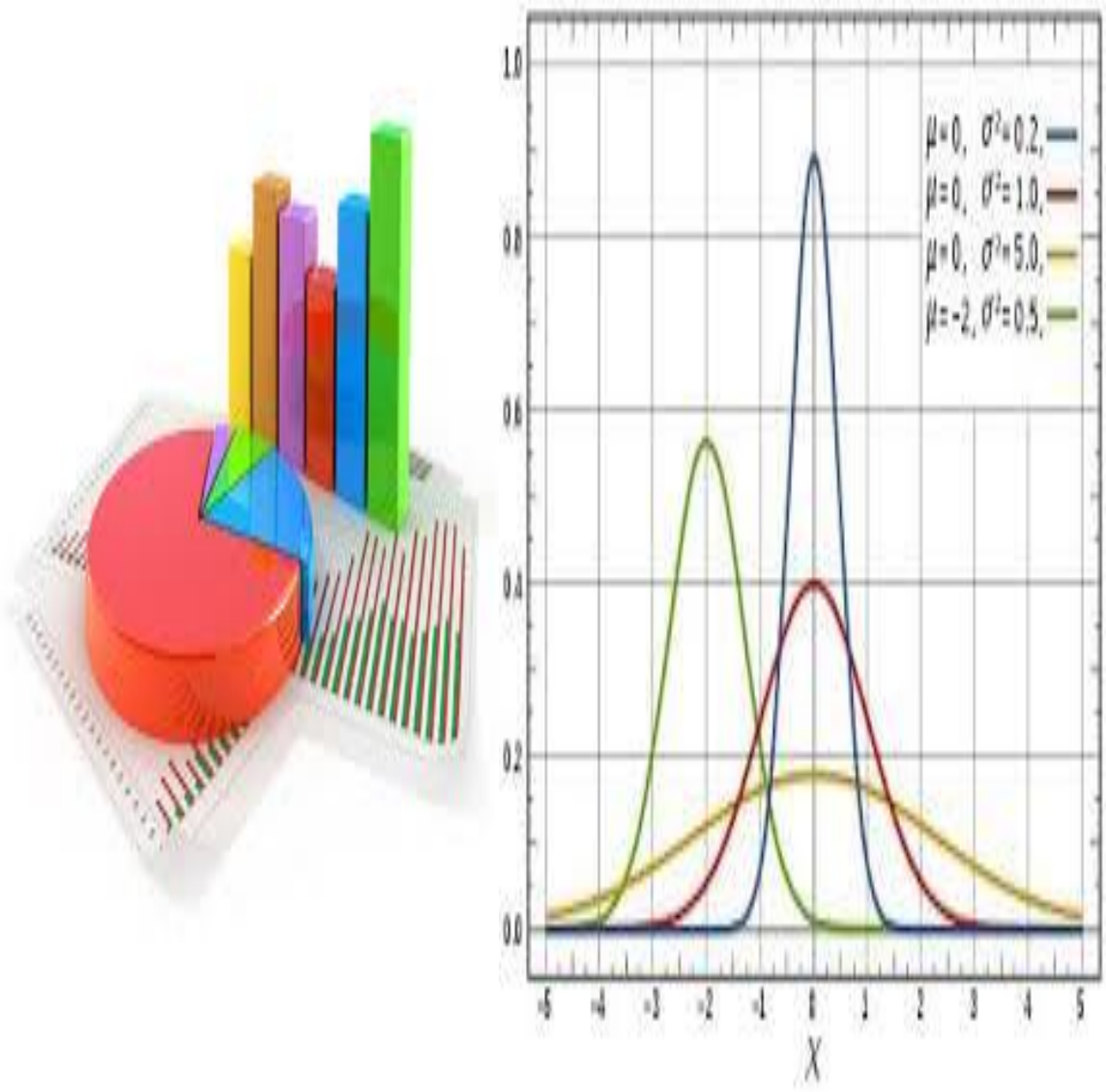
MCM-502

STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY



COMMERCE DEPARTMENT
SCHOOL OF MANAGEMENT STUDIES AND COMMERCE
UTTARAKHAND OPEN UNIVERSITY, HALDWANI

MCM-502

STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

UTTARAKHAND OPEN UNIVERSITY
University Road, Teenpani By-pass, Behind Transport Nagar, Haldwani-263139
Toll Free No..: 1800 180 4025
Fax No.: (05946)-264232, e-mail: info@uou.ac.in, som@uou.ac.in
http://www.uou.ac.in
www.blogsomcuou.wordpress.com

Board of Studies

Programme Coordinator and Editor

***Prof. Gagan Singh**, Department of Commerce, Uttarakhand Open University, Haldwani*

Units Written By	Unit No.
Dr. Surender Singh Kundu, Department of Commerce, Chaudhary Devi Lal University, Sirsa, Haryana	1-6,11-14
Dr. Umesh Maiya, Department of Management and Commerce, First Grade College, Udupi, Karnataka	7-10
Prof Ramendu Roy, Department of Management Studies, University of Allahabad	15-18
Dr. Anant Kumar Srivastava, Department of Management Studies, Shri Ram Murti College of Eng and Techno, Bareilly	19-22

Editing

Translation

Dr. H.K. Pant, Department of Management, Kumaun University, Bhimtal, Nainital	1-2, 7-18, 20
Dr. Brijesh Singh, Department of Commerce, Banaras Hindu University, Varanasi	3-6
Prof. Gagan Singh, Department of Commerce, Uttarakhand Open University, Haldwani	19
Dr. Aruna Shrivastava, Rajiv Gandhi College, Bhopal, M.P.	21
Dr. Dharmender Tiwari, Commerce Department, S.V. Government PG College, Lohaghat	22

ISBN :	**978-93-90845-87-3**
Copyright :	**Uttarakhand Open University**
Publication Year :	**2025**

Published by : Uttarakhand Open University, Haldwani, Nainital – 263139

Printed at : (……..)

MCM-502 STATISTICAL ANALYSIS AND RESEARCH METHODOLOGY

Block-1 **(Sampling and Sample Design)**

UNIT-1 (Introduction and Types of Sampling)

UNIT-2 (Point Estimation and Interval Estimation)

Block- 2 **(Probability and Theoretical Distribution)**

UNIT-3 (Approaches to Probability)

UNIT-4 (Theorems of Probability)

UNIT-5 (Binomial and Poisson Distribution)

UNIT-6 (Exponential, Beta & Normal Distribution)

Block-3 **(Hypothesis Testing and Significance Tests in Attributes & Variables)**

UNIT-7 (Procedure of Testing a Hypothesis)

UNIT-8 (Significance Test in Attributes)

UNIT-9 (Significance Test in Variables (Large Samples)

UNIT-10 (Significance Test in Variables (Small Samples)

Block-4 **(Regression, Correlation and Statistical Quality Control)**

UNIT-11 (Partial & Multiple Correlation)

UNIT-12 (Multiple Regression Analysis)

UNIT-13 (Types and Techniques of Statistical Quality Control)

UNIT-14 (Control Charts for Attributes and Variables)

Block-5 **(Non-Parametric Tests and Analysis of Variance)**

UNIT-15 (Chi-Square Test)

UNIT-16 (Sign Test & Median Test)

UNIT-17 (F-Test & Analysis of Variance (ANOVA)

UNIT-18 (**Multivariate Analysis Technique)**

Block-6 **(Research Methodology)**

UNIT-19 (Concepts, Approaches and Methods)

UNIT-20 (Research Design)

UNIT-21 (Measurement and Scaling Techniques)

UNIT-22 (Interpretation, Report Writing & Computer Applications in Research)

UNIT 1 INTRODUCTION AND TYPE OF SAMPLING

Structure

OBJECTIVES

After studying this unit, you will be able to understand:

- The basic concepts of sampling;
- Types of data collection methods;
- Types of sampling methods;
- Sampling and non-sampling errors; and the concept of sampling distribution.

1.1 INTRODUCTION

In all the spheres of life (such as Economic, Social and Business) the need for statistical investigation and data analysis is rising day by day. There are two methods of collection of statistical data: (i) Census Method, and (ii) Sample Method. Under census method, information relating to the entire field of investigation or units of population is collected; whereas under the sampling method, rather than collecting information about all the units of population, information relating to only selected units is collected. In modern times sampling method is an important and popular method of statistical inquiry. Besides, the economic and business world, this method is widely used in daily life. For example, a housewife comes to know of the coating of a whole lot of rice by observing two or three grains only. A doctor tests the blood of a patient by examining one or two drops of blood only. In the same way, you learn about the quality of a commodity while buying items of daily use like wheat, rice, pulses, etc. by observing the sample or specimen. In factories, a statistical quality controller inspects the quality of items by examining a few units produced in the factory. A teacher gets knowledge about the efficacy of his teaching by asking questions to a few students. In reality, there is scarcely any area left where sampling method is not used.

