypotheses? The answer is by using the following approach:

(a) Discussions with colleagues and experts about the problem, its origin and the objectives in seeking a solution;

(b) Examination of data and records, if available, concerning the problem for possible trends, peculiarities and other clues;

(c) Review of similar studies in the area or of the studies on similar problems; and

(d) Exploratory personal investigation which involves original field interviews on a limited scale with interested parties and individuals with a view to secure greater insight into the practical aspects of the problem.

Thus, working hypotheses arise as a result of a-priori thinking about the subject, examination of the available data and material including related studies and the counsel of experts and interested parties. Working hypotheses are more useful when stated in precise and clearly defined terms. It may as well be remembered that occasionally we may encounter a problem where we do not need working hypotheses, especially in the case of exploratory or formulative researches which do not aim at testing the hypothesis. But as a general rule, specification of working hypotheses in another basic step of the research process in most research problems.

4. Preparing the research design: The research problem having been formulated in clear cut terms, the researcher will be required to prepare a research design, i.e., he will have to state the conceptual structure within which research would be conducted. The preparation of such a design facilitates research to be as efficient as possible yielding maximal information. In other words, the function of research design is to provide for the collection of relevant evidence with minimal

expenditure of effort, time and money. But how all these can be achieved depends mainly on the research purpose. Research purposes may be grouped into four categories, viz., (i) Exploration, (ii) Description, (iii) Diagnosis, and (iv) Experimentation. A flexible research design which provides opportunity for considering many different aspects of a problem is considered appropriate if the purpose of the research study is that of exploration. But when the purpose happens to be an accurate description of a situation or of an association between variables, the suitable design will be one that minimises bias and maximises the reliability of the data collected and analysed.

There are several research designs, such as, experimental and non-experimental hypothesis testing. Experimental designs can be either informal designs (such as before-and-after without control, after-only with control, before-and-after with control) or formal designs (such as completely randomized design, randomized block design, Latin square design, simple and complex factorial designs), out of which the researcher must select one for his own project.

The preparation of the research design, appropriate for a particular research problem, involves usually the consideration of the following:

(i) The means of obtaining the information;

(ii) The availability and skills of the researcher and his staff (if any);

(iii) Explanation of the way in which selected means of obtaining information will be organised and the reasoning leading to the selection;

(iv) The time available for research; and

(v) The cost factor relating to research, i.e., the finance available for the purpose.

5. Determining sample design: All the items under consideration in any field of inquiry constitute a 'universe' or 'population'. A complete enumeration of all the items in the 'population' is known as a census inquiry. It can be presumed that in such an inquiry when all the items are covered no element of chance is left and highest accuracy is obtained. But in practice this may not be true. Even the slightest element of bias in such an inquiry will get larger and larger as the number of

observations increases. Moreover, there is no way of checking the element of bias or its extent except through a resurvey or use of sample checks. Besides, this type of inquiry involves a great deal of time, money and energy. Not only this, census inquiry is not possible in practice under many circumstances. For instance, blood testing is done only on sample basis. Hence, quite often we select only a few items from the universe for our study purposes. The items so selected constitute what is technically called a sample.

The researcher must decide the way of selecting a sample or what is popularly known as the sample design. In other words, a sample design is a definite plan determined before any data are actually collected for obtaining a sample from a given population. Thus, the plan to select 12 of a city's 200 drugstores in a certain way constitutes a sample design. Samples can be either probability samples or non-probability samples. With probability samples each element has a known probability of being included in the sample but the non-probability samples do not allow the researcher to determine this probability. Probability samples are those based on simple random sampling, systematic sampling, stratified sampling, cluster/area sampling whereas non-probability samples are those based on convenience sampling, judgement sampling and quota sampling techniques. A brief mention of the important sample designs is as follows:

(i) Deliberate sampling: Deliberate sampling is also known as purposive or non-probability sampling. This sampling method involves purposive or deliberate selection of particular units of the universe for constituting a sample which represents the universe. When population elements are selected for inclusion in the sample based on the ease of access, it can be called convenience sampling. If a researcher wishes to secure data from, say, gasoline buyers, he may select a fixed number of petrol stations and may conduct interviews at these stations. This would be an example of convenience sample of gasoline buyers. At times such a procedure may give very biased results particularly when the population is not homogeneous. On the other hand, in judgement sampling the researcher's judgement is used for selecting items which he considers as representative of the population. For example, a judgement sample of college students might be taken to secure reactions to a new method of teaching. Judgement sampling is used

quite frequently in qualitative research where the desire happens to be to develop hypotheses rather than to generalise to larger populations.

(ii) Simple random sampling: This type of sampling is also known as chance sampling or probability sampling where each and every item in the population has an equal chance of inclusion in the sample and each one of the possible samples, in case of finite universe, has the same probability of being selected. For example, if we have to select a sample of 300 items from a universe of 15,000 items, then we can put the names or numbers of all the 15,000 items on slips of paper and conduct a lottery. Using the random number tables is another method of random sampling. To select the sample, each item is assigned a number from 1 to 15,000. Then, 300 five digit random numbers are selected from the table. To do this we select some random starting point and then a systematic pattern is used in proceeding through the table. We might start in the 4th row, second column and proceed down the column to the bottom of the table and then move to the top of the next column to the right.

When a number exceeds the limit of the numbers in the frame, in our case over 15,000, it is simply passed over and the next number selected that does fall within the relevant range. Since the numbers were placed in the table in a completely random fashion, the resulting sample is random. This procedure gives each item an equal probability of being selected. In case of infinite population, the selection of each item in a random sample is controlled by the same probability and that successive selections are independent of one another.

(iii) Systematic sampling: In some instances the most practical way of sampling is to select every 15th name on a list, every 10th house on one side of a street and so on. Sampling of this type is known as systematic sampling. An element of randomness is usually introduced into this kind of sampling by using random numbers to pick up the unit with which to start. This procedure is useful when sampling frame is available in the form of a list. In such a design the selection process starts by picking some random point in the list and then every nth element is selected until the desired number is secured.

(iv) Stratified sampling: If the population from which a sample is to be drawn

does not constitute a homogeneous group, then stratified sampling technique is applied so as to obtain a representative sample. In this technique, the population is stratified into a number of nonoverlapping subpopulations or strata and sample items are selected from each stratum. If the items selected from each stratum is based on simple random sampling the entire procedure, first stratification and then simple random sampling, is known as stratified random sampling.

(v) Quota sampling: In stratified sampling the cost of taking random samples from individual strata is often so expensive that interviewers are simply given quota to be filled from different strata, the actual selection of items for sample being left to the interviewer's judgement. This is called quota sampling. The size of the quota for each stratum is generally proportionate to the size of that stratum in the population. Quota sampling is thus an important form of non-probability sampling. Quota samples generally happen to be judgement samples rather than random samples.

(vi) Cluster sampling and area sampling: Cluster sampling involves grouping the population and then selecting the groups or the clusters rather than individual elements for inclusion in the sample. Suppose some departmental store wishes to sample its credit card holders. It has issued its cards to 15,000 customers. The sample size is to be kept say 450. For cluster sampling this list of 15,000 card holders could be formed into 100 clusters of 150 card holders each. Three clusters might then be selected for the sample randomly. The sample size must often be larger than the simple random sample to ensure the same level of accuracy because is cluster sampling procedural potential for order bias and other sources of error are usually accentuated. The clustering approach can, however, make the sampling procedure relatively easier and increase the efficiency of field work, especially in the case of personal interviews.

Area sampling is quite close to cluster sampling and is often talked about when the total geographical area of interest happens to be big one. Under area sampling we first divide the total area into a number of smaller non-overlapping areas, generally called geographical clusters, then a number of these smaller areas are randomly selected, and all units in these small areas are included in the sample.

Area sampling is especially helpful where we do not have the list of the population concerned. It also makes the field interviewing more efficient since interviewer can do many interviews at each location.

(vii) Multi-stage sampling: This is a further development of the idea of cluster sampling. This technique is meant for big inquiries extending to a considerably large geographical area like an entire country. Under multi-stage sampling the first stage may be to select large primary sampling units such as states, then districts, then towns and finally certain families within towns. If the technique of random-sampling is applied at all stages, the sampling procedure is described as multi-stage random sampling.

(viii) Sequential sampling: This is somewhat a complex sample design where the ultimate size of the sample is not fixed in advance but is determined according to mathematical decisions on the basis of information yielded as survey progresses. This design is usually adopted under acceptance sampling plan in the context of statistical quality control.

In practice, several of the methods of sampling described above may well be used in the same study in which case it can be called mixed sampling. It may be pointed out here that normally one should resort to random sampling so that bias can be eliminated and sampling error can be estimated. But purposive sampling is considered desirable when the universe happens to be small and a known characteristic of it is to be studied intensively. Also, there are conditions under which sample designs other than random sampling may be considered better for reasons like convenience and low costs. The sample design to be used must be decided by the researcher taking into consideration the nature of the inquiry and other related factors.

6. Collecting the data: In dealing with any real life problem it is often found that data at hand are inadequate, and hence, it becomes necessary to collect data that are appropriate. There are several ways of collecting the appropriate data which differ considerably in context of money costs, time and other resources at the disposal of the researcher. Primary data can be collected either through experiment or through survey. If the researcher conducts an experiment, he

observes some quantitative measurements, or the data, with the help of which he examines the truth contained in his hypothesis. But in the case of a survey, data can be collected by any one or more of the following ways:

(i) By observation: This method implies the collection of information by way of investigator's own observation, without interviewing the respondents. The information obtained relates to what is currently happening and is not complicated by either the past behaviour or future intentions or attitudes of respondents. This method is no doubt an expensive method and the information provided by this method is also very limited. As such this method is not suitable in inquiries where large samples are concerned.

(ii) Through personal interview: The investigator follows a rigid procedure and seeks answers to a set of pre-conceived questions through personal interviews. This method of collecting data is usually carried out in a structured way where output depends upon the ability of the interviewer to a large extent.

(iii) Through telephone interviews: This method of collecting information involves contacting the respondents on telephone itself. This is not a very widely used method but it plays an important role in industrial surveys in developed regions, particularly, when the survey has to be accomplished in a very limited time.

(iv) By mailing of questionnaires: The researcher and the respondents do come in contact with each other if this method of survey is adopted. Questionnaires are mailed to the respondents with a request to return after completing the same. It is the most extensively used method in various economic and business surveys. Before applying this method, usually a Pilot Study for testing the questionnaire is conduced which reveals the weaknesses, if any, of the questionnaire? Questionnaire to be used must be prepared very carefully so that it may prove to be effective in collecting the relevant information.

(v) Through schedules: Under this method the enumerators are appointed and given training. They are provided with schedules containing relevant questions. These enumerators go to respondents with these schedules. Data are collected by filling up the schedules by enumerators on the basis of replies given by

respondents. Much depends upon the capability of enumerators so far as this method is concerned. Some occasional field checks on the work of the enumerators may ensure sincere work.

The researcher should select one of these methods of collecting the data taking into consideration the nature of investigation, objective and scope of the inquiry, financial resources, available time and the desired degree of accuracy. Though he should pay attention to all these factors but much depends upon the ability and experience of the researcher.

7. Execution of the project: Execution of the project is a very important step in the research process. If the execution of the project proceeds on correct lines, the data to be collected would be adequate and dependable. The researcher should see that the project is executed in a systematic manner and in time. If the survey is to be conducted by means of structured questionnaires, data can be readily machine-processed. In such a situation, questions as well as the possible answers may be coded. If the data are to be collected through interviewers, arrangements should be made for proper selection and training of the interviewers. The training may be given with the help of instruction manuals which explain clearly the job of the interviewers at each step. Occasional field checks should be made to ensure that the interviewers are doing their assigned job sincerely and efficiently. A careful watch should be kept for unanticipated factors in order to keep the survey as much realistic as possible. This, in other words, means that steps should be taken to ensure that the survey is under statistical control so that the collected information is in accordance with the pre-defined standard of accuracy. If some of the respondents do not cooperate, some suitable methods should be designed to tackle this problem. One method of dealing with the non-response problem is to make a list of the non-respondents and take a small sub-sample of them, and then with the help of experts vigorous efforts can be made for securing response.

8. Analysis of data: After the data have been collected, the researcher turns to the task of analysing them. The analysis of data requires a number of closely related operations such as establishment of categories, the application of these

categories to raw data through coding, tabulation and then drawing statistical inferences. The unwieldy data should necessarily be condensed into a few manageable groups and tables for further analysis. Thus, researcher should classify the raw data into some purposeful and usable categories. Coding operation is usually done at this stage through which the categories of data are transformed into symbols that may be tabulated and counted. Editing is the procedure that improves the quality of the data for coding. With coding the stage is ready for tabulation. Tabulation is a part of the technical procedure wherein the classified data are put in the form of tables. The mechanical devices can be made use of at this juncture. A great deal of data, especially in large inquiries, is tabulated by computers. Computers not only save time but also make it possible to study large number of variables affecting a problem simultaneously.

Analysis work after tabulation is generally based on the computation of various percentages, coefficients, etc., by applying various well defined statistical formulae. In the process of analysis, relationships or differences supporting or conflicting with original or new hypotheses should be subjected to tests of significance to determine with what validity data can be said to indicate any conclusion(s). For instance, if there are two samples of weekly wages, each sample being drawn from factories in different parts of the same city, giving two different mean values, then our problem may be whether the two mean values are significantly different or the difference is just a matter of chance. Through the use of statistical tests we can establish whether such a difference is a real one or is the result of random fluctuations. If the difference happens to be real, the inference will be that the two samples come from different universes and if the difference is due to chance, the conclusion would be that the two samples belong to the same universe. Similarly, the technique of analysis of variance can help us in analysing whether three or more varieties of seeds grown on certain fields yield significantly different results or not. In brief, the researcher can analyse the collected data with the help of various statistical measures.

9. Hypothesis-testing: After analysing the data as stated above, the researcher is in a position to test the hypotheses, if any, he had formulated earlier. Do the facts support the hypotheses or they happen to be contrary? This is the usual

question which should be answered while testing hypotheses. Various tests, such as Chi square test, t-test, F-test, have been developed by statisticians for the purpose. The hypotheses may be tested through the use of one or more of such tests, depending upon the nature and object of research inquiry. Hypothesis-testing will result in either accepting the hypothesis or in rejecting it. If the researcher had no hypotheses to start with, generalisations established on the basis of data may be stated as hypotheses to be tested by subsequent researches in times to come.

10. Generalisations and interpretation: If a hypothesis is tested and upheld several times, it may be possible for the researcher to arrive at generalisation, i.e., to build a theory. As a matter of fact, the real value of research lies in its ability to arrive at certain generalisations. If the researcher had no hypothesis to start with, he might seek to explain his findings on the basis of some theory. It is known as interpretation. The process of interpretation may quite often trigger off new questions which in turn may lead to further researches.

11. Preparation of the report or the thesis: Finally, the researcher has to prepare the report of what has been done by him. Writing of report must be done with great care keeping in view the following:

1. The layout of the report should be as follows: (i) the preliminary pages; (ii) the main text, and (iii) the end matter.

In its *preliminary* pages the report should carry title and date followed by acknowledgements and foreword. Then there should be a table of contents followed by a list of tables and list of graphs and charts, if any, given in the report. The main text of the report should have the following parts:

(a) *Introduction:* It should contain a clear statement of the objective of the research and an explanation of the methodology adopted in accomplishing the research. The scope of the study along with various limitations should as well be stated in this part.

(b) Summary of findings: After introduction there would appear a statement of findings and recommendations in non-technical language. If the findings are extensive, they should be summarised.

(c) Main report: The main body of the report should be presented in logical sequence and broken-down into readily identifiable sections.

(d) Conclusion: Towards the end of the main text, researcher should again put down the results of his research clearly and precisely. In fact, it is the final summing up.

At the end of the report, appendices should be enlisted in respect of all technical data. Bibliography, i.e., list of books, journals, reports, etc., consulted, should also be given in the end. Index should also be given specially in a published research report.

1. Report should be written in a concise and objective style in simple language avoiding vague expressions such as 'it seems,' 'there may be', and the like.
2. Charts and illustrations in the main report should be used only if they present the information more clearly and forcibly.
3. Calculated 'confidence limits' must be mentioned and the various constraints experienced in conducting research operations may as well be stated.

4.5.2 Research design

Design, in general, means organising a structural form of elements of any activity, keeping the purpose in view. Research design "is a plan of the proposed research work." It provides guidelines and directions in research investigations (Ghosh, 1984). A research design is a "Blue Print" for collection, measurement and analysis of data. It outlines how the research will be carried out. It is like glue which sticks together the entire process of research. It provides answers to various questions like - What techniques will be used to gather data. What kind of sampling will be used? How time and cost constraints be dealt with? Etc.

Kerlinger defines research design as "the plan, structure, and strategy of investigation conceived so as to obtain answers to research questions and to control variance. The Plan is the overall scheme or program of research. It prepares an outline of the investigations, formulating the hypothesis and to collect evidences for analysis for testing the hypothesis.

Research design provides a structure before data collection or analysis of data commences. In fact, research is not just a work plan which details what has to be done to complete the project but the work plan will flow from the project's research design.

"The function of a research design is to ensure that the evidence obtained enables to answer the initial question as unambiguously as possible. Obtaining relevant evidence entails specifying the type of evidence needed to answer the research question, to test a theory, to evaluate a program or accurately describe some phenomenon. In other words, when designing research we need to ask: given this research question (or theory), what type of evidence is needed to answer the question (or test the theory) in a convincing way." It is emphasised that research design deals with a logical problem and not a logistical problem." (DeVaus).

De Vaus also insists that research design is different from the method by which data are collected. He says that the way by which data is collected is irrelevant to the logic of the design. Research design is prepared at the commencement of the research project to serve as a blueprint for execution of the research effort.

The purpose of research design is two fold: 1) to provide answers on research as objectively, validly, accurately and economically as possible; and 2) to bring empirical evidence (i.e. derived from or guided by experience or experiment) to bear on the research problem by controlling variance. Controlling the variables refers to collecting evidences on one variable keeping the other variables constant. This method of controlling the variables is likely to give acceptable data for analysis and interpretation.

Composition of a Research Design

A research design generally comprises the following details:

- Statement of the problem of research;
- Specific questions to be answered or hypothesis to be tested;
- Significance of the problem;
- Objectives of research study;

- Assumptions, concepts and their operational definitions or variables;
- Kinds and sources of gathering data;
- Methods of gathering data;
- Data gathering instruments;
- Analysis and interpretations;
- Resources including personnel and budget; and
- Time scheme.

A simple and rather an unrefined example, to illustrate the above research design are given below. If the research effort is to find out the relatively sure method of building up a balanced collection for a research library, the steps are:

- Collection Building and Usage in Research Libraries;
- What methods of collection building would meet users' needs?

A working hypothesis is that the professional staff of the library that has the maximum contact with the users would select items that would circulate most frequently;

- The significance of this investigation is that it would give guidelines in building a most useful collection for the library;
- The objective is to identify the group that makes the best collection building among the three groups who get involved in this work viz. The heads of research project teams, the library professional staff and book vendors and jobbers who supply books on approval;
- The assumption underlying this research effort is that the purpose of a library is to develop the most useful collection possible, where useful is defined in terms of the circulation of the collection. Selection of materials is an independent variable in this project;
- The data to be collected are a selection of material items, their circulation

over a period of time;

- Data to be collected by using library records of circulation in a matrix of tabular statements;
- Use of a computer if data is stored in a computer;
- Appropriate statistical analysis for interpreting the data;
- An appropriate budget to get the work done and optimum staff required; and
- A suitable time frame for the work.

Attributes of a Research Design

It must be noted that the research hinges heavily on the steps 6 to 9 which deal with the collection of data for getting the right evidences and the subsequent operations of assembling them in a logical way, and organising the data for analysis and interpretation.A research design should have the attributes like objectivity, reliability, validity and generalisation to ensure a reasonable quality in the collection of data and recording them.

Objectivity refers to the method of collection of data to obtain accuracy in recording the scores. The measuring instrument should also measure accurately without any subjectivity. Reliability is the attribute of consistency in measurement. "A respondent is expected to give the same response to a particular item every time he is asked about it. In case, a respondent keeps on changing his response, then it would be difficult to decide as to which of the responses should be considered a genuine response.

There are different methods of determining reliability of responses given by a respondent. These include use of (a) check items (b) administering the same test repeatedly and (c) series of parallel forms." (Krishan Kumar, 1999) Validity pertains to the appropriate measuring instrument, it is taken to measure. For instance, a job satisfaction test should only measure job satisfaction and nothing else. There are different procedures adopted for establishing the validity of a test. These include validating the present data against 'concurrent' criterion of a

future principle or a theory etc.

From the application of the above attributes, a research design should use appropriate measuring instruments to yield objective, reliable and valid data. The data analysis should lead to generalisation, which may permit application with reference to a larger group of data and thus lead to some generalisation.

Design types

There are many ways to classify research designs, but sometimes the distinction is artificial and other times different designs are combined. Nonetheless, the list below offers a number of useful distinctions between possible research designs. A research design is an arrangement of conditions or collections.

Descriptive (e.g., case-study, naturalistic observation, survey)

Correlational (e.g., case-control study, observational study)

Semi-experimental (e.g., field experiment, quasi-experiment)

Experimental (experiment with random assignment)

Review (literature review, systematic review)

Meta-analytic (meta-analysis)

Sometimes a distinction is made between "fixed" and "flexible" designs. In some cases, these types coincide with quantitative and qualitative research designs respectively, though this need not be the case. In fixed designs, the design of the study is fixed before the main stage of data collection takes place. Fixed designs are normally theory-driven; otherwise, it is impossible to know in advance which variables need to be controlled and measured. Often, these variables are measured quantitatively. Flexible designs allow for more freedom during the data collection process. One reason for using a flexible research design can be that the variable of interest is not quantitatively measurable, such as culture. In other cases, theory might not be available before one starts the research.

Research design must contain the following aspects.

a. A clear statement of the research problem

b. Procedure and techniques to be used for gathering information from population to be studied.

c. Methods to be used for processing and analyzing data.

d. Means of obtaining the information.

e. The availability and skills of the researcher and his staff (if any)

f. Explanation of the way in which selected means of obtaining information will be organised and the reasoning leading to the selection

g. Time available for research

h. The cost factor relating to research, i.e., the finance available for the purpose.

4.6 Summary

This Unit gives an overview of research methodology that includes all the procedural efforts to conduct a research program. Beginning from identifying a problem of research through extensive studies of the literature, to select a problem for research investigations to the final effort of preparing a blue print for operating the research program, every process of research is described. Wherever possible illustrations are given to explain a particular point. The value of specialising in research methodology not only for taking research problems in library and information science, this exposure would enable information professionals to offer quality information support service to users in general and to researchers in particular. Research today is more and more team research and hence most research projects are operated by specialised institutions or departments of universities or research wings of industrial and business organisations. Those specialising in research methodologies would fit in these projects to gain valuable opportunities to enhance their professional competence and expertise.

4.7 Glossary

Causality: The relation between cause and effect.

Data: Factual information [as measurements or statistics] used as a basis for reasoning, discussion, or calculation

Questionnaire: Structured sets of questions on specified subjects that are used to gather information, attitudes, or opinions

Research: The systematic investigation into and study of materials and sources in order to establish facts and reach new conclusions.

4.8 Self Assessment Questions

I. What do you mean by Research? Explain its significance in modern times.

II. Briefly describe the different steps involved in a research process.

III. What do you mean by research? Explain its significance in modern times.

IV. Describe the different types of research

V. Defined the following terms:

a) Research

b) Research Problems

c) Research Methods

d) Research Techniques

e) g. Research Process

f) h. Research Design

4.9 Lesson End Exercises

Q1. Describe briefly the types of research.

Q2. Discuss in detail the steps of research.

Q3. What are the various applications of research in various functions of management?

Q4. Mention various types of business problems encountered by the researcher.

4.10 Further readings

Booth, W.C, Colomb, G.G. & Williams, J.M, (2008). The craft of Research. Chicago: University of Chicago Press.

Kumar, Ranjit., (2005). Research Methodology; A step by step Guide for Beginners. Newbury Park, CA: Sage Publication.

Cresswell, John, W., (2008). Research Design; Qualitative, Quantitative and Mixed Methods Approaches. Newbury Park, CA: Sage Publication.

Marczyk, G.R, DeMatteo, D. & Festinger D., (2005). Essentials of Research Design and Methodology, New York City, NY : Wiley.

4.11 References

Aaker, D. A., Kumar, V. & Day, G. S Marketing Research, 7th edn, John Wiley, New York.

Baker, M. J. Research for Marketing, Macmillan, London.

Boyd, H., Westfall, R. & Stasch, S. Marketing research: Text and cases. Boston: Irwin.

Bryman, A. Social Research Methods. London: Oxford University Press.

Churchill, G. Marketing research (3rd ed.). Hinsdale, Illinois: Dryden Press.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. Multivariate data analysis (6th ed.). New Jersey: Pearson Education International.

Hair, J. F., Celsi, M., Money, A., Samouel, P., & Page, M. Essentials of business research methods (2nd ed.). Armonk, NY: ME Sharpe.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. Journal of Management, 21(5), 967-988.

Kent, Ray Marketing Reearch: Approaches, methods and applications in Europe. London: Thomson.

Kline, R. B. Principles and practice of structural equation modeling (3rd

ed.). New York: The Guilford Press.

Kothari, C.R. Research Methodology Methods and Techniques, New Age International (P) Limited: New Delhi.

Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. Educational and Psychological Measurement, 30, 607-610

LESSON No. 5

MEASUREMENT AND SAMPLING UNIT-V

MEASUREMENT AND SAMPLING

Structure

5.1 Introduction

In any type of research, measurements play a crucial role towards generation of results. Measurements mostly apply to quantitative research. This kind of research encompasses the collection of data for specific issues. Measurements or scales help in classifying information into distinct categories. As a result, it becomes easier for the researcher to make an appropriate analysis. It is this kind of analysis that leads to rational conclusions. Measurement is the foundation of any scientific investigation. The four main types of scales used in research include ratio, interval, nominal and ordinal levels of measurements. These scales have different characteristics. In this paper, the different scales of measurements will be analyzed. Additionally, the specific role played by these scales in research will also be discussed.

Moreover, Sampling is one of the very important aspects from research perspective or in marketing research. From a general perspective, sampling involves selecting a relatively small number of elements (characteristics) from a larger defined group of elements and expecting that the information gathered from the small group of elements will provide accurate judgement about the lager group. Sampling is simply the process of learning about the population on the basis of a sample drawn from it. Thus in the sampling technique instead of every unit of the universe only a part of the universe is studied and the conclusions are drawn on that basis for the entire universe. Sampling is used in decision making almost every time. For example, before buying a book we flick through few pages and decide whether it suits our reading preferences. For a complex buy such as mobile phone, we first decide several features as essential and other as desirable. Then we decide on the brand and select the mobile phone on the brand, price of the product and several other such variables. While making the final decision making there are many such variables which we don't take into consideration. In a way, we use few elements (characteristics) of mobile phones and expect that they will cover most of what we desire. We use sampling when selecting a job, choosing a restaurant and even selecting a TV channels. As we consumers use sampling in our regular decision making process, practitioners or managers can also benefit by understanding sampling process in providing better

matched products with our needs. Furthermore, sampling provides several other benefits. For instance, as not every consumer of the products is being studied, the total cost of the research can be lowered with the use of sampling. A sample would require fewer fieldworkers. Hence, better personnel could be selected and trained and their work could be closely supervised. It is observed that the lesser administrative problems encountered in collecting data from the sample lead to more accurate data than could be obtained by collecting data from all the units. Therefore, it is very important for a research perspective to understand the concept of sampling and its concepts.

5.5.1 Concept of Measurement and Scale

Measurement is the process of describing some property of a phenomenon of interest, usually by assigning numbers in a reliable and valid way. The numbers convey information about the property being measured. When numbers are used, the researcher must have a rule for assigning a number to an observation in a way that provides an accurate description. Measurement is a generalized idea about a class of objects, attributes, occurrences, or processes. Everything we do begins with the measurement of whatever it is we want to study

Definition: measurement is the assignment of numbers to objects Example: When we use a personality test such as the EPQ (Eysenck Personality Questionnaire) to obtain a measure of Extraversion - 'how outgoing someone is' we are measuring that personality characteristic by assigning a number (a score on the test) to an object (a person).

Operational definition of measurement specifies what the researcher must do to measure the concept under investigation.

Measurement is used regularly in our daily lives. For example, if someone asks you of your favourite newspaper, your mind may create a list and you shall decide your favourite most newspaper from that. While deciding on that favourite newspaper your mind would have used several criteria such as your reading pattern, content of the newspaper, various other features such as writers involved, format, colour and pictures used, and columnists you prefer. Furthermore, your mind would have also told you the most preferred the second most preferred

and even least preferred newspaper. The criteria your mind is using in deciding the favourite newspaper is called measurement. In research terms, measurement is nothing but the assignment of numbers or other symbols to characteristics of objects according to certain pre-specified rules. One of the important things to note here is that researchers do not measure objects but some characteristics of it. So in reality, researchers do not measure consumers but their perceptions, beliefs, attitudes, preferences and so on. The idea of assigning numbers can be helpful in two ways in accurate understanding of a phenomenon; (1) it allows statistical testing and (2) it helps facilitate easier communication as people have a clear idea with regard to what 10% or 20% means worldwide. Furthermore, numbers also provide objectivity in understanding a phenomenon. This added accuracy due to numbers is essential to effective decision making.

The word scale or scaling is generally used for measuring something. It is in fact a device through which we measure various things. It is easy to apply scales in the field of physical science for measurement of physical phenomena. For example, for measuring the fluctuations of the weather, we use barometer. Thermometer is used for measurement of heat. Tapes, meters and other yard sticks containing millimeters, centimeters, meters, etc. are used. It means that different measurement or devices are available for measuring different aspects of physical phenomena but measurement of business and economic phenomena is not an easy task. Many aspects of industrial activity are not apparently discernible. It is not possible to measure emotions, attitudes, faiths, values etc., through the devices that are used for measuring the physical data. In business activity, there are two types of variables or factors that are responsible for change of environment. Some of these factors or variables can be measured through scales that are used for measuring various aspects of physical phenomena. For measuring them, they are converted into quantitative data and after that they are subjected to scaling or measurement.

Scaling defined: In the field of business research, measurement or scaling implies in conversion of the characteristics or qualitative data into quantitative data. After this conversion the scaling is done. This has to be done because qualitative data or measurement are mostly subjective and differ from investigator to investigator.

Unless it is converted into quantitative data, the measurement would not be possible. Various kinds of statistical measurements are used for conversion of qualitative one and measuring various aspects of business and industrial activity.