1.2 BASIC CONCEPTS OF SAMPLING

Before you make a detailed study the all aspects of sampling, you should understand some basic concepts related to them, which are as follows:

1.2.1 Universe or Population: In statistics, universe or population means an aggregate of items about which you obtain information. A universe or population means the entire field under investigation about which knowledge is sought. For example, if you want to collect information about the average monthly expenditure of all the 2,000 students of a college, then the entire kinds (i) Finite and (ii) Infinite. In a finite population, number of items is definite such as, number of students or teachers in a college. On the other hand, an infinite population has infinite number of items e.g., number of stars in the sky, number of water drops in an ocean, number of leaves on a tree or number of hairs on the head.

1.2.2 Sample: A part of population is called sample. In other words, selected or sorted units from a population are known as sample. In fact, a sample is that part of the population which you select for the purpose of investigation. For example, if an investigator selects 200 students from 2000 students of a college who represent all of them, then these 200 students will be termed as a sample. Thus, sample means some units selected out of a population which represent it.

1.3 DATA COLLECTION METHODS

There are two methods to collect statistical data, as follows:

1.3.1 Census Method: Census method is that method in which information or data is collected from every unit of the population relating to the problem under investigation and conclusions are drawn on their basis. This method is also called as Complete Enumeration Method. For example, some information (like Monthly Expenditure, Average Height, Average Weight, etc.) is to be collected regarding 2000 students of a college. For that purpose, if you collect data by inquiring each and every student of the college then this method will be called as census method. In this example, the whole college i.e., all 2000 students, will be considered as a population and every student as an individual will be called the unit of the population. The population in India is conducted after every ten years by using census method.

1.3.1.1 Merits of Census Method

(i) **Reliable and Accurate Data:** Data obtained by census method have more reliability and accuracy because in this method data are collected by contacting each and every unit of the universe.

(ii) **Extensive Information:** This method gives detailed information about each unit of the universe. For example, Indian population census does not only provide the knowledge about the number of persons but also information about their age, occupation, income, education, marital status, etc.

(iii) **Suitability:** This method is more suitable for the population with limited scope and diverse characteristics. Use of this method is also appropriate where intensive study is desired.

1.3.1.2 Demerits of Census Method

(i) **More Expensive:** Census method is an expensive one. More money is needed for it as information is collected from each unit of the population. This is why this method is used by Government mostly for very important issues like Census, etc.

(ii) **Time consuming:** This method involves much time for data collection because data are collected from each and every unit of the population. This results in delay in making statistical inferences.

(iii)**More use of Labour:** This method of data collection also involves very much labour. For this the enumerators in a large number are required.

(iv)**Not Suitable for Specific Problems:** This method is not suitable relating to certain specific problem and infinite population. For example, if the population is infinite or items of the population are perishable or very complex in nature, then the census method is not suitable.

1.3.2 Sampling Method: Sampling method is that method in which data is collected from a sample of items selected from the population and conclusions are drawn from them. For example, if a study is to be made regarding monthly expenditure of 2000 students of a college, then instead of collecting information from each student of the college, if you collect information by selecting some students like 100, then this will be called Sampling Method. Based on the sampling method, it is possible to study the monthly expenditure of all the students of the college. Sampling method has three main stages (i) to select a sample (ii) to collect information from it and (iii) to make inferences regarding the population.

1.3.2.1 Merits of Sampling Method

(i) **Less Expensive:** The Sampling method is less expensive. It saves money and labour because only a few units of the population are studied.

(ii) **Time Saving:** In sampling method, data can be collected more quickly as these are obtained from some items of the universe. Thus, much time is saved.

(iii) **Intensive Study:** As the number of items are less in the sampling method, they can be intensively studied.

(iv) **Organizational Convenience:** In this method, research work can be organized and executed more conveniently. More skilled and competent investigators can be appointed.

(v) **More Reliable Results:** If the sample is selected in such a manner that it represents totally the entire universe, then the results derived from it will be more accurate and reliable.

(vi) **More Scientific:** Sampling method is more scientific because data can be compared with other samples.

(vii) **Single Method:** In some fields where inquiry by census method is impossible, then in such a situation sampling method alone is more appropriate. If the population is infinite or too widespread or of perishable nature, then only the sampling method is used in such cases.

1.3.2.2 Demerits of Sampling Method

(i) **Less Accurate:** Sampling method has less accuracy because rather than making inquiry about each unit of the universe, partial inquiry relating to some selected units only is made.

(ii) **Wrong Conclusions:** If method of selecting a sample is not unbiased or proper caution has not been taken, then results are definitely misleading.

(iii) **Less Reliable:** Compared to census method, there is more likelihood of the bias of the investigator, which makes the results less reliable.

(iv) **Need for Specified Knowledge:** This is a complex method as specialized knowledge is required to select a sample.

(v) **Lack of Suitability:** The sampling method is not suitable in case of heterogeneity among the units of a population.

1.3.3 Difference between Census and Sample Method

The main differences between the census method and the sampling method are as follows:

(i) **Scope:** In census method, all items relating to a universe are investigated, whereas in sampling method only a few items are inquired.

(ii) **Cost:** Census method is expensive from the point of view of time, money and labour whereas sampling method economizes on them.

(iii) **Field of Investigation:** Census method issued in investigation with limited field whereas sampling method is used for investigations with large field.

(iv) **Homogeneity:** Census method is useful where units of the population are heterogeneous whereas sampling method proves more useful where population units are homogenous.

(v) **Type of Universe:** In such fields where the study of each and every unit of the universe is necessary, the census method is more appropriate. On the contrary, when the population is infinite or vast or liable to be destroyed as a result of complete enumeration, then the sampling method is considered to be more appropriate.

1.4 SAMPLING METHODS

The method of selecting a sample out of a given population is called sampling. In other words, sampling denotes the selection of a part of the aggregate statistical material with a view of obtaining information about the whole. Now a day, there are various methods of selecting sample from a population in accordance with various needs.

1.4.1 Probability Sampling Methods:

1. Simple Random Sampling
2. Stratified Random Sampling
3. Systematic Random Sampling.
4. Multistage Random Sampling
5. Cluster Sampling

1.4.2 Non-Probability Sampling Methods:

1. Judgment Sampling

2. Quota Sampling

3. Convenience Sampling

4. Extensive Sampling

1.4.1 Probability Sample Methods

Probability sampling methods are such methods of selecting a sample from the population in which all units of the universe are given equal chances of being included in the sample.

There are various variants of probability sampling methods, which are given below:

1 Simple Random Sampling: Simple random sampling is that method in which each item of the universe has an equal chance of being selected in the sample. Which item will be included in the sample and which not? such a decision is not made by an investigator on their will but selection of the units is left to chance. According to random sampling, there are two methods of selecting a random sample:

(a) Lottery Method: In this method, each unit of the population is named or numbered, which is marked on a separate piece of paper. Such chits are folded and put into some urn or bag. Thereafter, as many chits are made as selected by some person, as many units are to be included in a sample.

(b) Tables of Random Numbers: Some experts have constructed random number tables. These tables help in the selection of a sample. Of all such various tables, Tippett's Tables are most famous and are in use. Tippett has constructed a four-digit table of 10,400 numbers by using numbers as large as 41,600. In this method, first of all, all the items of a population are written serially. Thereafter by making use of Tippett's tables, in accordance with the size of the sample, numbers are selected. The selection of a sample with the help of Tippett's table can be made clear by an example:

An Extract of Tippett's Table

2952	6641	3992	9792	7979	5911
3170	5224	4167	9525	1545	1396
7203	4356	1300	2693	2370	7483

3408	2762	3563	6107	6913	7691
0560	5246	1112	9025	6008	8127

For example, suppose 12 units are to be chosen out of 5000 units. With Tippett's table, to decide such units, firstly 5000 units will be serially ordered from 1 to 5000 and then from Tippett's table, 12 numbers will be chosen from the beginning which will be less than 5000. These 12 numbers are follows:

2952	4167	4356	2370
3992	1545	1300	3408
3170	1396	2693	2762

The items of such serial numbers will be included in the sample. If units of the population are less than 100, then 4-digit random numbers will be made compact into two-digit numbers, and then such two-digit numbers will be selected. Like as to select 6 units out of 60 units, the units with serial numbers 29, 39, 31, 41, 15 and 13 will be included in the sample.

Merits

(i) There is no possibility of personal prejudice in this method. In other words, this method is free from personal bias.

(ii) Under this method, every unit of the universe gets an equal chance of being selected.

(iii) The use of this method saves time, money and labour.

Demerits

(i) If the sample size is small, then sample is not adequately represented.

(ii) If the universe is very small, then this method is not suitable.

(iii) If some items of the universe are so important that their inclusion in the sample is essential, then this method will not be appropriate.

(iv) This method will not be appropriate when population has units with diverse characteristics.

2. Stratified Random Sampling: This method is used when units of the universe are heterogeneous rather than homogeneous. Under this method, first of all, units of the population are divided into different strata in accordance with their characteristics. Thereafter by using random sampling, sample items are selected from each stratum. For example, if 150 students are to be selected out of 1500 students of a college, then firstly the college students will be divided into three groups based on Arts, Commerce, and Science. Suppose there are 500, 700, 300 students respectively in three faculties and 10 per cent sample is to be taken, then based on random sampling 50, 70 and 30 students, respectively will be selected by using random sampling. Thus, this method assumes equal representation for each class or group, and all the units of the universe get an equal chance of being selected in the sample.

Merits

(i) There is a greater likelihood of representation of units in this method.
(ii) A comparative study based on facts at different strata is possible under this method.
(iii) This method has more accuracy.

Demerits

(i) This method has limited scope because this method can be adopted only when the population and its different strata are known.
(ii) There is a possibility of prejudice if the population is not properly stratified.
(iii) If the population is too small in size, it is difficult to stratify it.

3 Systematic Random Sampling: In this method, all the items of the universe are systematically arranged and numbered, and then sample units are selected at equal intervals. For example, if 5 out of 50 students are to be selected for a sample, then 50 students would be numbered and systematically arranged. One item of the first 10 would be selected at random. Subsequently, every 10^{th} item from the selected number will be selected to frame a sample. If the first selected number is 5^{th} item, then the subsequent numbers would be 15^{th}, 25^{th}, 35^{th} and 45^{th}.

Merits

(i) It is a simple method. Samples can be easily obtained by it.
(ii) This method involves very little time in sample selection and results are almost accurate.

Demerits

(i) In this method, each unit does not stand the equal chances of being selected because only the first unit is selected on random sampling basis.
(ii) If all the units are different in characteristics, then results will not be appropriate.

4 Multi-stage Random Sampling: When the sampling procedure passes through many stages, then it is known as multi-stage sampling. In this method, firstly, the entire universe or population is divided into stages or sub-stages. At each stage, some units are selected on a random sampling basis. Thereafter, these units are sub-divided and based on random sampling, again, some sub-units are selected. Thus, this goes on with sub-division further and selection. For example, for a study regarding Adult Education in a State, first, some districts will be selected on a random basis. Thereafter, out of the selected districts, some tehsils and out of the tehsils, some wards and out of the wards, some households will be selected from whom the inquiry will be made concerning the problem at hand.

Merits

(i) This is the best method of studying a universe or population on a regional basis.
(ii) This method is suitable for those problems where decisions on the basis of the sample alone can't be taken.

Demerits

(i) This method requires many tests to correctly estimate the level of accuracy, which involves a lot of time and labour.
(ii) In this method, the level of estimated accuracy level is pre-decided, which does not seem logical.

5. Cluster Sampling: In this method, simply the universe is simply divided into many groups called cluster, and out of which a few clusters are selected on a random basis, and then the clusters are completely enumerated. This method is usually applied in industries, e.g., in the pharmaceutical industry, a machine produces medicine tablets in batches of a hundred each, then, for quality inspection, a few randomly selected batches are examined.

1.4.2 Non-Probability Sampling Methods

Non-probability sampling methods are those methods in which the selection of units is made on the basis of convenience or judgment of the investigator rather than on the basis of probability or chance. In such methods, selection of units is made in accordance with the specific objectives and convenience of the investigator.

1. Judgement Sampling: Under this method, the selection of the sample items depends exclusively on the judgement of the investigator. In other words, the investigator exercises his/her judgment in the choice and includes those items in the sample which he/she thinks are most typical of the universe with regard to the characteristics under study. For example, if a sample of 20 students is to be selected from a class of 80 students for analyzing the spending habits of the 10 students, the investigator would select 20 students, who in his/her opinion are representative of the class.

Merits

(i) This method is less expensive.
(ii) This method is very simple and easy.
(iii) This method is useful in those fields where almost similar units exist or some units are too important to be left out of the sample.

Demerits

(i) There is a greater chance of the investigator's own prejudice in this method.
(ii) This method is not very accurate and reliable.

2. Quota Sampling: In this method, the investigators are assigned definite quotas according to some criteria. They are instructed to obtain the required number to fill in each quota. The investigators select the individuals (i.e., sample items) to collect information on their personal judgements within the quotas. When all or a part of the whole quota is not available or approachable, the quota is completed by supplementing new respondents. Quota sampling is a type of judgement sampling.

Merits

(i) In this method, there is greater chance of important units being included.
(ii) Statistical inquiry is more organized in this method on account of the units of the quotas being fixed.

Demerits

(i) The possibility of prejudice shall remain.

(ii) There is a greater likelihood of sampling error in this method.

3. Convenience Sampling: In this type of non-probability sampling, the choice of the sample is left completely to the convenience of the investigator. The investigator obtains a sample according to the list of the teachers from the college prospectus and gets feedback from them regarding his/her publication. This method is less expensive and simpler but, is unscientific and unreliable. This method results in more dependence on the enumerators. This method is appropriate for sample selection where the universe or population is not clearly defined or a list of the units is not available, or sample units are not clear in themselves.

4.Extensive Sampling: In this method, sample size is taken almost as big as the population itself, e.g., 90 per cent of the population. Only those units are left out for which data collection is very difficult or almost impossible. Due to very large sample size, the method has greater level of accuracy. Intensive study of the problem becomes possible but this method involves heavy resources at disposal.

1.5 SAMPLING AND NON-SAMPLING ERRORS

The choice of a sample, though, may be made with utmost care, involves certain errors which may be classified into two types: (i) Sampling Errors, and (ii) Non-Sampling Errors. These errors may occur in the collection, processing, and analysis of data.

1.5.1 Sampling Errors

Sampling errors are those that arise due to the method of sampling. Sampling errors arise primarily due to the following reasons:

(i) Faulty selection of the sampling method.

(ii) Substituting one sample for another due to the difficulties in collecting the sample.

(iii) Faulty demarcation of sampling units.

(iv) Variability of the population, which has different characteristics.

1.5.2 Non-Sampling Errors

Non-sampling errors are those that occur due to human factors, which always vary from one investigator to another. These arise due to any of the following factors:

(i) Faulty planning.
(ii) Faulty selection of the sample units.
(iii) Lack of trained and experienced staff who collect the data.
(iv) Negligence and non-response on the part of the respondent.
(v) Errors in compilation.
(vi) Errors due to wrong statistical measures.
(vii) Framing of a wrong questionnaire.
(viii) Incomplete investigation of the sample survey.

1.6 CONCEPTS OF SAMPLING DISTRIBUTION

Now, you should understand some basics of sampling distribution, as follows:

1.6.1 Sampling Distribution: The purpose of selecting and studying a sample from the population is to estimate or make inference about some population characteristics. In this process, the knowledge of the sampling distribution is of vital importance.

1.6.2 Parameter (s): Any statistical measures computed from the population data is known as parameter. Thus, population mean, population standard deviation, population variance, population proportion, etc., are all parameters; parameters are denoted by the Greek letter such as μ, σ, σ^2 and P, respectively.

1.6.3 Statistic (s): Any statistical measure computed from sample data is known as statistic. Thus, sample mean, sample standard deviation, sample variance, sample proportion, etc., are all statistics; statistics are denoted by Roman letters such as $\bar{x}$, s, s^2 and p, respectively.

1.6.4 Sampling with and without replacement: Sample is a procedure of selecting a sample from the population. Sampling may be done with or without replacement. Sampling where each unit of a population may be chosen more than once is called sampling with replacement. If each unit cannot be chosen more than once, it is called sampling without replacement. In case of sampling with replacement, the total number of possible samples

each of size n drawn from a population of size N is N^n. but if the sampling is without replacement, the total number of possible samples will be ${}^{N}C_n = m$ (say).

1.7 SAMPLING DISTRIBUTION OF A STATISTIC

Sampling distribution constitutes the theoretical basis of statistical inference and is of considerable importance in business decision making. Sampling distribution of a statistic is the frequency distribution which is formed with various values of a statistic computed from different samples of the same size drawn from the same population. Suppose you draw all possible samples of size *n* from the population (N) with or without replacement. For each possible sample drawn from the population, you compute a statistic such as mean, median, standard deviation, variance, etc. The set of all possible values of a statistic is then classified and grouped into a frequency distribution (or probability distribution). The distribution so obtained is called the sampling distribution of a statistic. You could have various sampling distributions depending upon the nature of the statistic you have computed, e.g., if the particular statistic computed is the sample mean then the distribution is called sampling distribution of mean. If you compute variance of each sample then it is called the sampling distribution of variance. Similarly, you could have sampling distributions of proportion, median, standard deviation, etc.