Significance of Scaling

In business research studies, scales of measurement are very much important to make the studies exact and scientific. Though it is difficult to develop scales of measurement in business and economics, it does not make their utility less in any way. The business and industrial activity is complex, abstract and qualitative. While this makes measurement difficult, it increases the utility and importance of measurement scales. For a scientific study of business problems the importance of scales of measurement will be clear by following.

1. To Make Business Research Study Scientific: Pointing out the utility of scales of measurement in socio-economic investigation, W.J. Goode and P.K. Hatt have said, "All sciences move in the direction of greater precision. This makes many forms, but one fundamental form is measuring gradation". Thus business phenomena can be measured by means of socio-economic scales. Such scales of measurement unravel new sources of business information.

2. Objective Measurement: Measurement techniques help in maintaining objectivity of study. In business research as more and more scales of measurement are being developed, objectivity of study is increasing. Even if several factors contribute in hindering such objectivity, it has been achieved to such a limit that useful prediction has been possible.

Types of Scales: There are two types of scales.

a) **Concerning behaviour and personality -** These scales include:

i) Attitude Scales

ii) Moral Scales

iii) Characteristic Scales

iv) Social participation Scales

v) Psychoneurotic Inventories

b) Concerning socio-economic environment: These are another types of scales which are used to measure certain other aspects of the socio-economics environment. Some of these scales are used to study socio- economic status, communities, housing conditions and institutional framework.

Basis for Scale Classification

The scaling procedures may be broadly classified on one or more of the following basis:

a) **Scaling based on Subject Orientation:** A scale may be designed to measure the attributes of the respondent through stimulus presented to him. In such a scale the respondent's characteristic has been measured depending on his orientation to subject stimulus.

b) Scaling based on Response Form: The scaling procedure may be classified on the basis of response form whether categorical and comparative. Categorical scales are nothing but rating scale. Such scales are used when a respondent rates some object without direct reference to other objects. Under the comparative scales respondent is asked to compare several objects and respondent expected to rank that one object is superior to the other. Therefore, this scale is also known as ranking scale.

c) Scaling based on Subjectivity: There may a situation when the scale data is based on subjective personal preferences. Or simply make non preference judgments. For example, in former case, the respondent is asked to choose which food he prefers or which method he would like to be implemented, whereas in the latter case, respondent is simply asked judge which for is more delicious or which method will be economical.

d) Scaling based on properties: If scale has design considering the mathematical properties, one may classify the scale as nominal, ordinal, interval and ratio scales. These scales are discussed in detail:

1) Nominal Scale: A system in which numerical values are assigned to object in order to label them is called nominal scale. One ubiquitous example of nominal scale is the numbers on the jersey of football players which helps spectators to

identify them. Such a scale is a very convenient method to keep a track of objects, events and people. However, by using such a scale, not much statistical analysis can be achieved. All one can do is to use mode as a measure of central tendency. There is no generic measure of dispersion in case of nominal scales. Chi-square test is the most common test of statistical significance that can be utilized. For the measurement of correlation the contingency coefficient can be calculated.

Advantages

i) It is a simple scaling technique.

ii) It describes differences between objects as assigning them to categories

iii) They are widely used in surveys when data are being classified by major grouping.

Limitations

i) It has least powerful measure of scaled date

ii) It can not be statistically analyzed

iii) It has limited use in research.

ii) Ordinal Scale: The ordinal scale, as the name suggests, attempts to place objects or events in a particular order. In this scale, there is no attempt is made to have equal intervals of the scale. One popular example of such scale of this is the rank ordering utilized in qualitative business research. In fact, an ordinal scale can be depicted as a statement of greater than (>) or less than (<) without the ability to show how much greater or lesser. In the ordinal scale, equality (=) statement is also a possibility. In so far statistical measures are concerned, the measure of central tendency used with the ordinal scale is median. Dispersion can be measured by using either the percentile or quartile method. For measurement of statistical significance, non paramatric method is utilized.

Major limitation of ordinal scale is that they only permit the ranking of objects from lowest to highest. It must be noted that ordinal scales have no absolute value, and therefore absolute difference between adjacent ranks may not be equal.

iii) Interval Scale: In development of interval scale, the intervals are adjusted

in terms of some formula that has been evolved as a basis for making the intervals equal. The centigrade temperature scale is an example of an interval scale and can be used as an example of possibilities it provides. It can be said that an increase in temperature from 800c to 900C is same increase in temperature as an increase from 300c to 400c. However it can not be said that the temperature of 800c is twice as hot as the temperature of 400c because both numbers are dependent on fact that the zero on the scale is arbitrary and fixed as the temperature of the freezing point of water.

Interval scales provide more powerful measurement than ordinal scale because in interval scale the concept of equality of interval is inherent.

Mean is the appropriate measure of central tendency, while standard deviation is the most widely used measure of dispersion. Tests for statistical significance used in case of interval scales are the test and of "F" test. Correlation techniques utilized are product moments.

iv) Ratio Scale: Ratio scales have an absolute or true zero of measurement. For example, the zero point on an inch-tape indicates the complete absence of length or height. scales. In fact all mathematical operations possible with real numbers are possible with ratio scales. In case of ratio scales measures of central tendency is computed by calculating geometric and harmonic means.

With ratio scales it is possible to make statement such as, Rakesh's athletic prowess is twice as good as that of Mohans. Here the ratio involved thus facilitate a comparison which was not passable with Interval scales. Thus it can be concluded that starting from normal scale and progressing to ratio scales the quantum of information increases progressively.

e) Scaling based on number of dimensions: The scales may be unidimensional or multidimensional scales. Under unidimensional scale we measure only one attribute of the respondent whereas in later the object might be better described taking measurement of his attributes from multiple dimension.

In fact, ratio scale represents the actual quantum of variables. Scales evolved for measuring physical dimensions such as weight, height distance etc. are the examples. Large number of quantitative ratio scales techniques are used

to manipulate ratio

5.2 Objectives

After studying this chapter students should be able to understand:

- Delineate the concept of measurement and scale.
- Outline sampling, sample size, and sampling procedure.
- Demarcate the various types of probability and non probability sampling techniques.
- Understand and apply various methods of data collection.
- Mark out the steps required for the preparation of questionnaire.
- Illustrate the precautions while preparing a questionnaire.

5.3 Concept of Sample

A sample is that part of the universe which we select for the purpose of investigation. A sample exhibits the characteristics of the universe. The word sample literally means small universe. For example, suppose the microchips produced in a factory are to be tested. The aggregate of all such items is universe, but it is not possible to test every item. So in such a case, a part of the universe is taken and then tested. Now this quantity extracted for testing is known as sample.

Characteristics of Ideal Sample

A good sample has following qualities:

1. Representativeness: An ideal sample must be such that it represents adequately the whole populations. We would select those units which have the same set of qualities and features as are found in the whole data. It should not lack in any characteristic of the population.

2. Independence: The second feature of a sample is independence, that is interchangeability of units. Every unit should be available to be included in the sample.

3. Adequacy: The number of units included in a sample should be sufficient to

enable derivation of conclusions applicable to the whole population. A sample having 10% of the whole population is generally adequate.

4. Homogeneity: The units included in sample must bear likeness with other units, otherwise sample will be unscientific.

Reliability of the Sampling

As is evident from our discussions so far, sample should be reliable and free from bias. Then only it shall be possible for us to arrive at dependable results. It means that the size of sample, its relevance and stability to the problem, its representative character, its use for the study, etc. are the factors that determine its reliability. Reliability of samples may be tested on the following parameters.

1. Size of the sample: The size of the sample very much determines not only its representativeness but also its utility for study. The researcher must test that the size is adequate for scientific yet convenient study of the problem.

2. Representativeness of the sample: The representativeness of the sample should also be tested. It means that the sample selected should be representative and possess the characteristics of other units.

3. Parallel sampling: It means that apart from the sample that have been drawn, another sample may be drawn from the same universe for testing. On the basis of these tests, the reliability of the sample, initially selected may be tested.

4. Homogeneity of the samples: Samples, in order to be useful for study, should be homogeneous. This again means that they should possess all the characteristics that are present in the universe.

5. Unbiased selection: The selection of the sample should be done through a method which is free from bias and prejudices. For this, the researcher has to take the precautions.

Sampling: Basic Constructs

Universe/Population

From statistical analysis point of view, we often refer to two terms namely, universe and population. The term 'universe' implies the total of all the items or units of

analysis in the field of proposed research. On the other hand, term "population' implies the total number of items for which information is sought. Or in research terms, 'population' is defined as the totality of cases that confirm to some designated specifications. In research parlance, the units of research which possess the attributes required to achieve the research objectives are known as elementary units. The aggregation of such elementary unit is termed as population. Therefore, it is implied that all units of any research project will constitute universe, and all the elementary units (defined on the basis of research objective attributes) constitute population.

Universe - Finite and Infinite: If the number of units or items or members in a universe is definite and fixed it is a finite universe. For example, number of feature films produced in the world is finite and can be known. The number of districts in India is another example of finite universe. If the number of units or items is not fixed and finite in a universe it is known as Infinite universe. The examples are the number of stars in the sky, water in the sea etc.

Universe - Existent and Hypothetical: An existent universe is one which comprises of tangible items or units, the number of books in a library, the number of companies paying corporate tax. A hypothetical universe is one which is not existent. It is imaginary. The examples of hypothetical universe are the moves on a chess board. All possible and conceivable moves of chess can never be realized. They can only be imagined.

- **Sampling Unit**

Sampling units are the target population elements available for selection during the sampling process.

- **Sampling Frame**

The term, sampling frame or population frame refers to the listing of all items in the population with proper identification under study. A sampling frame is a representation of the elements of the target population. It consists of a list or set of directions for identifying the target population. Some common sources of sampling frame are the lists of voters, commercial directories, telephone directories, or even maps. Many commercial organisations provide a database

consisting of names, address, and telephone numbers of potential sampling frame for various studies. For example, if we want to find out the capital invested and number of workers working in small scale industries in Delhi, we must have a complete list of names and addresses of all the small scale firms. The list of names and addresses will be called sampling frame. Regardless of the sources, it is very difficult and expensive to obtain truly accurate or representative sampling frames.

- **Sampling Design**

A sampling design is a definite plan for obtaining a sample from the sampling frame. It refers to the technique or procedure of selecting some sampling units from which inferences about the population are drawn.

- **Statistics and Parameters**

A statistic is characteristic of a sample and a parameter is the characteristic of universe or population.

- **Sampling Errors**

In sample survey only a small part of the universe or population is studies, as such there is every possibility that its result would differ from each other. These differences constitute the errors due to sampling and are known as sampling errors.

- **Precision**

Precision is the range within which the population average for other parameter will lie in accordance with the reliability specified in the confidence level as percentage of the estimate ± or as a numerical quantity. For example, if the estimate is Rs. 5000 and the precision desired is ± 4% then the true value will be no less than Rs. 4800 and no more than Rs. 5200.

- **Confidence Level and Significance Level**

The confidence level which is also termed as, reliability is expressed in terms of percentage of times that an actual value will fall within the prescribed precision limit. For example, if we have to consider a confidence level of 90% which will

imply that if we repeat a particular exercise 100 times, 90 times the parameters of population under consideration will lie within the prescribed limit. The significant level, on the other hand, indicates the likelihood of the observation falling outside the prescribed range. Hence, if the confidence level (CL) is 90%, then significance level (SL) would be 10%. If the confidence level is 98%, then significance level is 2%. Therefore, significance level can mathematically be expressed as, SL % = 100% - CL%.

- **Sampling Distribution**

If we take certain number of samples and for each sample compute various statistical measures such as mean, standard deviation etc., then we can find out that each sample may give its own value for statistics under consideration. All such values of a particular statics, say, mean together with their relative frequencies will constitute the sampling distribution of mean standard deviation etc.

- **Theory of Sampling**

The theory of sampling emanates from the study of relationships existing in a population, compared to those in sample drawn from that particular population. In a sense, this theory is developed with a purpose of estimating the properties of a population from those obtained by studying a specific sample. It also deals with finding out the preciseness of the estimate. It is important to have random sample for this theory to be valid. The methodology of drawing inferences of the universe from random sampling is known as theory of sampling. Sampling theory deals with:

a) Statistical estimation: Sampling theory helps in estimating unknown population parameters from a knowledge of statistical measures based on sample studies. The estimation can be a point estimate or it may be an interval estimate. Point estimates is a single estimate in the form of a single figure but interval estimate has two limits viz., the upper limit and lower limit within which the parameter value may lie.

b) Testing of hypothesis: The second objective of the sampling theory is to accept or reject the hypothesis. One great help that sampling theory provides is

in observance of differences in estimation are due to chance and if they are of any real significance.

c) Statistical Inferences: Sampling theory uses statistical underpinnings to make a generalized statement about a specific population extended from the studies conducted on sample obtained from it. This is what is known as, statistical inferences about a particular population and accuracy thereof.

- **Sampling and Non Sampling Errors**

To use surveys based on samples it is essentials that one appreciate the concept of sampling and non-sampling errors. Sampling errors arises out of the fact that inferences for the entire population is drawn on the basis of few sample observations. On the other hand non-sampling errors happens due to errors of computation at the sate of classification and processing of data.

- **Sampling Errors**

Sampling errors are generally of two types viz biased and un-biased.

1. Biased Errors: The process of selection and estimation of samples may have some bias which leads to these errors. For instance if instead of using simple random sampling, judgment sampling is used in a research survey some bias is introduced in the result due to judgment of the researcher in selecting the judgement sample. Such errors are biased sampling errors.

2. Unbiased Errors: Causes of these errors is due to chance disagreement between the population units selected in the sample and those not selected. An error in final result is due to the fact of difference in the unit.

Causes of Bias: Bias may arise due to following reasons:

a) Faulty selection of sample

b) Substitution

All of the above factors can lead to bias in the representative nature of the sample. Bias can be avoided naturally by eliminating the above sources of bias. The easiest and definite way of avoiding bias in the sampling selection process is by selecting the sample absolutely randomly. Generally sampling errors

are reduced by increasing the sample size.

- **Non-Sampling Errors**

Non-sampling errors typically arise due to following factors:

a) Inadequate and inconsistent data specification.

b) Faulty method of interviews and observations.

c) Lack of experience and training of investigators.

d) Errors in data processing operations.

e) Errors in tabulation and classification of data.

The above list is not comprehensive but does give indication of possible sources of errors. Non-sampling errors can generally be reduced by controlling the factors which were listed above. However, it must be noted that converse phenomena happens, which is through sampling errors decrease by increasing sample size, non-sampling errors increase with increase in sample size. Therefore, size of sample should be optimum so as to minimize a sum of sampling and non-sampling errors.

5.3.2 Sample Size and Sampling Procedures

Sample size is the essential aspect that must be considered by researchers. For proper study of the problem, it is necessary to have proper sampling. It means that the sample should be of proper size. If the sample is either too small or too big, it shall make the study difficult. What should be the size of the sample, is a question which should be answered only after taking into account the various factors of research problem at hand.

Factors to be considered in Sample Size

The following factors should be considered while deciding the sample size:

i) The size of the universe: The large the size of the universe, the bigger should be the sample size.

ii) The resources available: If the resources available are vast, a large sample size could be taken. However, in most cases resources constitute a big constraint

on sample size.

iii) The degree of accuracy or precision desired: The greater the degree of accuracy desired the larger should be the sample size. However, it does not necessarily mean that bigger samples always ensure greater accuracy.

iv) Homogeneity or heterogeneity of the Universe: If the universe consists of homogeneous units, a small sample may serve the purpose but if the universe consists of heterogeneous units, a large sample may be required.

v) Nature of study: For an intensive and continuous study a small sample may be suitable. But for studies which are not likely to be repeated and are quite extensive in nature, it may be necessary to take larger sample size.

vi) Method of sampling adopted: The size of samples is also influenced by the type of sampling plan adopted. For example, if the sample is a simple random sample it may necessitate a bigger sample size. However, in a properly drawn stratified sampling plan, even a small sample may give better results.

vii) Nature of respondents: Where it is expected a large number of respondents will not co-operate and send back the questionnaires, a larger sample should be selected.

The above factors have to be properly weighted before arriving at the sample size. However, the selection of optimum sample size is not the simple as it might seem to be.

Determination of Sample Size

An important question that will arise in any research study using sampling is, what should be the size of the sample? In determining the sample size, either one or both of the following considerations are taken into account:

1. Cost and time,

2. Permissible error in estimation.

Usually in many sampling surveys, cost and time are considered in arriving at a suitable sample size. Statistical methods can be employed in determining the

sample size for probability sampling designs, if the latter consideration is taken into account. The basic principle behind these methods can briefly be stated as follows:

Through sampling, we are getting an estimate for a parameter of the population (or universe). It is rational on our part to expect that the estimate to be as close as possible to the value of (unknown) parameter. This factor is to be considered in determining the actual sample size. But, the basic question is, that how much close the estimate should be to the parameter. One ideally wishes that the estimate should exactly be the same as the parameter. This is physically impossible in almost all cases, since it involves considering the whole population i.e., the sample is the population itself. Hence in determining the actual sample size. But, the basic question is, that how much close the estimate should be to the parameter. One ideally wishes that the estimate should exactly be the same as the parameter. This is physically impossible in almost all cases, since it involves considering the whole population, i.e., the sample is the population itself. Hence in determining the sample size, the researcher should fix up the margin of error, like his estimate should be within plus or minus 5 units of the value of the parameter. But the value of parameter is not known. Hence, one cannot be sure whether the given size of the sample results into the required amount of accuracy. However, one can make a probability statement like: there should be a chance (probability) of 95% that the estimate should be within plus or minus 5 units of the parameter.

In some design it is possible to incorporate the cost aspect also. In practice, in research studies a number of parameters are estimated. In such cases, for determining the sample size, one should consider a key parameter among those parameters.

A number of formulae have been devised for determining the sample size depending upon the availability of information. A formula is given below:

$$n = (z.\ ?/d)2$$

Where n = Sample size

Z = Value at a specified level of confidence or desired degree of precision.

? = Standard deviation of the population

d = Difference between population mean and sample mean.

The steps in computing the sample from the above formula are:

i) Select the desired degree of precision, i.e., specified level of confidence and designate it as small, 'z' (at 1% level of significance or 99% confidence level the value of, 'z' is 2.58, and at 5% level of significance or 95% confidence level 1.96).

ii) Multiply the, 'z' selected in step 1 by the standard deviation of the universe which may be assumed.

iii) Divide the product of the proceeding step by the difference between population and same mean. Square the resultant quotient. The result is the size of sample required.

Solved Example

Determine the sample size if standard deviation of population is 8, population mean = 30, sample mean = 28 and the confidence level of 99%

Solution

$$n = (z.?/d)2$$

? = 8, d = (30-28) = 2

z = 2.576 (at 1% level the z value is 2.576)

Substituting the values:

$$n = 2.567/2 = 106$$

Similarly, there has been substantial debate over what represents a satisfactory sample size with no simple and definitive rule to define a suitable sample size. Different researchers have recommended diverse sample sizes as appropriate. The sample size can be determined by the reflection of the following scholar's contentions for example, Hinkin (1995) recommended ratios of

statements to responses from 1:4 to 1:10. Similarly, Hair et al. (2006) suggested that every item requires minimum 5 respondents and maximum 10 respondents. Similarly researchers such as Kline (2011) argue that ratios of free model parameters estimated to responses from 1:5 to 1:10 is needed.

Sampling Procedures: Sampling is a process or technique of choosing a sub-group from a population to participate in the study; it is the process of selecting a number of individuals for a study in such a way that the individuals selected represent the large group from which they were selected. There are two major sampling procedures in research. These include probability and non probability sampling. These are discussed in section 5.3.3.

5.3.2 Types of Sampling Techniques - Probability and Non-Probability Sampling

Probability Sampling Method

The units which constitute a probability sample are selected randomly, with each unit having a known chance of selection. Thus, before a probability sample can be drawn, the project will need to define a sampling 'frame' for the population. Such a frame will need to ensure that each unit is included only once, and that no unit is excluded; thus, all units have an equal chance of selection. The frame should cover the entire population and be convenient to use. A probability sample should attempt to be representative of the entire population, but it can never be an exact replica. However, by applying the rule of probability, generalizations concerning the population may be made and calculations made about the degree of confidence with which the results can be viewed. Sample error, for probability samples, stems from the variability of the sample and/or the size of that sample. Types of Probability Sample are simple random sampling (SRS), stratified sampling and systematic Sampling

Advantages of probability sampling: The following are the advantages of this type of sampling methodology.

i) Probability sampling does not require detailed information about the universe, to be effective.

ii) Probability sampling provides estimates which can be measured precisely and are inherently unbiased.

iii) It is possible to evaluate the relative efficiency of various sample designs when probability sampling is used.

Limitations of probability sampling: The limitations of this method are:

a) It requires a high degree of skill level and expertise.

b) It requires a lot time to plan and determine a probability sample.

c) The costs involved in probability sampling are higher as compared to non-probability sampling.

Non-Probability Sampling Method

The researcher does not know the chances of a unit's selection if non-probability sampling techniques are employed. Therefore, the ability to generalize about a population, using the laws of probability, is much reduced and it is not possible to calculate the degree of confidence in the results. The sample is chosen at the convenience of the consultant or to fulfil the demands of some predetermined purpose. Non-random sampling is process of sample selection without the use of randomization. In other words, a non-random sample is selected on a basis other than probability consideration such as convenience, judgment etc. Non-probability sampling methods are judgment or purposive or deliberate sampling, convenience sampling, quota sampling.

Characteristics of Non-probability Sampling:

The following are the main characteristics of non-probability sample:

i). There is no idea of population in non-probability sampling.

ii). There is no probability of selecting any individual.

iii). Non-probability sample has free distribution.

iv). The observations of non-probability sample are not used for generalization purpose.

v). Non-parametric or non-inferential statistics are used in non probability sample.

vi). There is no risk for drawing conclusions from non-probability sample.

Simple Random Sampling (SRS)

Sampling procedure is known as simple random sampling where the individual units constituting the sample are selected at random. Random sampling method of selection assures each individual element or units in universe and equal chance of being chosen. In other words, if in a sample size of n, all the possible combinations of n elementary items have the same probability of being included; it is called simple random sampling. Individual units are assigned a number, a sample of these numbers is then selected either by using a 'lottery' system, or by the use of random number tables. The method is simple to use and it obeys the laws of probability; however, it may produce samples which are not representative of the population.

The following precautions should be taken in random sampling:

a) The universe or the population to be sampled should be clearly defined.

b) List of all the units that are available for the purpose of selection, should be prepared.

c) Units to be selected that are ready for the purpose of selection should be approximately of equal size.

d) These units should be independent and not dependent upon one another. It means that if one unit is selected, it should not be necessary to select another unit for the sake of complete information.

e) Every unit should be accessible.

f) Replacement of the selected unit should not be done

Simple random sampling is of two types: a) Sampling with equal probability

i) with replacement, and

ii) without replacement.

b) Sampling with probability proportional to size of the sample unit.

Selecting a Random Sampling: A random sample can generally be selected in

the following four methods:

i) Lottery Method

ii) Tippet's number of method

iii) Selection from sequential method

iv) Grid system

i) **Lottery Method:** In this method, a lottery is drawn by writing the numbers or the name of various units and putting them in a container. They are thoroughly mixed and certain numbers are picked up from the container, and those which are picked up are taken up for

sampling.

ii) **Tippet's Number Method:** It is called Tippet's numbers method because it was evolved by L.H.C. Tippet who constructed a list of 10,400 four-digit numbers written at random. From these numbers it is not very difficult to draw samples at random. For example, if 50 persons are to be selected for study out of the total number of 500, then we can open any page of Tippet's numbers and select first 50 random numbers that are below the value of 500 and take them up for study. On the basis of the experiments carried out through this technique, it has been found that the results that are drawn on the basis of this method or random sampling, are quite reliable.

iii) Selection from Sequential List: In this method, the names are arranged serially according to a particular order. The order may be alphabetical, geographical or only serial. Then out of the list any number may be taken up. Beginning of selection may be made from anywhere. For example, if we want to select 10 persons, we can start right from the 10th and select 10, 20, 30, 40 and so on.

iv) Grid System: This method is generally used for selecting the sample of an area and so on in this method, a map of the entire area is drawn. After that a screen with squares is placed upon the map and some of the squares are selected at random. Then screen is placed upon the map and the area falling within the

selected squares are taken as samples. This is today possible by electronic means.

Advantage of simple random sampling method: The method of random sampling is simple and said to be a very easy form of the method employed for the study of business research problems. In this method, the investigator can keep himself away from prejudices, bias and other elements of subjectivity. It has the following advantages:

a) It is quite simple, and follows mathematical procedures.

b) It is free from bias and prejudices.

c) It is said to be more representative because in this method, each unit has equal chance of being selected.

d) In this method if some error has crept in, it will not be difficult to detect in case the sampling has been done strictly according to random sampling.

Disadvantage of simple random sampling method: Although the random sampling method has some advantages, it suffers from certain demerits as well. These demerits are:

a) Selection according to strictly random basis is not possible: Sometimes instead of random sampling, we resort to substitution of selected sample which vitiates the whole procedure.

b) Lack of control of research: In this method, the researcher has no control over the selection of the unit. Therefore, the units selected may be widely dispersed and possibility of contact may become a problem.

c) Random sampling does not suit heterogeneous groups: Random sampling is a useful method if all the units are heterogeneous, in case units are heterogeneous in nature, the random sampling method may not be very useful.

Stratified Sampling

This method accepts the variability of the population and, by stratifying it before the sample is taken, tries to reduce its potential unrepresentativeness. Stratified sampling is distinguished by the two-step procedures it involves. In the first step the population is divided into mutually exclusive and collectively exhaustive sub-

populations, which are called strata. In the second step, a simple random sample of elements is chosen independently from each group or strata. This technique is used when there is considerable diversity among the population elements. The major aim of this technique is to reduce cost without lose in precision. Stratified random sampling adopts the position that each group/stratum is a population in its own right and then extracts a sample, by simple random means, from each of them. There are two main types of stratified random sampling (a) proportionate stratified sampling and (b) disproportionate stratified sampling discussed below.

Types of stratified sampling:

a) Proportionately stratified sampling: In this kind of stratified sampling the number of units should be drawn from each strata in proportion to size of the strata. In other words, in proportionate stratified sampling, the size of each sub-sample taken from a particular stratum is proportionate to the size of that stratum in the population. Thus, if 25 per cent of the population is aged between 35 and 45, then 25 per cent of the sample should be composed of people in that age group.

b) Disproportionate stratified sampling: In this type of stratified sampling an equal number of cases are taken from each stratum without any consideration to the size of strata in proportion to universe. In other words, in disproportionate stratified sampling the proportion of a characteristic as possessed by the population, is not reflected to the exact extent in the size of the sub-sample. Such a deliberate 'distortion' of the size of the sub-sample may improve the quality of the data if certain strata have an unusually large influence in the situation under investigation and need to be given a more significant role. Here, not every unit has an equal chance of selection, but the chance of selection is still known, thus the laws of probability still rule and appropriate weighting (s) can be used when calculating the results.

c) Stratified weight sampling: In this method, an equal number of units are selected from each stratum and averages are drawn, but in doing so they are given weight in proportion to the size of the stratum in relation characteristics."

Importance of Strata: The method very much depends upon the process of

stratification. If correct stratification is done then the method shall be successfully employed. Stratification should be done with following kept in mind:

a) Each stratum in the universe should be large enough in size so that selection of items may be done on random basis.

b) There should be a perfect homogeneity among different units of stratum.

c) The ratio of number of items to be selected from each unit of strata should be the same as the total number of the units in the strata bearing the units of the entire universe.

d) Stratification should be well-defined and clear-cut. It means that each unit stratum should be free from influence of the other.

Merits of the stratified sampling method: The merits of stratified sampling are described below:

a) Greater control of the investigator: In this method, the investigator has greater control over the selection of the samples. In random sampling although every group has a chance of being selected in the sample but there is every possibility but it can so happen that certain important groups are left unrepresented, but in stratified sampling no important group is likely to be left out.

b) Easy to achieve representative character: In this method it is possible to achieve representative character with fewer items. In case of homogeneity in stratum, selection of few units fulfills the representative character.

c) Replacement of units is possible: Normally in random sampling particular unit is not accessible for study; it is difficult to replace it by another but in stratified sampling replacement of an inaccessible case by an accessible case is possible. Stephen has said: "By providing that fixed proportion of the sample shall come from each geographic area or income class, stratification automatically brings about a replacement of persons lost to the sample by persons of the same stratum. Thus partly correcting the bias would result if there were no replacement of losses".

Demerits of stratified sampling method: The disadvantages of stratified sampling method are described below:

a) Possibility of bias: In this method if a stratification has not been done properly there is every possibility of bias creeping in.

b) Difficult to attain proportion: It is very difficult to attain proportion through design. In random sampling it is achieved automatically. Attainment of proportion becomes particularly difficult when there is wide variance in size of different strata.

c) Difficulty in making the sample representative: In disproportionate type of stratified sampling, the element of weighting introduces the factor of selection. If underweighting has been done, the sample becomes unrepresentative.

d) Difficulty in placing cases under stratum: If the strata's are not very clear-cut, it may not be easy to decide in which unit or stratum a particular case is to be placed. This upsets the applecart of this method.

Systematic Sampling

Systematic sampling is a variation of simple random sampling. It requires the universe or a list of its units may be ordered in such a way that each element of the universe can be uniquely indentified by its order. A voter list, a telephone directory, a card index system would all generally satisfy this condition. Suppose there are 5000 cards (and hence, 5000 units of the universe), and we want a sample of 500. We can select a number between (and including) 1 and 10 at random, say 8. Then we can select the units whose cards are in the following position: 8, 18, 28, 38, …….928,1008, …..4998. This would be a systematic random sample or commonly known as a systematic sample.

Sampling Interval/Ration

Use natural ordering of universe select random starting point between 1 and the nearest integer to the sampling ratio, select items at intervals of nearest integer to sampling ratio. Mathematically,

$$K = N/n$$

Where K = Sampling ratio

N = Universe size

n = Sample size

Advantage of systematic sampling: The main advantage of systematic sampling are the following:

i) Simplicity in drawing a sample, which is easy to check.

ii) Except for populations with periodic behaviour, systematic sampling variances are often somewhat smaller than those for alternative procedures.?

iii) If population is ordered with respect to a relevant property, giving a stratification effect, this helps in reducing variability compared to simple random sampling.

Disadvantage of systematic sampling: The following are the main disadvantages of systematic sampling:

i) If sampling interval is related to a periodic ordering of the universe, increased variability may be introduced.

ii) Estimates of error likely to be high where there is stratification effect.