An important property of the sampling distribution of a statistic is that if random samples of large size (n>30) are taken from a population which may be normally distributed or not, then the sampling distribution of the statistic will approach to normal distribution.

1.8 STANDARD ERROR OF A STATISTIC

The standard deviation of the sampling distribution of a statistic is known as the standard error (S. E.) of a statistic. As there are various types of sampling distributions, you could have various types of standard errors depending on the nature of sampling distribution. The standard deviation instead of using the term standard deviation for measuring variation, you may use a new term called standard error of mean. The standard error of mean measures the extent to which the sample mean differs from the population mean. Thus, the basic difference between the standard deviation and standard error of mean is that the first measures the extent to which the individual items differ from the central value and the last measures the extent to which individual sample mean differ from the population mean. Like

the standard error of the means, you could have the standard error of the median, standard deviation, proportion, variance, etc.

The standard error is used in a large number of problems, which are discussed as follows:

(i) Reliability of a sample: The standard error gives an idea about the reliability and precision of a sample. That is, it indicates how much the estimated value differs from the observed values. The greater the standard error, the greater is the deviation between the estimated and observed values and lesser is the reliability of a sample. The smaller the standard error, the smaller is the deviation between the estimated and observed values and greater is the reliability of a sample.

(ii) Tests of significance: The standard error is also used to test the significance of the various results obtained from small and large samples. In case of large sample, if the difference between the observed and the expected value is greater than 1.96 S.E. then you reject the hypothesis at 5% and conclude that sample differs widely from the population. But if the difference between the observed and the expected value is greater than 2.58 S.E. then you reject the null hypothesis at 1% and conclude that the sample differs widely from the population.

(iii) To determine the confidence limits of the unknown population mean: The standard error enables us to determine the confidence limits within which a population parameter is expected to lie with a certain degree of confidence. The confidence limits of the unknown population mean μ are given by.

Large Sample	*Small Sample*
95% *confidence limits for* μ	95% *confidence limit for* μ
$\bar{x} - 1.96\ S.E.\ and\ \bar{x} + 1.96\ S.E.$	$\bar{x} \pm\ t_{.05}\ S.E.$
99% *confidence limits for* μ	99% *confidence limit for* μ
$\bar{x} - 2.58\ S.E.\ and\ \bar{x} + 2.58\ S.E.$	$\bar{x} \pm t_{.01}\ S.E.$

1.9 SAMPLING DISTRIBUTION OF MEANS

It is important to note that sampling distribution is widely used in sampling theory. Draw all possible samples of size n with or without replacement from a population of size N with mean μ and variance σ^2. For each possible sample drawn from the population, you compute

the mean $\bar{x}$ of each sample. The mean will vary from sample to sample. The set of all possible means obtained from different samples are called the sampling distribution of means.

The following are the important properties of the sampling distribution of means:

(i) The mean of the sampling distribution of means is equal to the population mean (μ).

Symbolically, $\mu_{\bar{x}} = \mu$ or $E(\bar{x}) = \mu$

This property can be proved as follows:

Let $x_1, x_2, \ldots x_n$ present a random sample (with replacement) of size n from a finite population of size N having its mean μ and variance σ^2, then

$$\bar{x} = \frac{x_1 + x_1 + \cdots x_n}{n}$$

$$E(\bar{x}) = E\left[\frac{\Sigma x}{n}\right] = E\left[\frac{x_1 + x_1 + \cdots + x_n}{n}\right]$$

$$= \frac{1}{n}\{E(x_1) + E(x_2) + \cdots + E(x_n)\}$$

$$= \frac{1}{n}\{\mu + \mu + \mu + \cdots + \mu\} = \frac{1}{n}.n\mu = \mu$$

Thus, the mean of the sampling distribution of means is equal to the population mean.

(ii) The standard error of the sampling distribution of means is obtained as:

$$S.E._{\bar{x}} \ or \ \sigma_{\bar{x}} = \frac{S.D. of \ Population}{\sqrt{Size \ of \ the \ sample}} = \frac{\sigma}{\sqrt{n}}$$

This property can be proved as follows:

$$Var(\bar{x}) = Var\left(\frac{\Sigma x}{n}\right) = Var\left(\frac{x_1 + x_1 + \cdots + x_n}{n}\right)$$

$$= \frac{1}{n^2}[Var(x_1) + Var(x_2) + \cdots Var(x_n)]$$

$$= \frac{1}{n^2}[\sigma^2 + \sigma^2 + \cdots \sigma^2]$$

$$= \frac{1}{n^2} . n\sigma^2 = \frac{\sigma^2}{n}$$

Where, σ^2 is the population variance, x is the sample size.

Because, n $>$ 1 , obviously, $\frac{\sigma^2}{n} < \sigma^2 \Rightarrow V(\bar{x}) <$ Population variance.

$$\therefore \qquad S.E._{\bar{x}} \ or \ \sigma_{\bar{x}} = \sqrt{Var(\bar{x})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

This formula holds only when sampling is with replacement.

Note: When the population is finite and the samples are drawn without replacement then $S.E._{\bar{x}}$ is obtained as:

$$S.E._{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

(iii) The sampling distribution of means is approximately a normal distribution with mean μ and variance σ^2, provided the sample is large (n>30).

(iv) The following formula is used to find the probability of the sampling distribution of means.

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

Example 1.1: Consider a population consisting of three values: 2, 5, and 8. Draw all possible samples of size 2 with replacement from the population. Construct a sampling distribution of means. Also, find the mean and standard error of the distribution.

Solution: The population consists of three values. The total number of possible samples of size 2 drawn with replacement are $N^n=3^2=9$. All possible random samples and their sample mean are shown is the following table:

Sample No.	**Sample Values**	**Sample Mean $\bar{x}$**

1.	(2,2)	$\frac{1}{2}(2+2)=2$
2.	(5,2)	$\frac{1}{2}(5+2)=3.5$
3.	(8,2)	$\frac{1}{2}(8+2)=5$
4.	(2,5)	$\frac{1}{2}(2+5)=3.5$
5.	(5,5)	$\frac{1}{2}(5+5)=5.0$
6.	(8,5)	$\frac{1}{2}(8+5)=6.5$
7.	(2,8)	$\frac{1}{2}(2+8)=5.0$
8.	(5,8)	$\frac{1}{2}(5+8)=6.5$
9.	(8,8)	$\frac{1}{2}(8+8)=8.0$

On the basis of the means ($\bar{x}$) of all the 6 possible samples, the sampling distribution of means is given below:

Sample Means ($\bar{x}$)	f	$f(\bar{x})$	$d=(\bar{x})-\mu_{\bar{x}}$	d^2	fd^2
2	1	2	-3	9	9
3.5	2	7	-1.5	2.25	4.50
5.0	3	15	0	0	0
6.5	2	13	1.5	2.25	4.50
8.0	1	8	+3	9.0	9.0
	$\Sigma f=9$	$\Sigma f(\bar{x})=45$			$\Sigma fd^2=27$

Mean of the Sampling Distribution of Means

$$\mu_{\bar{x}} = \frac{\Sigma f(\bar{x})}{\Sigma f} = \frac{45}{9} = 5$$

Variance of the Sampling Distribution of Means

$$Var(\bar{x}) = \frac{\Sigma f(\bar{x} - \mu_{\bar{x}})^2}{\Sigma f} = \frac{\Sigma f d^2}{\Sigma f} = \frac{27}{9} = 3$$

Hence,

$$S.E._{\bar{x}} = \sigma_{\bar{x}} = \sqrt{3} = 1.732$$

Example 1.2: Construct a sampling distribution of the sample means from the following population:

Population Unit:	1	2	3	4
Observation:	22	24	26	28

If random sample of size 2 is taken from population without replacement, then find the mean and standard error of the distribution.

Solution: The population consists of four values (22, 24, 26, 28). The total number of possible sample of size 2 drawn without replacement is ${}^4C_2$=6. All the possible random samples and their sample means are shown in the table given below:

Sample No.	**Sample Values**	**Sample Mean $\bar{x}$**
1.	(22,24)	$\frac{1}{2}(22+24) = 23$
2.	(22,26)	$\frac{1}{2}(22+26) = 24$
3.	(22,28)	$\frac{1}{2}(22+28) = 25$
4.	(24,26)	$\frac{1}{2}(24+26) = 25$

5.	(24,28)	$\frac{1}{2}(24+28)=26$
6.	(26,28)	$\frac{1}{2}(26+28)=27$

On the basis of the means ($\bar{x}$) of all the 6 samples without replacement, the sampling distribution of means is given below:

Sampling Distribution of Means without Replacement

Sample Means ($\bar{x}$)	f	$f(\bar{x})$	$d=(\bar{x})-\mu_{\bar{x}}$	d^2	fd^2
23	1	23	-2	4	4
24	1	24	-1	1	1
25	2	50	0	0	0
26	1	26	1	1	1
27	1	27	2	4	4
	$\Sigma f=6$	$\Sigma f(\bar{x})=150$			$\Sigma fd^2=10$

Mean of the Sampling Distribution of Means

$$\mu_{\bar{x}}=\frac{\Sigma f(\bar{x})}{\Sigma f}=\frac{150}{6}=25$$

Variance of the Sampling Distribution of Means

$$Var(\bar{x})=\frac{\Sigma f(\bar{x}-\mu_{\bar{x}})^2}{\Sigma f}=\frac{\Sigma fd^2}{\Sigma f}=\frac{10}{6}=\frac{5}{3}$$

Hence,

$$S.E._{\bar{x}}=\ \sigma_{\bar{x}}=\sqrt{Var(\bar{x})}=\sqrt{\frac{5}{3}}=1.29$$

Alternatively, The sampling distribution of means can also be written in terms of probability as below:

Sample Means ($\bar{x}$)	23	24	25	26	27
Probability (p)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Since, 25 occurs twice, its probability of occurrence is $\frac{2}{6}$. Each of the other sample mean occurs only with probability $\frac{1}{6}$.

Mean of the Sampling Distribution of Means

$$E(\bar{x}) = \Sigma p\bar{x} = \frac{1}{6} \times 23 + \frac{1}{6} \times 24 + \frac{2}{6} \times 25 + \frac{1}{6} \times 26 + \frac{1}{6} \times 27$$

$$= \frac{1}{6}.[23 + 24 + 50 + 26 + 27] = \frac{150}{6} = 25$$

Variance of the Sampling Distribution of Means

$$Var(\bar{x}) = E(\bar{x}^2) - [E(\bar{x})]^2$$

$$E(\bar{x}^2) = 23^2 \times \frac{1}{6} + 24^2 \times \frac{1}{6} + 25^2 \times \frac{2}{6} + 26^2 \times \frac{1}{6} + 27^2 \times \frac{1}{6}$$

$$= \frac{1}{6}.[529 + 576 + 1250 + 676 + 729]$$

$$= \frac{3760}{6} = 626.17$$

$$Var(\bar{x}) = E(\bar{x}^2) - [E(\bar{x})^2] = 626.17 - 625 = 1.67$$

Hence,

$$S.E._{\bar{x}} = \sqrt{Var(\bar{x})} = \sqrt{1.67} = 1.29$$

Example 1.3: A population consists of four elements: 3, 7, 11, and 15. Consider all possible samples of size two which can be drawn with replacement from this population Find (i) the population mean μ (ii) the population variance $\sigma^{2,}$ (iii) the mean of the sampling distribution of means, (iv) standard error (or S.D.) of the sampling distribution of means. Verify (iii) and (iv) by using (i) and (ii) and by use of a suitable formula.

Solution: (i) μ = population mean $= \frac{\Sigma X}{N} = \frac{3+7+11+15}{4} = \frac{36}{4} = 9$

(ii) σ^2 = population variance $= \frac{\Sigma(X-\mu)^2}{N} = \frac{(-6)^2+(-2)^2+(2)^2+(6)^2}{4} = \frac{80}{4} = 20$

$\therefore$ $\sigma = S.D. = \sqrt{20}$

(iii) All possible random sample of size two with replacement is

$N^n = 4^2 = 16$ and their sample means are shown in the

following table:

Sample No.	**Sample Values**	**Sample Mean $\bar{x}$**	**Sample No.**	**Sample Values**	**Sample Mean $\bar{x}$**
1	(3,3)	3	9	(11,3)	7
2	(3,7)	5	10	(11,7)	9
3	(3,11)	7	11	(11,11)	11
4	(3,15)	9	12	(11,15)	13
5	(7,3)	5	13	(15,3)	9
6	(7,7)	7	14	(15,7)	11
7	(7,11)	9	15	(15,11)	13
8	(7,15)	11	16	(15,15)	15

On the basis of the means ($\bar{x}$) of all the 16 samples without replacement, the sampling distribution of ($\bar{x}$) can be written as:

Sampling Distribution of Means without Replacement

Sample Means $(\bar{x})$	f	$f(\bar{x})$	$d = (\bar{x}) - \mu_{\bar{x}}$	d^2	fd^2
3	1	3	-6	36	36
5	2	10	-4	16	32
7	3	21	-2	4	12
9	4	36	0	0	0
11	3	33	+2	4	12
13	2	26	+4	16	32
15	1	15	+6	36	36
	$\Sigma f = 16$	$\Sigma f(\bar{x}) = 144$			$\Sigma fd^2 = 160$

Mean of the Sampling Distribution of Means

$$\mu_{\bar{x}} = \frac{\Sigma f(\bar{x})}{\Sigma f} = \frac{144}{16} = 9$$

Variance of the Sampling Distribution of Means

$$Var(\bar{x}) = \frac{\Sigma f(\bar{x} - \mu_{\bar{x}})^2}{\Sigma f} = \frac{160}{16} = 10$$

Hence,

$$S.E._{\bar{x}} = \sigma_{\bar{x}} = \sqrt{Var(\bar{x})} = \sqrt{10}$$

Using the formula, $\mu_{\bar{x}} = \mu$ and $V(\bar{x}) = \frac{\sigma^2}{n}$, you get the mean of the sampling distribution of means $\mu_{\bar{x}} = \mu = 9$ and variance of the sampling distribution of means $(\sigma_{\bar{x}})^2 = \frac{\sigma^2}{n} = \frac{20}{2} = 10$.

Hence, the results of (iii) and (iv) are verified by using the results of (i) and (ii).

1.10 LAW OF LARGE NUMBERS AND CENTRAL LIMIT THEOREM

Law of Large Numbers and the Central Limit Theorem both serve the basis for the development of sampling distribution of a statistic.

1.10.1 Law of Large Numbers: The law of large numbers states that as the sample size increases, the sample mean will be closer and closer to the population mean. It does not guarantee that if the sample size is increased sufficiently, the sample mean will be equal to the population mean. There are two implications of the law of large numbers (i) the difference between sample mean and population mean can be reduced by increasing the sample size, and (ii) variation from one sample mean to another sample mean (of the same size) also decreases as the size of the sample increases.

1.10.2 Central Limit Theorem: It is widely used in the field of estimation and inference. This theorem states that if you select random sample of large size n from any population with mean μ and standard deviation σ and compute the mean of each sample, then the sampling distribution of mean $\overline{x}$ approaches normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. This is true even if the population itself is not normal. The utility of this theorem is that it requires virtually no conditions on the distribution pattern of the population.

1.11 SUMMARY

In all, there are two methods of collection of statistical data: (i) Census Method, and (ii) Sample Method. Presently, the sampling method is an important and popular method of statistical inquiry. By an investigator, the sample is chosen by him/her by using different sampling methods to save time and money. Therefore, it may be stated that the sampling method is more useful in all spheres of life.

1.12 GLOSSARY

Non-probability sampling methods: These are the methods in which the selection of units is based on probability, but based on convenience or the decision of the researcher.

Aggregate: In statistics, aggregate means the collected(total) objects/things about which you want to receive information.

1.13 CHECK YOUR PROGRESS

1. Under --- ----------information relating to only selected units is collected, rather than collecting information about all the units of the aggregate.
2. A portion of the selection of the whole is called a ----------
3. The use of the census method is useful where the population is in units---------- of population.
4. In the ----------method, first of all, the whole is divided into levels or sub-levels.
5. Non-sampling errors are those which caused by ---------- factors.

1.14 ANSWERS TO CHECK YOUR PROGRESS

1. **Sampling Method 2. Sampling 3. Heterogeneous 4. Multi-stage Random Sampling 5. human**

1.15 TERMINAL QUESTIONS

1. Explain data collection methods along with their merits and demerits.

2. Describe sampling methods. Also discuss their relative merits and demerits.

3. Write a short note on sampling and non-sampling errors.

4. Discuss sampling distribution of means.

5.Differentiate standard deviation and standard error.

6. Explain the law of large number and the central limit theorem.

7. A population consists of the following elements: 2, 4, 5, 8 and 11. Find:

(a) How many different samples of size 3 are possible when sampling is done without replacement?

(b) List all of the possible different samples.

(c) Compute the mean of each of the samples given in part (b).

(d) Find the sampling distribution of the sample mean $\overline{x}$.

(e) If all the elements are equally likely, compute the value of the population mean μ.

ANSWERS

CHECK YOUR PROGRESS

7. (a) The total number of possible samples of size 3 drawn without replacement is ${}^{5}C_{3}=10$.

(b) All the possible different samples and their sample means are shown in the following table.