In practice, it is essential to use systematic sampling only when one is sufficiently acquainted with the data to be able to demonstrate that periodicities do not exist, or sampling interval is not a multiple or sub multiple of the period.

Convenience Sampling

Here the sample is chosen for the convenience of the research worker. Convenience sampling is commonly known as unsystematic, accidental or opportunistic sampling. The researcher makes assumption that the target population is homogenous and the individuals interviewed are similar to the overall defined target population. This in itself leads to considerable sampling error as there is no way to judge the representativeness of the sample. Moreover, the results generated are hard to generalize to a wider population.

Advantages:

i). Convenience sampling is least time consuming and least costly among all

the methods.

ii). The sampling units are accessible, easy to measure and cooperative.

iii). It is a method useful in exploratory research. A street interviewer who needs to sample 50 people, for example, might question the first 50 people who walk past the street corner where the interviewer is standing.

Limitations:

i). Many potential sources of selection bias are present including respondent self-selection.

ii). Convenience samples are not representative of any definable population. Hence it is not theoretically meaningful to generalize to any population from a convenience sample, and convenience samples are not appropriate for marketing research projects involving population inferences.

Applications (Using) of Convenience Sampling:

A convenience sampling may be used in the following cases:

i) When universe is not well defined,

ii) When sampling unit is not clear, and

iii) When complete list of the source is not available.

iv) Convenience samples can be used for focus groups, pretesting questionnaires, or pilot studies.

v) However, this technique is sometimes used even in large surveys.

Judgement Sampling

Judgement sampling, also known as purposive sampling is an extension to the convenience sampling. This makes an attempt to ensure a more representative sample than that gathered using convenience techniques. Research consultants use their expertise, or consult an expert, to evaluate populations and to make recommendations as to which particular units should be sampled. With small populations, accurate assessments and guidance as to a unit's selection, judgement sampling can render samples with less variable error than might result with a

sample chosen using a simple random technique, though this cannot be conclusively proved. In this method of sampling the choice of sample items depends primarily on the judgment of the researcher. In other words, the researcher determines and includes those items in the sample which he thinks are most typical of the universe with regard to the characteristics of research project. For example, if sample of 100 Teachers is to be selected from university having 500 teachers for analyzing the spending habits of teachers, the researcher would select 100 teachers who, in his judgment, are representative of the university.

Merits: This method is sometimes used in solving many types of economic and business problems. The use of judgment sampling is justified by following premises:

i) If there are a small number of sampling units is in the universe, judgment sampling enables inclusion of important units.

ii) Judgment stratification of population helps in obtaining a more representative sample in case research study wants to look into unknown traits of the population.

iii) Judgment sampling is a practical method to arrive at some solution to everyday business problems.

Limitations:

i) The judgment sampling involves the risk that the researcher may establish conclusions by including those items in the sample which conform to his preconceived ideas.

ii) There is no objective way of evaluating the reliability of sample results.

Applications of Judgment Sampling

i) Test markets selected to determine the potential of new products.

ii) Purchase engineers selected in industrial marketing research because they are considered to be representative of the company

iii) Bellwether precints selected in voting behaviour research

iv) Expert witnesses used in court

Applications of Judgment Sampling:

It may be useful if broad population inferences are not required. Judgement samples are frequently used in commercial marketing research projects.

Quota Sampling

Quota sampling is a procedure that restricts the selection of the sample by controlling the number of respondents by one or more criterion. The restriction generally involves quotas regarding respondent's demographic characteristics (e.g. age, race, income), specific attitudes (e.g. satisfaction level, quality consciousness), or specific behaviours (e.g. frequency of purchase, usage patterns). These quotas are assigned in a way that there remains similarity between quotas and populations with respect to the characteristics of interest. This attempts to reflect the characteristics of the population in the chosen sample, and in the same proportions. From national statistics, researchers gather the percentages for such 'stratifiers' as age groupings, income levels etc. and use them to construct 'cells'. This results in statements such as 23 per cent of the population is female, aged between 30 and 40 and earning £12 000-15 000 per annum'. The sample would then be collected, and 23 per cent of it would have to fulfil those demands. Quota controls must be available, easy to use and current. Quota 'stratifiers' shouldn't be used merely because they are available - they must be relevant to the project. This method may be cheaper to operate than a probability-based method, it is quick to use and relatively simple to administrate - it does not require a sampling frame. However, there is the possibility that the interviewer shows bias in the way the individual units are selected and in the difficulty that may arise in uncovering relevant and available quota controls. Quota sampling is most commonly used in marketing survey and election polls. Quota sampling is often called as the most refined form of non- probability sampling.

Merits

i) Reduces cost of preparing sample and field work, since ultimate units can be selected so that they are close together.

ii) Introduces some stratification effect.

Demerits

i) Introduces bias of investigator is not involved at any stage, the errors of the method cannot be estimated by statistical procedures.

ii) Since random sampling is not involved at any stage, the errors of the method cannot be estimated by statistical procedures.

5.4. Types of Data: Secondary and Primary Data

Data Defined

Data in the plural sense implies a set of numerical figures usually obtained by measurement or counting. Data refers to numerical description of quantitative aspects of things. For example, data of students of a college include count of the number of students, and separate count of number of various types of students such as, male and female, married and unmarried, or under graduates and post graduates. It may also include such measures as their height and weights.

Characteristics of Data

In order that numerical description may be called data, they must possess the following characteristics.

i) Data is aggregate of facts: For example, single unconnected figures can not be used to study the characteristics of a business activity.

ii) Data is affected to a large extent by multiplicity of factors: For example in business environment the observations recorded are affected by a number of factors (controllable and uncontrollable).

iii) Data is estimated according to reasonable standard of accuracy: For example in the measurement of length one may measure correct upto 0.01 of a cm., the quality of the product is estimated by certain tests on small samples drawn from big lots of products.

iv) Data is collected in a systematic manner for a predetermined objective: Facts collected in a haphazard manner and without a complete awareness of the objective will be confusing and can not be made the basis of valid

conclusions. For example, collected data on price serves no purpose unless one knows whether he wants to collect data on wholesale or retail prices and what are the relevant commodities under considerations.

v) The Data must be related to one another: The data collected should be comparable, otherwise these can not be placed in relation to each other, e.g. data on the yield of crop and quality of soil are related but the crop yields cannot have any relation with the data on the health of the people.

vi) Data must be numerically expressed: That is, any facts to be called data must be numerically or quantitatively expressed. Qualitative characteristics such as beauty, intelligence etc. are called attributes, and must be scaled to express in numeric terms.

Secondary Data

This is also known as published data. Data which are not originally collected but rather obtained from published sources and statistically processed are known as secondary data. For example data published by Reserve Bank of India, Ministry of Economic Affairs, Commerce Ministry as well as international bodies such as World Bank, Asian Development Bank, International Labour Organization etc.

Advantage

i) Less costly as data is already available.

ii) It is faster to collect and process as compare to primary data.

iii) It provides valuable insights and contextual familiarity with the subject matter.

iv) It provides a base on which further information can be collected to update it and finally use it for the purpose of research.

Disadvantage

i) Locating appropriate source and finally getting access to the data could be time consuming.

ii) The data available might be too vast and a lot of time may be spent going through it.

iii) It might have been originally collected for some purpose which is specific and not known to the present researcher. To that extent, it might be erroneous to use it.

iv) The accuracy of secondary data as well as its reliability would depend on its source.

v) It might not be updated and not of much use in a dynamically changing environment.

Characteristics of Secondary Data

The secondary data should possess the following characteristics:

a) Reliability of Data: Reliability of data can be established from the following information:

i) Who collected the data and from which sources

ii) The method used in collecting the data

iii) Whether census or sampling method is used in collecting the data.

iv) Whether compiler and source both are dependable

v) The purpose for which the data were originally collected.

b) Suitability of Data: The secondary data should be suitable for the enquiry. Even if the data are reliable they should not be used if the same are found to be unsuitable for enquiry. Data may be suitable for one enquiry and may not be suitable for another. For checking suitability of data one should see that.

i) What is the object of enquiry?

ii) The definitions of various terms and units of collection must be carefully scrutinized.

iii) What is the standard of accuracy aimed at?

v) Age of the data. At what time frame the data was collected.

v) Do the data refer to homogenous population?

c) **Adequacy of the data:** The secondary data may be reliable and suitable but

the same may be inadequate for the purpose of current investigation. The data collected earlier may refer to an area which is narrower or wider than the area required for the present enquiry, and if it is such, the data should not be used at all. The secondary data may not cover the period suitable to the enquiry. The degree of accuracy of the original data may be found to be inadequate for the enquiry.

Sources of Secondary Data

Following are the main sources of secondary data.

i) **Official Publications:** Publications of the Central and State Governments, Government of foreign countries or international bodies, etc.

ii) **Semi Official Publications:** Publication of the Semi Government bodies, e.g., Municipal/District Board, Corporation etc.

iii) **Publication relating to Trade:** Publication of the trade associations, chamber of commerce, banks, cooperative societies, stock exchange, trade unions etc

iv) **Journal/Newspapers etc. :** Some newspapers/ journals collect and publish their own data, e.g., Indian Journal of Economics, Economist, Economic Times, Far Eastern Review etc.

v) **Data Collected by Research Agencies**: Research agencies like MARG Nielsen and Gallys also collect useful data which are available as data bases upon payment.

vi) **Unpublished Data:** Data may be obtained from several companies, organizations, universities etc. working in the same areas, and who have done very good work. For example Data on Energy Conservation by The Energy Resources Institute (TERI) can be utilized by private and public sector companies active in this area.

Primary Data

These data are collected first time as original data. Primary data collection requires researchers to get directly involved in the data collection process for the issue at

hand. It may be qualitative or quantitative in nature. Primary data can be collected either through experiment or through survey. If the researcher conducts an experiment, he observes some quantitative measurements, or the data, with the help of which he examines the truth contained in his hypothesis. But in the case of a survey, data can be collected by any one or more of the following ways:

Methods of Collecting Primary Data:

1. Direct Personal Interviews.
2. Indirect Oral Interviews.
3. Information from Correspondents.
4. Mailed Questionnaire Methods.
5. Schedule Sent Through Enumerators.
6. By Observation

1. Direct Personal Interviews:

A face to face contact is made with the informants (persons from whom the information is to be obtained) under this method of collecting data. The interviewer asks them questions pertaining to the survey and collects the desired information. Thus, if a person wants to collect data about the working conditions of the workers of the Tata Iron and Steel Company, Jamshedpur, he would go to the factory, contact the workers and obtain the desired information. The information collected in this manner is first hand and also original in character. There are many merits and demerits of this method, which are discussed as under:

Merits:

1. Most often respondents are happy to pass on the information required from them when contacted personally and thus response is encouraging.
2. The information collected through this method is normally more accurate because interviewer can clear doubts of the informants about certain questions and thus obtain correct information. In case the interviewer apprehends that the informant is not giving accurate information, he may cross-examine him and thereby try to obtain the information.

3. This method also provides the scope for getting supplementary information from the informant, because while interviewing it is possible to ask some supplementary questions which may be of greater use later.

4. There might be some questions which the interviewer would find difficult to ask directly, but with some tactfulness, he can mingle such questions with others and get the desired information. He can twist the questions keeping in mind the informant's reaction. Precisely, a delicate situation can usually he handled more effectively by a personal interview than by other survey techniques.

5. The interviewer can adjust the language according to the status and educational level of the person interviewed, and thereby can avoid inconvenience and misinterpretation on the part of the informant.

Demerits:

1. This method can prove to be expensive if the number of informants is large and the area is widely spread.

2. There is a greater chance of personal bias and prejudice under this method as compared to other methods.

3. The interviewers have to be thoroughly trained and experienced; otherwise they may not be able to obtain the desired information. Untrained or poorly trained interviewers may spoil the entire work.

4. This method is more time taking as compared to others. This is because interviews can be held only at the convenience of the informants. Thus, if information is to be obtained from the working members of households, interviews will have to be held in the evening or on week end. Even during evening only an hour or two can be used for interviews and hence, the work may have to be continued for a long time, or a large number of people may have to be employed which may involve huge expenses.

2. Indirect Oral Interviews:

Under this method of data collection, the investigator contacts third parties generally called 'witnesses' who are capable of supplying necessary information.

This method is generally adopted when the information to be obtained is of a complex nature and informants are not inclined to respond if approached directly. For example, when the researcher is trying to obtain data on drug addiction or the habit of taking liquor, there is high probability that the addicted person will not provide the desired data and hence will disturb the whole research process. In this situation taking the help of such persons or agencies or the neighbours who know them well becomes necessary. Since these people know the person well, they can provide the desired data. Enquiry Committees and Commissions appointed by the Government generally adopt this method to get people's views and all possible details of the facts related to the enquiry. Though this method is very popular, its correctness depends upon a number of factors such as

1. The person or persons or agency whose help is solicited must be of proven integrity; otherwise any bias or prejudice on their part will not bring out the correct information and the whole process of research will become useless.
2. The ability of the interviewers to draw information from witnesses by means of appropriate questions and cross-examination.
3. It might happen that because of bribery, nepotism or certain other reasons those who are collecting the information give it such a twist that correct conclusions are not arrived at.

Therefore, for the success of this method it is necessary that the evidence of one person alone is not relied upon. Views from other persons and related agencies should also be ascertained to find the real position .Utmost care must be exercised in the selection of these persons because it is on their views that the final conclusions are reached.

3. Information from Correspondents:

The investigator appoints local agents or correspondents in different places to collect information under this method. These correspondents collect and transmit the information to the central office where data are processed. This method is generally adopted by news paper agencies. Correspondents who are posted at different places supply information relating to such events as accidents, riots, strikes, etc., to the head office. The correspondents are generally paid staff or

sometimes they may be honorary correspondents also. This method is also adopted generally by the government departments in such cases where regular information is to be collected from a wide area. For example, in the construction of a wholesale price index numbers regular information is obtained from correspondents appointed in different areas. The biggest advantage of this method is that, it is cheap and appropriate for extensive investigation. But a word of caution is that it may not always ensure accurate results because of the personal prejudice and bias of the correspondents. As stated earlier, this method is suitable and adopted in those cases where the information is to be obtained at regular intervals from a wide area.

4. Mailed Questionnaire Method:

Under this method, a list of questions pertaining to the survey which is known as 'Questionnaire' is prepared and sent to the various informants by post. Sometimes the researcher himself too contacts the respondents and gets the responses related to various questions in the questionnaire. The questionnaire contains questions and provides space for answers. A request is made to the informants through a covering letter to fill up the questionnaire and send it back within a specified time. The questionnaire studies can be classified on the basis of:

i. The degree to which the questionnaire is formalized or structured.

ii. The disguise or lack of disguise of the questionnaire and

iii. The communication method used.

When no formal questionnaire is used, interviewers adapt their questioning to each interview as it progresses. They might even try to elicit responses by indirect methods, such as showing pictures on which the respondent comments. When a researcher follows a prescribed sequence of questions, it is referred to as structured study. On the other hand, when no prescribed sequence of questions exists, the study is non-structured.

When questionnaires are constructed in such a way that the objective is clear to the respondents then these questionnaires are known as non- disguised; on the other hand, when the objective is not clear, the questionnaire is a disguised one.

On the basis of these two classifications, four types of studies can be distinguished:

1. Non-disguised structured,
2. Non-disguised non-structured,
3. Disguised structured and
4. Disguised non-structured.

There are certain merits and demerits of this method of data collection which are discussed below:

Merits:

1. Questionnaire method of data collection can be easily adopted where the field of investigation is very vast and the informants are spread over a wide geographical area.
2. This method is relatively cheap and expeditious provided the informants respond in time.
3. This method has proved to be superior when compared to other methods like personal interviews or telephone method. This is because when questions pertaining to personal nature or the ones requiring reaction by the family are put forth to the informants, there is a chance for them to be embarrassed in answering them.