Sample No.	**Sample Values**	**Sample Mean $\bar{x}$**
1.	(2,4,5)	$\frac{1}{3}(2+4+5)=3.67$
2.	(2,4,8)	$\frac{1}{3}(2+4+8)=4.67$
3.	(2,4,11)	$\frac{1}{3}(2+4+11)=5.67$
4.	(2,5,8)	$\frac{1}{3}(2+5+8)=5.0$
5.	(2,5,11)	$\frac{1}{3}(2+5+11)=6.0$
6.	(2,8,11)	$\frac{1}{3}(2+8+11)=7.0$
7.	(4,5,8)	$\frac{1}{3}(4+5+8)=5.67$
8.	(4,5,11)	$\frac{1}{3}(4+5+11)=6.67$
9.	(4,8,11)	$\frac{1}{3}(4+8+11)=7.67$
10.	(5,8,11)	$\frac{1}{3}(5+8+11)=8.0$

(c) In the above table, you have 10 possible samples of size 3 without replacement. Since, 5.67 occurs twice, its probability of occurrence is $\frac{2}{10}$. Each of the other sample means occur only once with probability $\frac{1}{10}$.

(d) Sampling distribution of means $\bar{x}$ (i.e., the probability distribution of sample mean $\bar{x}$) is given below:

Sampling Distribution of $\bar{x}$

Sample Mean $\bar{x}$	3.67	4.67	5	5.67	6	6.67	6	7.67	8.0
Probability (p)	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

(e) Population consists of the values (2, 4, 5, 8, 11). Since each value occurs equally likely, the probability of occurrence of each value is $\frac{1}{5}$. Hence,

Sample Mean $\bar{x}$	2	4	5	8	11
Probability (p)	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

$$\text{Population Mean} = \mu = 2 \times \tfrac{1}{5} + 4 \times \tfrac{1}{5} + 5 \times \tfrac{1}{5} + 8 \times \tfrac{1}{5} + 11 \times \tfrac{1}{5}$$

$$= \frac{1}{5}.[2 + 4 + 5 + 8 + 11] = \frac{30}{5} = 6$$

1.16 SUGGESTED READINGS

1. Roy Ramendu, 'Principles of Statistics' Prayag Pustak Bhawan, Allahabad.

2. Gupta S. P. & Gupta M. P., 'Business Statistics' Sultan Chand & Sons, New Delhi.

3.Shukla S. M. & Sahai S. P., 'Advanced Statistics' Sahitya Bhawan Publications, Agra

4. Goon, Gupta and Dasgupta, 'Basic Statistics' World Press Limited – Calcutta.

5. Fundamentals of Business Statistics – Sanchethi and Kappor.

6. Srivastava, Shenoy and Guptha, 'Quantitative Methods in Management

UNIT 2 : POINT ESTIMATION AND INTERVAL ESTIMATION

Structure

OBJECTIVES

After studying this unit, you will be able to understand:

- The basic concept of estimation;
- Properties of a good estimator;
- Types of estimators;
- Applications of point estimation and
- The applications of interval estimation

2.1 INTRODUCTION

In day-to-day life, there is a need to make an estimate of a population parameter from the sample statistic. For example, suppose you are willing to find out the average amount of Coca-Cola drunk per day by the students in a university. It is difficult to find out the average of all the students. To solve the problem, you can take a sample and the average amount of Coca Cola drunk is found out. This sample mean is then used to find the average of the population. In fact, you can estimate the population average based on the sample average. The theory of estimation deals with the estimation of the unknown population parameters (such as population mean and population variance) from the corresponding sample statistics (such as sample mean and sample variance).

2.2 BASIC CONCEPTS OF STATISTICAL ESTIMATION

The following terms are used in the study of statistical estimation:

2.2.1 Estimators and Estimates:

To estimate population parameters, you can use various sample statistics. Those sample statistics, like the sample mean $\bar{x}$ sample median *m***,** sample variance s^2 etc., which are used to estimate the unknown population parameters like population mean μ, population variance σ^2, etc., are called estimators and the actual value taken by the estimators are called estimates. If $\hat{\theta}$ (read as theta hat). Thus, θ is an estimator of the population parameter θ.

2.2.2 Point Estimate and Interval Estimate

An estimate of the population parameter can be done in two ways:

(1) Point Estimate: A single value of a statistic that is used to estimate the unknown population parameter is called a point estimate, e.g., the sample mean $\bar{x}$ which you can use for estimating the population mean μ, is a point estimator of μ. Similarly, the statistic s^2 is a point estimator of σ^2 whereas the value of s^2 is computed from a random sample. The point estimate is a single point on the real number scale and named as point estimator.

(2) Interval Estimate: An interval estimate refers to the probable range within which the real value of a parameter is expected to lie. The two extreme limits of such a range are called fiducial or confidence limits, and the range is called a confidence interval. These are determined based on sample studies of a population. Thus, based on sample studies when you estimate that the average monthly expenditure of students staying in hostel is between Rs. 1000 and Rs. 2000 it will be a case of interval estimate and the figures of Rs. 1000 and 2000 will be the two extreme limits within which the actual expenditure of the students would lie.

2.3 PROPERTIES OF A GOOD ESTIMATOR

There can be more than one estimators of a population parameter, e.g., the population mean (μ) may be estimated either by the sample mean ($\bar{x}$) or by sample median (m) or by sample mode (z) etc. Similarly, the population variance (σ^2) may be estimated either by the sample variance (s^2) sample S.D. (s), sample mean deviation, etc. Therefore, it becomes necessary to determine a good estimator out of a number of available estimators. A good estimator is close to the true value of the parameter as possible. Good estimators possess the following characteristics or properties:

2.3.1 Unbiased Estimator

An estimator $\hat{\theta}$ is said to be an unbiased estimator of the population parameter θ if the mean of the sampling distribution of the estimator $\hat{\theta}$ is equal to the corresponding population parameter θ.

Symbolically, $\mu_{\hat{\theta}} = \theta$

In terms of mathematical expectation $\hat{\theta}$ is an unbiased estimator of θ if the expected value of the estimator is equal to the parameter being estimated.

Symbolically, $E(\hat{\theta}) = \theta$

Example 2.1: Sample mean $\bar{x}$ is an unbiased estimator of the population mean μ because, the mean of the sampling distribution of the means $\mu_{\bar{x}}$ or E ($\bar{x}$) is equal to the population mean μ.

Symbolically, $\mu_{\bar{x}} = \mu$ or $E(\bar{x}) = \mu$

Example 2.2: Sample variance s^2 is a biased estimator of the population variance σ^2 because the mean of the sampling distribution of variance is not equal to the population variance.

Symbolically, $\mu_s^2 \neq \sigma^2 \; or \; E(s^2) \neq \acute{o}^2$

However, the modified sample variance (s^2) is an unbiased estimator of the population variance σ^2 because

$$E(\hat{s}^2) \neq \acute{o}^2 \qquad Where\ \hat{s}^2 = \frac{n}{n-1} \times s^2$$

Example 2.3: Sample proportion p is an unbiased estimator of the population proportion P because, the mean of the sampling distribution of proportion is equal to the population proportion.

Symbolically, $\mu_P = p$ or $E(p) = p$

2.3.2 Consistent Estimator

An estimator is said to be consistent if the estimator approaches the population parameter as the sample size increases. In other words, an estimator $\hat{\theta}$ is said to be a consistent estimator of the population parameter θ, if the probability that $\hat{\theta}$ approaches to θ is 1 as n becomes larger and larger.

Symbolically, $P(\hat{\theta} \rightarrow \theta) \rightarrow 1$ as $n \rightarrow \infty$

Note: A consistent estimator need not to be unbiased

A sufficient condition for the consistency of an estimator is that

(i) $E(\hat{\theta}) \rightarrow \theta$

(ii) $Var(\hat{\theta}) \rightarrow 0$ as $n \rightarrow \infty$

Example 2.4: Sample mean $\bar{x}$ is a consistent estimator of the population mean μ because the expected value of the sample mean approaches to the population mean and the variance of the sample mean approaches to zero as the size of the sample is sufficiently increased.

Symbolically, $(i)E(\bar{x}) \to ì$

$$(ii)Var\ (\bar{x}) = \frac{ó^2}{n} \to 0 \quad as \quad n \to \infty$$

Example 2.5: The Sample median is also a consistent estimator of the population mean because:

$$(i)E(m) \to ì$$

$$(ii)Var\ (m) \to 0 \quad as \quad n \to \infty$$

2.3.3 Efficient Estimator:

Efficiency is a relative term. The efficiency of an estimator is generally defined by comparing it with another estimator. Suppose you are considering two unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of θ. Out of them, the estimator $\hat{\theta}_1$ will be called an efficient estimator of θ if the variance of $\hat{\theta}_1$ is less than the variance of $\hat{\theta}_2$.