Demerits:

1. This method can be adopted only where the informants are literates so that they can understand written questions and lend the answers in writing. 2. It involves some uncertainty about the response. Co-operation on the part of informants may be difficult to presume.
3. The information provided by the informants may not be correct and it may be difficult to verify the accuracy.

However, by following the guidelines given below, this method can be made more effective:

The questionnaires should be made in such a manner that they do not become an

undue burden on the respondents; otherwise the respondents may not return them back.

i). Prepaid postage stamp should be affixed

ii). The sample should be large

iii). It should be adopted in such enquiries where it is expected that the respondents would return the questionnaire because of their own interest in the enquiry.

iv). It should be preferred in such enquiries where there could be a legal compulsion to provide the information.

5. Schedules Sent Through Enumerators:

Another method of data collection is sending schedules through the enumerators or interviewers. The enumerators contact the informants, get replies to the questions contained in a schedule and fill them in their own handwriting in the questionnaire form. There is difference between questionnaire and schedule. Questionnaire refers to a device for securing answers to questions by using a form which the respondent fills in him self, whereas schedule is the name usually applied to a set of questions which are asked in a face-to face situation with another person. This method is free from most of the limitations of the mailed questionnaire method.

Merits:

The main merits or advantages of this method are listed below:

1. It can be adopted in those cases where informants are illiterate.
2. There is very little scope of non-response as the enumerators go personally to obtain the information.
3. The information received is more reliable as the accuracy of statements can be checked by supplementary questions wherever necessary.

This method too like others is not free from defects or limitations. The main limitations are listed below:

Demerits:

1. In comparison to other methods of collecting primary data, this method is quite costly as enumerators are generally paid persons.
2. The success of the method depends largely upon the training imparted to the enumerators.
3. Interviewing is a very skilled work and it requires experience and training. Many statisticians have the tendency to neglect this extremely important part of the data collecting process and this result in bad interviews. Without good interviewing most of the information collected may be of doubtful value.
4. Interviewing is not only a skilled work but it also requires a great degree of politeness and thus the way the enumerators conduct the interview would affect the data collected. When questions are asked by a number of different interviewers, it is possible that variations in the personalities of the interviewers will cause variation in the answers obtained. This variation will not be obvious. Hence, every effort must be made to remove as much of variation as possible due to different interviewers.

6. By Observation

This method implies the collection of information by way of investigator's own observation, without interviewing the respondents. The information obtained relates to what is currently happening and is not complicated by either the past behaviour or future intentions or attitudes of respondents. This method is no doubt an expensive method and the information provided by this method is also very limited. As such this method is not suitable in inquiries where large samples are concerned. It is further discussed in the next section of the same unit.

Advantage: The advantages of primary data are:

i) Primary data is more accurate and gives detailed information according to the requirement.

ii) The explanation of terms, definition, and concepts are incorporated in

primary data.

iii) Methods of collection, its limitations and other allied aspects are highlighted.

iv) It is more reliable and less prone to errors.

v) It often includes a copy of the schedule and description of the procedure used in selecting the sample and collecting the data.

Limitations: The limitations are given below:

i) It is expensive to collect primary data.

ii) It is time consuming method of data collection.

iii) It requires experts/trained personnel to collect the primary data. Otherwise it may lead to wrong observations/unreliable data collection.

5.4.1 Various Methods of Collection and Data

The various methods of data collection are as under:

i) Observation method

ii) Personal Interview

iii) Schedules

iv) Documented sources of data

v) Case study method

vi) Questionnaire method

(i) Observation Method

This is the most commonly used method data collection especially in studies relating to behavioural sciences. In observation studies, the researcher observes the behaviour of individuals in real-life settings. Accurate watching and noting of phenomenon as they occur in nature with regard to cause and effect or mutual relation is called observation method of data collection. This type of research originated in anthropology and has percolated into many other field of research. There is a still debate among researchers as to whether observation is a

quantitative or qualitative technique. Observation methods are widely used in organisation research to examine how people behave in groups, in teams and as organisation members. The observation studies are extremely useful in collecting behavioural data as oppose to attitudinal data. The main characteristic of all observation techniques is that researcher must rely heavily on their powers of observing rather than actually communicating with the people to collect primary data.

Characteristics of Observation Method

These are as follows:

i) Direct Method: In observation method data is collected through direct contact with phenomenon under study. In this method sensory organs particularly eye, ear, voice are used.

ii) Source of Primary Data: That is a classical method for collection of primary data.

iii) Requires In-depth study: In this method, the observer goes to the field and makes the study of the phenomenon in an in-depth fashion to acquire data.

iv) Collection follows observation: In this method the investigator first of all observes the things and then collects the data.

v) Relationship between the cause and affect: Observation method leads to development of relationship between the cause and effect of the events.

vi) Scientific method for collecting dependable data: This is the most scientific method for collection of dependable data. Observations are planned and recorded systematically. There should be checks and balances on this methodology.

vii) Selective and Purposeful collection: The observations are made with definite purpose. Collection of materials is done according to a particular purpose.

Merits of Observation Method

These are as follows

a) Common method: The method of observation is common to all the discipline of research.

b) Simplicity: The method is very simple to use.

c) Realistic: Since observation is based on actual and first hand experience, its data are more realistic than the data of those techniques which are indirect and secondary source of information.

d) Formulation of hypothesis: In all the business operations, the method of observation is used as the basis of formulating hypothesis, regarding business research problem.

e) Verification: For verification of hypothesis, again we depend upon observation. Therefore, it can be said that the problem presents itself and resolves itself through observation method.

f) Greater reliability of conclusions: The conclusions of observations are more reliable than non-observation conclusions, because they are based on first hand perception by the eyes and can be verified by any one by visual perception.

On account of the above advantage, the observation method is called Classical Technique of Investigation for research purposes.

Limitations of Observations Method

i) Some events can not be objects of observation: There are certain events which are microscopic, indefinite and may not occupy any definite space or occur at a definite time and can not be noticed for observation purposes. For example, it is not possible to observe emotions and sentimental factors, likes and dislikes etc.

ii) Illusory observation: Since we have no depend upon our eyes for observation, we can never be sure if what we are observing is the same as it appears to our eyes, Eyes are prone to deception. It is well known that eyes see a mirage in desert at noon.

iii) Self-consciousness in the observed: In observation method, the atmosphere

tends to become artificial and this leads to a sense of self consciousness among the individuals who are being observed. This hampers their naturalness in behaviour and thus the purpose of observation which is to know the behaviour of individuals under normal conditions get defeated.

iv) Subjective explanation: The final results of observation depend upon, the interpretation and understanding of the observer, the defects of subjectivity in the explanation creep in description of the observed and deductions from it. For example, if we see a man coming out of a wine shop, quite drunk, and he starts firing at random, we may believe that liquor induces irrational violence in a man, which may not be the case always.

v) Slowness of Investigation: The slowness of observation methods lead to disheartening, disinterest among both observer and observed,.

vi) Expensive methodology: Being a long drawn process, the technique of observation is expensive.

vii) Inadequacy: The full answer cannot be obtained by observation alone; observation must be supplemented by other methods of study.

(ii) Interview Method

Under this method of collecting data there is a face to face contact with the persons from whom the information is to be obtained (known as informants). This method allows the researcher to collect both attitudinal and behavioural data from the respondents from all the time frames (past, present and future). The interviewer asks them questions pertaining to the survey and collects the desired information. Thus if a person wants to collect data about the working conditions of the workers of Hindustan Unilevers Ltd., Mumbai, he would go to works at Mumbai, contact the workers and obtain the information. The information is obtained at first hand and is original in character.

Characteristics of Interview Method

The following are the main characteristics of interview method.

a) It is a close contact or interaction including dialogue between two or more persons.

b) There is a definite object of interview, such as knowing the ideas and views of others.

c) There is a face to face contact or primary relationship between the individuals.

d) This is the most suitable method of data collection for business and economic problems.

Merits of Interview Method

As compared with other research methods, the method of interview possesses some unique qualities. These are:

1. Direct research: In the interview method, the researcher has not to confine himself to the external aspects of human behaviour; but can probe into the internal aspects as well. Furthermore, in the interview the barrier between the researcher and the respondent is eliminated any they know each other directly.

2. In depth research: This characteristic follows from the above mentioned fact that the interview studies the internal aspects of the research problem, which are inaccessible to other methods. Accordingly, in comparison with other methods, the interview method is a method of in-depth research.

3. Knowledge of past and future: In interview we also learn about the outlook, aspirations and future goals of human beings and their present abilities. Accordingly, by interview we unravel the hidden past and prognosticate about the future.

4. Mutual encouragement: In an interview there is inflow and outflow of ideas between the interviewer and the interviewee. This exchange proves encouraging for both the researcher and the respondent.

5. Supra-observational: An interview gives us knowledge of facts which are inaccessible to observation. The emotional attitude, secret motivation and incentives governing human life come to surface in an interview though these are unobservable. Therefore interview has a quality which may be called

supra-observational.

6. Examination of known data: The information given by the interviewee, if suspect, can be tested through cross-examination of the interviewee. Moreover, body language accompanying the responses give a clue to the interviewer about the veracity or otherwise of the answer being given.

Limitations of Interview Method

Inspite of the above-mentioned merits of the method of interview, it suffers from certain limitations which are:

1. Inadequate information: There are certain matters which can be written in privacy but about which one does not speak before others. If these matters are subject of an interview, the likelihood is that only a disguised version of these will be presented. Again, there are people who are temperamentally unable to discuss things though they are powerful writers. These persons are also unlikely to present true facts in an interview.

2. Defects due to interviewee: If an interviewee is of low level intelligence he is unfit to give correct information. Some persons are in the habit of talking in around about manner and it is impossible to decipher what they say.

3. Prejudices of Interviewer: The prejudices of interviewers are as much problem of research as are the inadequacies of the interviewees. If the interviewer is unable to suppress his prejudices, his understanding and interpretation of the information given in interview will be defective.

4. One-sided and incomplete research: In the interview, certain aspects of human behaviour get over-emphasized at the expense of others. There is a tendency to give too much importance to personal factors and minimize the role of the environment factors. For these reasons, the research by interview is liable to suffer from one-sidedness.

5. Interviewing is an art rather than science: Another limitation of interview method is that its procedures cannot be standardized; there is too much room for improvision. The success of an interview is more due to skill and tact than due to knowledge. However, the success in interview depends

exclusively on the intelligence and skill of the interviewer. Therefore, the method of interview is more an art than science.

6. Difficulty in Persuading the Interviewee: Many people are unwilling to participate in interviews. Under these circumstances, the first problem before the interviewer is to persuade the prospective interviewee to extend his cooperation for the research project and agree for being interviewed.

Types of Interviews

The chief types of interviews are as follows:

1. Classification according to formalness: The formal classification of interview gives us two main types.

a) **Formal Interview**: In this type of interview, the interviewer presents a set of well defined questions and notes down answers of information in accordance with prescribed rules.

b) **Informal Interview:** In contrast with the formal interview the interviewer has full freedom to make suitable alterations in the questions to suit a particular situation in formal interview. He may revise, re-order or paraphrase the question to suit the needs of the respondents.

2. Classification according to the number: Anther classification of interview is according to the number of persons taking part in it. Following are its main types.

i) **Personal Interview**: In personal interview single individual is interviewed. The personal interview help to establish close personal contacts between the interviewer and the interviewee and as a result detailed knowledge about intimate and personal aspects of the individual can be had.

ii) **Group Interview:** As the name makes it plain the group interview is the opposite of the personal, because in it two or more persons are interviewed. The interview is suited for gathering routine information.

3. **Classification according to purpose:** The interviews have also been classified by the purpose for which they are held. Following are the types

of this classification:

i) **Diagnostic Interview**: As the name makes clear, this type of interviewers, try to understand the cause or causes of a malady. In clinical psychology and psychoanalysis, the preliminary interviews with the patients are held with a purpose to grasp the nature and cause of disease.

ii) **Research interview:** These interviews are held to gather information pertaining to certain problems. The questions to be asked to gather the desired information are pre-determined and by asking them of the informations the data is collected. In as such as this data is gathered for the purpose of research into a problem, these are called research interviews.

iii) **Interviews to fulfil curiosity:** These interviews, as the name implies, are held to satisfy some question lurking in the mind of a scientist. For example, if a scientist gets an idea that good lectures are delivered extempore, he has to interview some reputedly good lecturers whether they make extensive notes for delivering a lecture or not.

4. **Classification according to the period of contact:** The different types of problem require different amount of time for contact with respondents. The time can be short or long. Accordingly, two types according to time are follows:

i. **Short-contact interview:** For filling-up schedules etc., a single sitting of small duration suffices. Therefore, in researches of this type short contact interview suffices.

ii. **Prolonged contact interview:** In contrast with research by schedule, the case-history method requires prolonged interviews. In this establishment of close personal relations between the interviewer and interviewee is very likely.

5. Classific ation according to subject-matter: The classification of interviews according to subject-matter gives us the following three types:

i. Qualitative interview: The qualitative interviews are about complex and non-quantifiable subject-matter. For example, interviews held for case studies

are qualitative, because the interviewer has to range over past, present and future to know enough about a case

ii. **Quantitative interview**: The quantitative interviews are those in which certain set facts gathered about large number of persons. The census interviews are its example

iii. **Mixed interview:** In certain interviews both types of data the routine and specialized is sought, part of it is quantifiable while the rest is not. Therefore it is known as mixed interview.

(iii) Using Schedule Method for Data Collection

Schedule is the name usually applied to a Performa containing a set of questions which are asked and filled by an interviewer in a face situation with a respondent. It is a standardized device or tool of observation to collect the data in a objective manner. In this method the interviewer puts certain questions and the respondent furnishes certain answers and the interviewer records them as in a research instrument called schedule.

Purposes/Objectives of the Schedule

The main objectives of the schedules are as follow:

i) **Delimitation of the topic:** A schedule is always about a definite item of research study its subject is a single subject item rather than the research subject in general. The schedule delimits and specifies the subject of inquiry.

ii) **Aids memory:** It is not possible for the interviewer to keep in mind or memorize all the information that be has collected from different respondents. If no standardized tool is available he might put different questions to different persons and thereby get confused when he has to analyse and tabulate the data. Schedule acts as a aide memoire.

iii) **Aid to classification and analysis:** Another objective of schedule is to tabulate and analyse the data collected in a scientific manner. Through schedules, he can collect the matter in a homogeneous manner.

Types of Schedules

Schedules that are used in business research are classified as under:

i) **Observation schedule:** The schedules that are used for observation are known as observation schedule. In these schedules observer records the activities and responses of a worker or a group under specific conditions. The main purpose of the observation schedule is to verify information.

ii) **Rating schedule:** In the fields of business guidance, psychological research , and social research ,the rating schedules are used to assess the attitude, opinions, attitudes, preferences, inhibitions and other like elements. as is evident from the term rating , in these schedules the value and the trend of the above mentioned qualities is measured on a rating scale.

iii) **Document schedule**: The schedule of this type are used to obtain data regarding return evidence and case histories from autobiography, diary, case histories or final records of governments etc. It is a good method for collecting exploratory data or preparing source list.

iv) Interview schedule: In an interview schedule an interviewer presents the questions of the schedule to interviewee and records their responses on blank places.