Symbolically, $\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$

Then $\hat{\theta}_1$ is called an efficient estimator.

Example 2.6: Sample mean $\bar{x}$ is an unbiased and efficient estimator of the population mean (or true mean) than the sample median *m* because the variance of the sampling distribution of the means is less than the variance of the sampling distribution of the medians.

The relative efficiency of the two unbiased estimators is given below:

You know that $Var\ (\bar{x}) = \frac{ó^2}{n}, Var\ (m) = \frac{ð}{2}.\frac{ó^2}{n}$

Efficiency $= \frac{Var(\bar{x})}{Var\ (m)} = \frac{\frac{ó^2}{n}}{\frac{ðó^2}{2n}} = \frac{2}{ð} = \frac{14}{22} = \frac{7}{11} = 0.64 \left[ð = \frac{22}{7}\right]$

$$Var\ (\bar{x}) = 0.64.Var\ (m)$$

Therefore, the sample mean $\bar{x}$ has 64% more efficiency than the sample median. Hence the sample mean is more efficient estimator of the population mean as compared to sample median.

2.3.4 Sufficient Estimator

The last property that a good estimator should possess is sufficiency. An estimator $\hat{\theta}$ is said to be a sufficient estimator of a parameter θ if it contains all the information in the sample regarding the parameter. In other words, a sufficient estimator utilizes all information available in the sample as furnished about the population. Sample means $\bar{x}$ is said to be a sufficient estimator of the population mean μ.

2.4 APPLICATION OF POINT ESTIMATION

Now, you will study the applications of point estimation as follows:

2.4.1 Point Estimation in case of Single Sampling

When a single independent random sample is drawn from an unknown population is called single sampling. The point estimator of the population parameter can be illustrated by the following examples:

Example 2.7: A sample of 10 measurements of the diameter of a sphere gave a mean $\bar{x}$ = 4.38 inches and a variance =.06 inches. Determine the unbiased and efficient estimates of (a) the true mean (i.e. population mean) and (b) the true variance (i.e. population variance).

Solution: You are given n=10, $\bar{x}$=4.38, s^2=.06

(a) The unbiased and efficient estimate of the true mean μ is given by:

$\bar{x} = 4.38$

(b) The unbiased and efficient estimate of the true variance σ^2 is:

$$\hat{s}^2 = \frac{n}{n-1}.s^2$$

Putting the values you get

$$\hat{s}^2 = \frac{10}{10-1} \times .06 = 1.11 \times 0.06 = .066$$

Thus, $\mu = 4.38, \acute{o}^2 = 0.666$.

Example 2.8: The following five observations constitute a random sample from an unknown population:

6.33,6.37,6.36,6.32 and 6.37 centimeters.

Find out unbiased and efficient estimates of (a) true mean and (b) true variance.

Solution:

(a) The unbiased and efficient estimate of the true mean (i.e. population mean) is given by the value of

$$\overline{x} = \frac{\sum x}{n} = \frac{6.33 + 6.37 + 6.36 + 6.32 + 6.37}{5} = \frac{31.75}{5} = 6.35$$

(b) The unbiased and efficient estimate of the true variance (i.e. population variance) is:

$$\hat{s}^2 = \frac{\sum(x - \overline{x})^2}{n-1}$$

Where $\hat{s}^2$ =modified sample variance.

$$= \frac{(6.33-6.35)^2(6.37-6.35)^2(6.36-6.35)^2(6.32-6.35)^2(6.37-6.35)^2}{5-1}$$

$$= \frac{.0022}{4} = .00055\ cm^2$$

Example 2.9: The following data relate to a random sample of 100 students in a university classified by their weights (kg):

Weight (kg)	60-62	63-65	66-68	69-71	72-74
No. of Students	5	18	42	27	8

Determine unbiased and efficient estimates of (a) population mean and (b) population variance.

Solution:

Calculation of Mean and variance

Weight	No. of Students (*f*)	M.V. (*m*)	A=67, d=*m*-A	*d'*=*d*/3	*fd'*	$f\,d'^2$
60-62	5	61	-6	-2	-10	20
63-65	18	64	-3	-1	-18	18
66-68	42	67	0	0	0	0
69-71	27	70	+3	+1	+27	27
72-74	8	73	+6	+2	+16	32

	$n = 100$				$\sum fd' = 15$	$\sum fd'^2 = 97$

(a) The unbiased and efficient estimate of the population mean is given by the value:

$$\overline{x} = A + \frac{\sum fd'}{n} \times i$$

$$= 67 + \frac{15}{100} \times 3 = 67 + (0.45) = 67.45$$

(b) The unbiased and efficient and efficient estimate of the population variance is:

$$\hat{s}^2 = \frac{n}{n-1} . s^2$$

Where,

$$s^2 = \frac{\sum fd'^2}{n} - \left(\frac{\sum fd'}{n}\right)^2 \times i^2$$

$$= \left[\frac{97}{100} - \left(\frac{15}{100}\right)^2\right] \times 3^2$$

$$= [0.97 - .0225] \times 9 = 8.5275$$

Now,

$$\hat{s}^2 = \frac{n}{n-1} s^2 = \frac{100}{99} \times 8.5275 = 8.6136$$

Thus,

$$ì = 67.45, ó^2 = 8.6136$$

2.4.2 Point Estimation in Case of Repeated Sampling

When more than one random sample of same size are drawn from a population with or without replacement is called repeated sampling. It may be understood by the following examples:

Example 2.10: A population consists of five values: 3,4,5,6 and 7. List all possible samples of size 3 without replacement from this population and calculate the mean ($\bar{x}$) of each sample. Verify that sample mean ($\bar{x}$) is an unbiased estimate of the population mean.

Solution: The population consists of the five values: 3,4,5,6 and 7. The total number of possible samples of size 3 without replacement are ${}^5C_3 = 10$ which are shown in the following table:

Sample No.	Sample Values	Sample Mean ($\bar{x}$)
1	(3,4,5)	$\frac{1}{3}(3+4+5) = \frac{12}{3}4$
2	(3,4,6)	$\frac{1}{3}(3+4+6) = \frac{13}{3}4.33$
3	(3,4,7)	$\frac{1}{3}(3+4+7) = \frac{14}{3}4.67$
4	(3,5,6)	$\frac{1}{3}(3+5+6) = \frac{14}{3}4.67$
5	(3,5,7)	$\frac{1}{3}(3+5+7) = \frac{15}{3}5.0$
6	(3,6,7)	$\frac{1}{3}(3+6+7) = \frac{16}{3}5.33$
7	(3,5,6)	$\frac{1}{3}(3+5+6) = \frac{15}{3}5.00$
8	(3,5,7)	$\frac{1}{3}(3+5+7) = \frac{16}{3}5.33$
9	(3,6,7)	$\frac{1}{3}(3+6+7) = \frac{17}{3}5.67$

10	(3,6,7)	$\frac{1}{3}(3+6+7)=\frac{18}{3}6.00$
Total	k=10	$\sum \bar{x}=50$

Mean of Sampling Distribution of Means = $ì_{\bar{X}}=\sum\frac{\bar{x}}{k}=\frac{50}{10}=5.$

Population Mean (μ)= $\frac{3+4+5+6+7}{5}=5$

Therefore, it may be stated that $ì_{\bar{X}=ì}$ sample mean ($\bar{x}$) is an unbiased estimate of the population mean μ.

Example 2.11: Consider a hypothetical population comprising three values: 1, 2 and 3. Draw all possible samples of size 2 with replacement. Calculate the mean ($\bar{x}$) and variance s^2 for each sample. Examine whether the two statistics (($\bar{x}$) and s^2) are unbiased and efficient for the corresponding parameters.