Characteristics of A Good Schedule

The following are the essentials or characteristics of a good schedule.

1. **Accurate communication**: It means that the questions that are given in schedule should be such that the respondent is able to understand them in the light in which they are asked. For accurate communication, the questions should be of the following types.

a) **Questions should be interlinked:** It means that if information about different aspects is required, the questions asked should be such that their answers may present compact picture of the information.

b) **Suggestive questions:** The questions should be suggestive. There should be questions on each topic, but the questions should be so designed that

the respondent may be stimulates to give the correct answers

2. **Accurate response:** It means that the schedule, should be such that the re-quired information may be easily secured. For this the interviewer has to prepare the schedule in a scientific manner and also make efforts to inspire the respondent to give answers. For this the following steps should be taken.

a) **The size of the schedule should be attractive:** It should not be too lengthy to make the respondent bored.

b) **The questions of the schedule should be clearly worded and be unambiguous**: Even if some unpalatable information is to be collected the framer of the schedule should couch his questions in such a language that the information is secured without injuring the feelings of the respondent.

c) **The questions free from subjective evaluation:** The questions should be relevant and pointed. They should not deal with the subjective evaluation

d) **Information sought should be capable of being tabulated**: Questions of the schedule should be so framed that the information collected through them should be capable of being tabulated and if needed, subject to statistical analysis.

Suitability of Schedule Method

This method is generally employed in following situations:

a) The field of investigation is wide

b) Where the researcher/investigator requires quick results at low cost.

c) Where the respondents are educated.

d) Where trained and educated investigates are available.

Limitations of Schedule Method

Schedule method of data collection, like all other methods, has limitations. Some of these are:

1. **Costly and time consuming**: As compared with questionnaire, schedule

method is generally costlier and more time -consuming. But this factor becomes a serious limitation when the correspondents are physically scattered over a wide area. To approach all of them is prohibitively costly, besides involving excessive time

2. **Requirement of a large number of well-trained field workers:** The schedule method requires very large number of well trained and experienced field workers. Therefore is becomes very difficult if not impossible, to hire a large number of experienced workers. It involves great cost and sometimes so many workers are not easily available and one cannot depend upon in experienced hands.

3. **Adverse Effect of Personal Presents:** Where as the personal presents proves helpful in assuring the response, removing their doubts, it also becomes an inhibiting factor. Many people can note down certain facts on papers but cannot say them in presents of others.

4. **Organizational Difficulty:** If the field of research is geographically wide, it becomes difficult to organize research. To gather workers who are well acquainted with various geographical region and different types of people is a mammoth task. Though not beyond achievement, it is certainly difficult.

(iv) Documented Sources of Data

Meaning of document: Document is a very important, dependable and valuable source of information. Many researchers have made use of this vital source. Document is nothing but a written record that contains important information about a problem or aspect of study. It may be a report, a diary, letter, history, history, official and non-official records, proceedings of the legislature, committees, societies, surveys, journals, periodicals, speeches etc.

Types of Document: Strictly speaking it is very difficult to classify the documents. All the documents have different types and traits and elements in them. For the convenience of the study they have been classified under the following two heads:

i. Personal documents

ii. Public documents

Personal Documents

These documents are recorded by the individuals. An individual may record his view and thought about various problem. He may do so because of his personal interest in those problems and without knowing that these documents at a later date may form it a subject or source of study.

Types of personal documents: Personal documents may be categorized or divided under the following heads for the convenience of the study:

i) Life history

ii) Spontaneous autobiography

iii) Voluntary autobiography or self-record

iv) Letters

v) Memoires

In area of business research these personal document are of not of much use except in case these pertain to business leaders. Even in such a case they have limited use.

Public Documents

Public documents are quite different from personal documents. They deal with the matters of different interest. Public documents may be divided into the following two categories.

i) Unpublished records: Such records, although they deal with the matters of public interest, are not available to people in published form. It means that everybody cannot have access to them. Proceedings of the meetings, nothing on the files and memoranda etc., form the category of unpublished records. It is said that these records are very reliable. Since there is no fear of their being made public, the writers give out their views clearly.

ii) Published records: These records are available to people for investigation and perusal. Survey reports, report of enquiries and such other documents fall under this category.

The data contained in these documents are considered by some people as quite reliable because the collecting agency knows that it shall be difficult to test while others are of the view that if the data are to be published , the collecting or publishing agency does some window-dressing as a result of which the accuracy is sometimes doubtful.

Most of the information that is now available to people and researchers in regard to business environment, are to be found in the form of reports. The reports published by government are considered as more dependable. On the other hand some people think that the reports that are published by certain private individuals and agencies are more dependable and reliable.

(v) Case Study Method

Case study method may be defined as small inclusive and intensive study of an individual in which investigator brings to bear all his skills and methods or as a systematic gathering of enough information about a person to permit one to understand how he or she functions as until of society. The case study is a form of qualitative how he or she functions as unit of society. The case study is a form of qualitative analysis involving a very careful and complete observation of a person, situation or institution.

Case study is a method of exploring and analyzing business aspects of an industrial unit, even entire industry.

Characteristics of a case study: The important characteristics of case study method are as under:

i) **Study of a unit:** The case study method studies a subject matter which forms a cohesive, whole and may be treated as a unit. The unit can be an individual, a family, an institution.

ii) **Intensive or** In-Depth Study: Case study attempts a deep and detailed study of the unit. It is a method of study in depth rather than breadth. It places more emphasis on the full analysis of a limited number of events or conditions and their inter-relations.

iii) **Knowledge of behaviour patterns:** The case study method deals with

both what and why of the subject. It tries to describe the complex behavioural pattern of a unit and having done this, tries to discover the factors which will rationally account for them. In brief, case study method aims at description as well as explanation of the unit it studies. It also explains the place and role of a unit in its surrounding social milieu.

iv) **The study of the whole unit:** The case study method tries to perceive the unitary forces of the subject matter and organizes it into an integral whole.

Basic Assumptions of the case study method: Following are the basic assumptions of the case study method.

a) **Totality of the being:** In this method, the unit of the study which may be an individual or a business group is studied as a unit. This study is confined to a particular time or situation but it is in totality or in all its aspects.

b) **Underlying Unit:** It is believed that a unit is the representative of a type and it should be studied as a type rather than as an individual unit. This assumption involves that the units are the same and there is no difference in studying a particular unit.

c) **Complexity of the business environment:** This study is based on the assumption that the business environment is a very complex affair and so deeper study is required. Therefore the case data are gathered of an entire life cycle of an industrial product or unit.

d) **Influence of the time factor:** Business environment gets influenced by the time factor, as it is dynamic in nature. Accordingly, the study can not be worthwhile unless it is long range and over considerable period of time so that institutions and groups in their various aspects may become well known.

e) **Similarity of response in human beings:** The case study method believes that the fundamental responses from human beings will be more or less the same.

f) **Resources or circumstances:** Besides believing in the fundamental of consistency of human nature, it also believes that the business conditions and circumstances tend to recur from time with marginal changes.

Advantages of case study method: The case study method is very popular method of collecting data about industrial units and industries as a whole. The main advantages of case study methods are as follows:

i) **Intensive and deep study of the problems is possible:** The case study method enables one to understand fully the behaviour pattern of the concerned unit. In this method the problem is recognized as a unit and various aspects of the problem are subjected to deep and detailed study.

ii) **Study of the subjective aspects:** Through case study a researcher can obtain a real and enlightened record of specific experiences, which would reveal a business units problems and motivation that drives the unit to adopt a certain pattern of behaviour.

iii) **Comparative Study:** In this method, all the aspects of life of an industrial unit are studied. Through this study the characteristics of one particular industry or firm may be differential from the characteristics of an other. Thus the method is helpful in comparative study.

iv) **Formulation of valid hypothesis:** Once the various cases are extensively studied and analysed, the researcher can deduce various generalizations, which may be developed into useful hypothesis.

v) **Useful in framing in questionnaire and schedules:** Case study is of great help in framing questionnaire, schedule and other forms. This in turn helps in getting prompt response.

vi) **Sampling:** Case study is helpful in stratification of the sample. By studying the individual units the researcher can put them in definite classes or types and thereby facilitate the perfect stratification of the sample.

viii) **Study of process:** In cases where the problem under study constitutes a process and not one incident, e.g. merger process, cartel formation etc., the case study is the appropriate method.

viii) **Use of several research methods:** The researcher can use one or more of the several research methods under the case study method depending upon the prevalent circumstances.

Limitations of the Case Study Method

i) **Unrealistic assumptions**: Case study method is based on several assumptions which may not be very realistic at times, and as such the usefulness of case data is always subject to doubt.

ii) **Problem of finance, time and energy:** Through this method, it is not possible to cover the large area for study. It requires large finance time and energy to complete the study.

iii) **False generalization:** Generalization are formulated on the basis of data collected. If the data collected are wrong, the generalization shall be wrong.

iv) **Difficult to test reliability of the validity of the data collected**: This study is based on the information that is given by an individual or a firm. If the information furnished is wrong, there is no method on the basis of which researcher can test the reliability of the information.

v) **Not possible to apply sampling method:** In this method individual is recognized as a unit and the entire study is concentrated on that unit. It is necessary that the unit may be representative of the universe. Through the study of a particular unit, it is not possible to apply the knowledge of other units.

vi) **Defective records**: If the records on the basis of which data are collected are defective and not based on objectivity the generalization becomes defective.

vi) **Lack of quantitative study:** It is not possible to quantify the feelings, emotions, reactions, values etc., and as such it lacks scientific temper.

VI) Questionnaire Method: This method of data collection is discussed in section 5.5

5.5. Preparation or Process of Questionnaire

Questionnaire Method: Under this method, a list of questions pertaining to the survey (known as questionnaire) is prepared and sent to the various informants by post. Questionnaire contains the questions and provides the space for answers.

A request is made to respondents through a covering letter to fill up the questionnaire and send it back within a specified time.

Designing questionnaire has been always an issue of debate in marketing research as some researchers view it as art which is based on experience of the researcher,14 while others consider it as a science based on sound theoretical development.15 While the debate is still going on with regard to what a questionnaire design is all about, there is consensus among the research community that the designing process involves some established rules of logic, objectivity and systematic procedures. While the systematic procedure provides guidelines to avoid major mistakes, each questionnaire requires a customized path for development. The generic structure in developing questionnaire is described as follows:

(a) Specification of the information needed in researchable format

(b) Selection of interview method

(c) Determination of question composition

(d) Determination of individual question content

(e) Developing question order, form and layout

(f) Pilot testing the questionnaire

(a) Specification of the information needed in researchable format

The first step in developing a questionnaire is to specify the information needed in researchable format. The researcher should also look at the research objectives and hypotheses. At this stage, it is very important to have a clear idea of target population and sample. The characteristics of the respondents have a great influence on questionnaire design. For example, questions which are appropriate for elderly consumers might not be appropriate for young consumers. Unclear understanding of the information needed could lead to the development of an improper questionnaire which has direct effect on the analysis and the final results.

(b) Selection of interview method

We discussed various methods of interview including personal, mail, telephone

and internet based interviews. The type of interviewing method also plays an important role in questionnaire design. For example, in personal interview situations, respondents are able to see the questionnaire and interact in person with the interviewer. This provides an opportunity to ask varied questions involving complexities because instant feedback mechanism is available. Due to the personal interaction it is also possible sometimes to ask lengthy questions. In telephone interviews, because the respondent cannot see the questionnaire it is quite hard to ask complex and lengthy questions. Therefore, the questions should be short and to the point involving little complexity. Even with the use of computer assisted telephone interviews (which involves sophisticated skip patterns and randomization) the questions have to be kept simple. The length related issues can be dealt with in mail questionnaire however because in this situation the respondent is left on his or her own it is recommended that the questions be kept simple. Internet based questionnaire provide high level of interactivity however, as the respondent is trying to tackle each question on his or her own, the researcher must take this into consideration in questionnaire development process. The interview method also has an effect on the scaling technique due to the issue of complexity. In personal interviews most complex scales can easily be used however, in telephone interviews researchers tend to prefer nominal scales. At times researchers have used other scales in telephone interviews with varied effects. In mail interviews complex scales can be used however, detailed explanation with examples is always desirable. Similar pattern is also observed in internet based interviews.

(c) Determination of question composition

Once the information is specified in the researchable format and the interview method is decided, the next stage for the researchers will be to determine what kind of question are they going to ask to the respondents. There are two major types of question structures: unstructured (also called open ended questions) and structured (also called close ended questions).

Unstructured questions (or open-ended questions) are questions in which respondents are asked to answer the questions in their own words. These types

of questions allow the respondents to express their general attitude and opinions and provide rich insights relating to the respondents views about a certain phenomenon. Unstructured questions are highly used in exploratory research. While unstructured questions provide freedom of expression there are inherent disadvantages associated with them with regard to interviewer bias. If the interviewer is recording the answers by writing the summary down while respondents speaks, the recording may be biased as its based on skills of interviewer on deriving the main points. It is always advisable to use audio recording if possible. Another disadvantage of this questioning is creating coding and interpretations. The overall coding of unstructured questions is costly and time consuming. To avoid mistakes of response recording and coding related errors, researchers use pre-coding wherein they identify possible answers and assign responses to the categories they have identified.

Most conclusive studies employ structured (or close-ended) questions. These types of questions allow the respondents to answer the questions in a pre-defined format. There are three main types of structured questions, dichotomous, multiple choice and scale questions. This type of question format reduces the amount of thinking and effort required by respondents. Interviewer bias in eliminated with unstructured questions because either the interviewer or respondents themselves have to check a box or a line, circle a category, hit a key on a keyboard or record a number.18 In simple words, structured format gives the researcher an opportunity to control the respondent's thinking and allows simplicity. Of the three major types of structured questions, dichotomous question is the simple most questioning category. A dichotomous question has only two response alternatives, yes or no, male or female and so on. Sometimes, a neutral alternative is also added in the questions such as 'don't know' or 'no opinion'. While simplicity is the greatest advantage of dichotomous questions, the response bias becomes a great disadvantage also. Dichotomous questions are good when considering collecting demographic information however, with attitude measurement they are of little use. Multiple choice questions provide an extension to the dichotomous question wherein a respondent is provided with a set of alternatives and is allowed to choose more than one alternative. Multiple choice

questions also have an inherent position and order bias wherein respondents tend to choose the first or last statement in the list. To avoid such bias several forms of the questions with the same alternatives should be prepared. This can easily be handled when interviewing respondents on internet or on telephone using CATI. Another disadvantage of multiple choice questions is the effort required in developing an effective question. A theoretical exploration as well as an exploratory study can assist in such process.

(d) Determination of individual question content

Each individual question is unique from its content perspective and therefore must be treated with caution in the development process. Using components such as words, order, tenses and so on, each question attempts to fulfil the overarching research objectives.

One of the most important components of any question is words. Researchers have to be very clear in the choice of words which can easily be understood in the correct manner by respondents. If the researchers and respondents do not assign the same meaning to the used words, the response will be biased. Wording of a question could create problems such as ambiguity, abstraction, and connotation

To avoid these problems researchers can take several steps such as:

1) Use ordinary words which can easily be understood by the respondents

(2) Avoid ambiguous words

(3) Avoid leading questions

(4) Avoid implicit questions

(5) Avoid generalizations

(6) Avoid double barrelled questions

(e) Developing question order, form and layout

The question order, format and layout can have a significant impact on respondent engagement. Questionnaire with unclear order, format and layout generally get very low response rate and in turn become costly exercise. The questionnaire

can be divided in three main parts generally: forward and opening questions; generic information questions; specific information questions.

The forward and opening questions are highly important in gaining respondents' trust and making them feel comfortable with the study. It also improves the response rate among the respondent if they find it worthwhile and interesting. Questions pertaining to opinion can give a good start to most questionnaires as everyone likes to give some opinion about issues at hand. At times, when it is necessary to qualify a respondent (i.e. determine if they are part of the defined target population), opening questions can act as qualification questions.

Generic information questions are divided into two main areas: classification information questions and identification information questions. Most socioeconomic and demographic questions (age, gender, income group, family size and so on) provide classification information. On the other hand, respondent name, address, and other contact information provide identification information. It is advisable to collect classification information before identification information as most respondents do not like their personal information collected by researchers and this process may alienate the respondent from the interview.

The specific information questions are questions directly associated with the research objectives. They mostly involve various scales and are complex in nature. This type of questions should be asked later in the questionnaire after the rapport has been established between the researcher and the respondent. Most researchers agree that it is good to start with forward and opening questions followed progressively by specific information question and concluding with classification and identification information questions.

The format and layout of the questionnaire has a direct impact on respondent engagement. It is always suggested that the questionnaire format and layout should have some type of symmetry. This can lead to higher response rate.

(f) Pilot testing the questionnaire

Once the preliminary questionnaire has been developed using the above stated

process a researcher should assign coding (discussed in the next chapter) to every question and test the questionnaire on a small sample of respondents to identify and eliminate potential problems. This sampling process is called pilot testing. It is advised that, a questionnaire should not be used in the field survey without being adequately pilot tested. A pilot test provides testing of all aspects of a questionnaire including, content, wording, order, form and layout.20 The sample respondents selected for the pilot test must be similar to those who will be included in the actual survey in terms of their background characteristics, familiarity with the topic and attitudes and behaviours of interest. An initial personal interview basedpilot test is recommended for all types of surveys because the researcher can observe respondents' attitudes and reactions towards each question. Once the necessary changes have been made using the initial personal interview based pilot test, another pilot test could be conducted for mail, telephone or internet based survey. Most researchers recommend a pilot test sample between 15 and 30 respondents. If the study is very large involving multiple stages, a larger pilot test sample may be required. Finally, the response obtained from the pilot test sample should be coded and analysed. These responses can provide a check on the adequacy of the data obtained in answering the issue at hand.

Types of Questionnaire

The questionnaire may be of following types:

a) Structured Questionnaire: structured Questionnaires are those in which a question is presented the respondents with fixed response categories.

b) Un-structured questionnaire: Here every question is not necessarily presented to the respondent in the same wording and does not have fixed responses. Respondents are free to answer the question the way they like.

c) Disguised typed questionnaire: Here the questions are direct and therefore respondents may give answer to certain sensitive questions whose accuracy may be questioned.

c) Mixed Questionnaire: This is a questionnaire which is neither completely structured nor un-structured. It consists of both the types of questions.

Advantage of Questionnaire Method

This method is an indirect method of data collection. It has certain advantages as compared to other methods. It is merits are as follows:

i) Economical: In comparison to other methods of data collection (observation methods, case study, interview etc.) the mailed questionnaire method is cheapest and quickest method. The cost in this method is only that of getting the questionnaire prepared and the postage expense. There is no need to visit the respondents personally or continue the study over a long period.

ii) Less skill of administration: The questionnaire method requires less skill to administer than an interview, observation or case study method of data collection.

iii) Research in wide area: If the informants or the respondents are scattered in a large geographical areas, the Questionnaire method is only means of research. The other methods of data collection such as schedule, interview or observation method do not prove to be successful. Even after spending large amount of money, it may not be possible to collect the information quickly but through questionnaire method, large area can be covered. Some times certain agencies also cooperative in the task of dispatches or sending of the questionnaire to the informants.

iv) Time Saving: Besides saving money, questionnaire method saves time. Simultaneously hundreds of persons are approached through it whereas if they are to be interviewed it may take a long time.

v) More reliable in special cases: This is a method of collecting data in an objective manner through standardized impersonal questions. The respondents give free, frank and reliable information. Moreover the informants or respondents are free to give information as and when they want. Because of this freedom, the information that is provided is more dependable and reliable.

vi) Free from external influence: In questionnaire method, informants or respondents are free from external influences, as researcher is not present. They provide reliable, valid and meaningful information based on his knowledge, views and attitudes.

vii) Suitable for special type of responses: The information about certain problems can be best obtained through this method. For example, the research about sexual habits, marital relations, dreams etc. can easily be obtained by keeping the name of respondents anonymous.

viii) Less errors: Chances of errors are very low, because the supply of information is done by respondent himself.

ix) Originality: The informants are directly involved in the supply of information, so the method is more original.

x) Uniformity: The impersonal nature of questionnaire ensures uniformity from one measurement situation to another.

xi) Collection of information relevant to the objective: Through this method, the questionnaire are framed according to the object, hence data collection is also accordingly to that objective.

Disadvantage of Questionnaire Method

The method has the following disadvantages/limitations.

a) Lack of interest: Lack of interest on the part of respondents is very common. The respondents gets disinterested due to large number of questions.

b) Incomplete response: Some persons give answers which are so brief that the full meaning is incomprehensible.

c) Illegibility: Some persons write so badly that even they themselves find it difficult to read their own handwriting/.

d**) Useless in-depth research problems:** If a problem requires deep and long study, it can not be studied through this method.

e) Inelastic: This method is very rigid since no alteration may be introduced.

f) Prejudices and bias of the researcher influences the questions: Since researcher frames the questions his personal views, prejudices and the influence the questions and he instead of becoming objective and impersonal becomes bias and prejudicious.

g) Poor response and lack of reality: All the informants do not give answer or do not fill the questionnaire. There is a large percentage of those who do not send back the questionnaire. This makes the study unreliable.

h) The incompleteness of the form of questionnaire: Sometimes the questionnaire is itself incomplete and some of the important aspects about which the information is required are not given, hence data collected is neither reliable nor helpful for the study.

i) Lack of personal contact: There is no provision in this method for coming face to face with the respondent. This may result manipulation of replies by the respondents.

5.5.1 Precautions of Questionnaire and Collection of Data

The researcher or the framer of the questionnaire should take certain precautions while constructing or framing the questionnaire. For sake of emphasis these precautions are listed as follows:

1. Questions should be simple be simple, unambiguous and clear. The questions should not be ambiguous or couched in difficult words and unknown phraseology. The questions should be simple and suit the level of the intelligence of the informants. Very complicated questions be avoided. Unless it is done, the questionnaire is not likely to be useful.

2. Stimulating for the informants: Since answers to the questions are to be furnished by respondents. If the questions are not stimulating enough the informants are not likely to provide relevant answers the whole purpose shall be defeated.

3. Limited number of questions: If there are a large number of questions, the respondents shall lose interest in them. Generally the informants do not want to be bothered with too many questions. If they feel that they are being subjected to unnecessary work, they start giving unrelated and needless answers.

4. Technical and special words should be clearly explained: If the questionnaire contains certain technical and special terminology if should be clearly explained at the beginning of questionnaire.

5. Hypothetical questions should not be asked: While formulating the questionnaire the framer should always keep in mind not to include hypothetical questions. Answers to such questions are small but the investigator does not gain anything from them. Subjective and qualitative questions should be avoided as far as possible.

5.6 Summary

In this unit we focused on a very important construct in the field of measurement, scale and sampling. The four main types of measurement scales used in research include ratio, interval, nominal and ordinal levels of measurements. These specific attributes of these scales have been discussed throughout the paper. Fundamental examples of nominal scale of measurement include attributes like gender, color and yes on no answers. When data is categorized as male or female, it does not necessarily imply that males are superior to females. Ordinal measurement is a method that encompasses ranking information. Some types of data have specific attributes which may be classified into distinct ranks.

In most cases, the ranks used in ordinal scale of measurement provide clear distinction between the smallest and largest. As the name suggests, regular intervals separate categories or classes from one another. While investigating on the implementation of conflict resolution framework in community colleges, the interval scale may be used in different ways. The different ranks used in ordinal scales of do not necessarily reflect any measurable distance or disparities between variables. As a result, this scale of measurement relies heavily on approximation. On the other hand ratio scales are different in that they provide a predetermined zero spot for all characteristics. They facilitate for the inclusion of absent measurements in data collection.

Both probability and non-probability methods will be discussed in details in this chapter with advantages and disadvantages associated with each technique. It will also focus on what criteria should be kept in mind when selecting an appropriate sampling technique.

As detailed in the chapter sampling is quite a common phenomenon in decision making process. Before developing deeply into the sampling process one must

be aware of several basic constructs involved with sampling namely; population, target population, elements, sampling unit and sampling frame. Determining the final sample size for research involves various qualitative and quantitative considerations.

There are two basic techniques of selecting sample; probability sampling techniques and non-probability sampling techniques. Probability sampling techniques are more robust in comparison to non-probability sampling. Findings based on non-probability are hard to generalize to a wider population.

Probability sampling is sub-divided into simple random sampling, systematic sampling and stratified sampling. While being robust probability sampling techniques are resource intensive in terms of cost and time involved. Non-probability sampling is sub-divided into convenience sampling, judgement sampling and quota sampling. Non-probability sampling techniques are less costly and less time consuming however they have problems relating to selection bias also.

Selecting an appropriate sampling technique depends on various factors such as research objectives, available resources, knowledge of target population and scope of research, degree of accuracy and statistical analysis required for result interpretation.

Further the various methods of data collection i.e., observation method, personal interview. By schedules or questionnaire method, case study method, documentary sources of data method have been explained. Their relative advantages, limitations, applications, procedure and precautions in use have been described in detail.

5.7 Glossary

Measurement: It is the process of describing some property of a phenomenon of interest, usually by assigning numbers in a reliable and valid way.

Scaling: In the field of business research, measurement or scaling implies in conversion of the characteristics or qualitative data into quantitative data Sampling: is simply the process of learning about the population on the basis of a sample

drawn from it.

Non-Probability Sampling: The sample is chosen at the convenience of the consultant or to fulfil the demands of some predetermined purpose.

Observation: implies the collection of information by way of investigator's own observation, without interviewing the respondents

Population: implies the total number of items for which information is sought.

Probability Sampling: The units which constitute a probability sample are selected randomly, with each unit having a known chance of selection.

Sample: is that part of the universe which we select for the purpose of investigation. A sample exhibits the characteristics of the universe. The word sample literally means small universe.

Sampling Design: is a definite plan for obtaining a sample from the sampling frame. It refers to the technique or procedure of selecting some sampling units from which inferences about the population are drawn.

Sampling Errors: In sample survey only a small part of the universe or population is studies, as such there is every possibility that its result would differ from each other.

Sampling Frame: The term, sampling frame or population frame refers to the listing of all items in the population with proper identification under study.

Sampling Unit: Sampling units are the target population elements available for selection during the sampling process.

Simple Random sampling: of selection assures each individual element or units in universe and equal chance of being chosen.

Systematic sampling: is a variation of simple random sampling. It requires the universe or a list of its units may be ordered in such a way that each element of the universe can be uniquely indentified by its order.

Universe: implies the total of all the items or units of analysis in the field of proposed research.

Questionnaire: A measuring device used to query a population/sample in order to obtain information for analysis.

Respondent: An element or member of the population selected to be sampled.

5.8 Self Assessment Questions

1. Define the term sampling and explain its main elements. .

2. Explain the following terms:

    a) Sample

    b) Sampling frame

    c) Sampling design

    d) Sampling distribution

    e) Measurement and Scale

3. Differentiate between population and sample.

4 .Explain the main essentials of a sample. What factors to be considered in selecting the size of sampling?

5.Explain the difference between probabilities and non-probabilistic methods of sampling.

6. Distinguish between:

    a) Sample and stratified sampling method

    b) Systematic and cluster sampling

    c) Judgment and Quota sampling

    d) Random sampling and Judgment sampling

7. Difference between primary data and secondary

5.9 Lesson End Exercises

1. List the 2 situations where you would use (a) Convenience Sampling (b) Judgmental Sampling and (c) Quota Sampling.

2. Outline the various precautions of questionnaire and collection of data.
3. List few research pursuits where the observation method of data collection would be useful?
4. List few research pursuits where the interview method of data collection would be effective?
5. If you were researching on economic development of India, post-independence what kind of documents would you look into?
6. Explain the different sources of secondary data and the precautions in using secondary data.

5.10 Further Readings

Beri, G.C. Business Statistics, IIIrd Ed. Tata McGraw Hill Pvt. Ltd.; India.

Cooper, Donald R. & Schindler, Pamela S. Business Research Methods, Tata McGraw Hill Compnies; India.

Jhunjhunwala, B. Business Statistics, S Chand & Co. New Delhi.

Sachdeva, J.K. Business Research Methodology, Himalaya Publishing House; New Delhi.

Shajahan, S. Research Methods for Management, Jaico Publishing House, Delhi; India.

Singh, D. & Chaudhary F.S. Theory and Analysis of Sample Survey Designs, New Age International (P) Limited: New Delhi.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. Multivariate data analysis: A global perspective (7th ed.). Upper Saddle River: Pearson Education.

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. A primer on partial least squares structural equation modeling (PLS-SEM). Thousand Oaks: Sage.

5.11 References

Aaker, D. A., Kumar, V. & Day, G. S Marketing Research, 7th edn, John Wiley, New York.

Baker, M. J. Research for Marketing, Macmillan, London.

Boyd, H., Westfall, R. & Stasch, S. Marketing research: Text and cases. Boston: Irwin.

Bryman, A. Social Research Methods. London: Oxford University Press.

Creswell, J. W. (2009). Research design: Qualitative, quantitative, and mixed methods approaches (3rd ed.). Thousand Oaks, CA: Sage Publications.

Churchill, G. Marketing research (3rd ed.). Hinsdale, Illinois: Dryden Press.

Cozby, P. C. (2012). Methods in behavioral research (11th ed.).Boston:McGraw Hill Higher Education.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. Multivariate data analysis (6th ed.). New Jersey: Pearson Education International.

Hair, J. F., Celsi, M., Money, A., Samouel, P., & Page, M. Essentials of business research methods (2nd ed.). Armonk, NY: ME Sharpe.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. Journal of Management, 21(5), 967-988.

Kent, Ray Marketing Reearch: Approaches, methods and applications in Europe. London: Thomson.

Kline, R. B. Principles and practice of structural equation modeling (3rd ed.). New York: The Guilford Press.

Kothari, C.R. Research Methodology Methods and Techniques, New Age International (P) Limited: New Delhi.

Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. Educational and Psychological Measurement, 30, 607-

Kumar, V., Aaker, D. A., & Day, G. S. Essentials of marketing research New York: John Wiley & Sons, Inc.

Malhotra, N. K. Marketing Research: An Applied Orientation, 3rd edn, Prentice-Hall International, London

Malhotra, N.K. Marketing Research- An Applied Orientation, Pearson Education, Singapore.

Malhotra, N. K. Marketing Research (4th ed.). Harlow: Prentice Hall.

Malhotra, N. Marketing Research (4th ed.). New Jersey: Pearson.

Schmidt, M. J., & Hollensen, S. Marketing research an international approach. Harlow: Pearson Education.

Trochim, W. M. & Donnelly, J. P. (2006). Research methods knowledge base,OH: Cengage Learning

Willis, K. In-depth interviews, in the handbook of international market research techniques, Robin Birn, Ed. London: McGraw-Hill.