Solution: The population consists of three values: 1, 2, and 3. The total numbers of possible samples of size 2 with replacement are $N^n=3^2=9$ are as given under:

Sample No.	Sample Values	Sample Mean ($\bar{x}$)	Sample Variance $s^2=\frac{1}{2}[(x_1-\bar{x})^2+(x_2-\bar{x})^2]$	Modified Sample Variance $\left(\hat{s}^2=\frac{n}{n-1}s^2\right)$
1	(1,1)	$\frac{1}{2}(1+1)=1.0$	$\frac{1}{2}[(1-1)^2+(1-1)^2]=0.00$	0.00
2	(1,2)	$\frac{1}{2}(1+2)=1.5$	$\frac{1}{2}[(1-1.5)^2+(2-1.5)^2]=0.25$	0.50

3	(1,3)	$\frac{1}{2}$ (1+3)=2.0	$\frac{1}{2}$ [(1-2)^2 +(3-2)^2] =1.0	2.00
4	(2,1)	$\frac{1}{2}$ (2+1)=1.5	$\frac{1}{2}$ [(2-1.5)^2+(1-1.5)^2]=0.25	0.5
5	(2,3)	$\frac{1}{2}$ (2+2)=2.0	$\frac{1}{2}$ [(2-2)^2+(2-2)^2]=0.00	0.00
6	(2,3)	$\frac{1}{2}$ (2+2)=2.5	$\frac{1}{2}$ [(2-2.5)^2+(3-2.5)^2]=0.25	0.50
7	(3,1)	$\frac{1}{2}$ (3+1)=2.0	$\frac{1}{2}$ [(3-2)^2 +(1-2)^2] =1.00	2.00
8	(3,2)	$\frac{1}{2}$ (3+2)=2.5	$\frac{1}{2}$ [(3-2.5)^2 + (2-2.5)^2] =0.25	0.50
9.	(3.3)	$\frac{1}{2}$ (3+3)=3.0	$\frac{1}{2}$ [(3-3)^2 +(3-3)^2]=0.00	0.00
Total	k=9	$\sum(\bar{x}) = 18$		$\sum \hat{s}^2 = 6$

(a) Mean of Sampling Distribution of Means=$\mu_{\bar{x}} = \sum \frac{\bar{x}}{k} = \frac{18}{9} = 2$. Here k=No. of sample.

Population Mean μ = $\frac{1+2+3}{3} = 2$.

Since $\mu_{\bar{x}}$ =μ, sample mean $\bar{x}$ is an unbiased estimate of the population mean μ.

(b) Mean of the sampling Distribution of Variance =$\mu_{s^2} = \sum \frac{s^2}{k} = \frac{3}{9} = \frac{1}{3}$

Population Variance σ^2 =

$$\frac{(1-2)^2+(2-2)^2+(3-2)^2}{3} = \frac{2}{3}$$

Since $\mu_{s^2} \neq \sigma^2$, sample variance s^2 is not an unbiased estimator of the population variance (σ^2). But the modified sample variance is defined as $\hat{s}^2 = \frac{n}{n-1}.s^2$ will be an unbiased estimate of the population variance σ^2 because:

$$ì_{\hat{s}^2} = \sum \frac{\hat{s}^2}{k} = \frac{6}{9} = \frac{2}{3}$$

$$\sigma^2 = \frac{2}{3}$$

Therefore, $ì_{\hat{s}^2} = \sigma^2$

Since $ì_{\hat{s}^2} = \sigma^2$ the modified sample variation is an unbiased estimate of the population variance.

Example 2.12: Show that the sample mean ($\overline{x}$) is an unbiased estimate of the population mean.

or

An independent random sample $x_1, x_2, x_3, \ldots\ldots, x_n$ is drawn from a population with mean μ. Prove that the expected value of the sample mean ($\overline{x}$) equals the population mean μ.

Solution: A random sampling is one where each sample has an equal chance of being selected. You may draw a random sample of size 'n'.

Then,

$$E(\overline{x}) = E\left[\frac{x_1 + x_{2+} \ldots + x_n}{n}\right], \text{ Where } x_1 \text{is the sample observation.}$$

$$= \frac{1}{n}\left[E(x_1) + E(x_2) + \cdots E(x_n)\right]$$

Now the expected values of x_i (a member of the population) is population mean μ.

Therefore, $E(\overline{x}) = \frac{1}{n}\left[ì + \cdots + ì\right]$ Because $[E(x_1) = (x_2) = E(x_n) = ì]$

$= \frac{1}{n}.[nì] = ì$ Because, $[\sum C = c_1 + c_2 + \cdots c_n = nC]$

Thus, the sample mean $\overline{x}$ is an unbiased estimate of the population mean.

2.5 INTERVAL ESTIMATION (OR CONFIDENCE INTERVAL)

In the theory of interval estimation, you will find an interval or two numbers within which the value of the unknown population parameter is expected to lie with a specified probability. The method of interval estimation consists of the determination of two constants t_1 and t_2 such that P [t_1 <θ< t_2 for given value of t]=1 - α, where α is the level of significance. The interval [t_1 and t_2] within which the unknown value of parameter θ is expected to lie is known as the confidence interval, and the limits t_1 and t_2 so determined are known as confidence limits and 1-α is called the confidence coefficient depending upon the desired precision of the estimate, e. g., α=0.05 (or 0.01) gives 95% (or 99%) confidence limits.

Now, you will study the procedure for setting up confidence interval (or interval estimation) or limits for a population parameter. For the purpose, the following steps enable you to compute the confidence interval or confidence limits for the population parameter θ in terms of the sample statistic *t*.

(i) compute or take the appropriate sample statistic *t*;

(ii) obtain the S.E. (*t*), the standard error of the sample statistic *t*; and

(iii) select the confidence level and corresponding to that specified level of confidence or you can note down the critical value of the statistic *t.*

2.6 APPLICATIONS OF INTERVAL ESTIMATION

The applications relating to interval estimation (or confidence interval) are studied under the following heads:

2.6.1 Interval Estimation (or Confidence Interval) for Large Samples ($n \geq 30$)

In large sample, the interval estimation is further divided under the following heads:

- Confidence Interval or Limits for Population Mean
- Confidence Interval or Limits for Population Proportion
- Confidence Interval or Limits for Population Standard Deviation
- Determination of a proper Sample Size for Estimating μ or P.

(1) Confidence Interval or Limits for Population Mean μ (when $n \geq 30$)

The determination of the confidence interval or limits for the population mean μ in case of large sample ($n > 30$) requires the use of normal distribution.

(i) $(1 - á)$ 100% Confidence limits for μ are given by:

$$\overline{x} \pm Z_{á/2}.S.E_x$$

or $\overline{x} \pm Z_{á/2}.\frac{ó}{\sqrt{n}}$ Where, ó is known.

or $\overline{x} \pm Z_{á/2}.\frac{s}{\sqrt{n}}$ When ó is not known [for large sample, $ó = s$]

(ii) $(1 - á)$ 100% confidence interval for μ is given by:

$$\overline{x} - Z_{á/2}\frac{ó}{\sqrt{n}} < μ < \overline{x} + Z_{á/2}.\frac{ó}{\sqrt{n}}$$

or $\overline{x} - Z_{á/2}\frac{s}{\sqrt{n}} < ì < \overline{x} + Z_{á/2}.\frac{s}{\sqrt{n}}$ Where, ó is not known.

In particular, 95% confidence limits for μ are:

$\overline{x} \pm 1.96\frac{ó}{\sqrt{n}}$ [For large sample $ó = s$]

Similarly, 99% confidence limit for μ are

$$\overline{x} \pm 2.58\frac{ó}{\sqrt{n}}$$

Procedure: The construction of confidence interval for population mean μ involves the following steps:

(i) Compute $\overline{x}$ or take $\overline{x}$

(ii) Compute the $S.E._{\bar{x}}$ by using the following formula:

(a) $S.E._{\bar{x}} = \frac{ó}{\sqrt{n}}$,when ó is known.

(b) $S.E._{\bar{x}} = \frac{s}{\sqrt{n}}$,when ó is not known.

(iii) Select the desired confidence level and corresponding to that level of confidence you will find the value of $Z_{á/2}$.

(iv) Substitute the value of $\bar{x}, S.E._{\bar{x}}$ and $Z_{á/2}$ in the above stated formula.

Note:

1. If the population $S.D.$ is not known then the sample $S.D.(s)$ is used for large samples.

2. The values of $Z_{á/2}$ (for large samples) corresponding to various level of confidence are given below:

Confidence Level (1-α) 100% Z-Value	90% ± 1.64	95% ± 1.96	96% ± 2.06	98% ± 2.33	99% ± 2.58	Without any reference to the confidence level ± 3

Note: Where no reference to the confidence interval is given then you should take $Z_{á/2} = 3$. This value corresponds to 99.73% level of confidence.

Example 2.13: A random sample of 100 observations yields sample mean $\bar{x} = 150$ and sample variance $s^2 = 400$. Compute 95% and 99% confidence interval for the population mean.

Solution: You are given: $n = 100, \bar{x} = 150, s^2 = 400 \Rightarrow s = 20$

$$S.E._{\bar{x}} = \frac{s}{\sqrt{n}} \qquad \text{[For large sample } ó = s\text{]}$$

$$= \frac{20}{\sqrt{100}} = 2$$

At 95% confidence level, the value of $Z_{á/2} = 1.96$

At 99% confidence level the value of $Z_{á/2} = 2.58$

(a) 95% confidence Interval or limits for μ are:

$$\bar{x} \pm 1.96\, S.E._{x}$$

Putting the values, you will get

$$150 \pm 1.96 \times 2 = 150 \pm 3.92 = 153.92 \; or \; 146.08$$

Thus, $146.08 < \mu < 153.92$

(b) 99% Confidence interval or Limits for μ are:

$$\bar{x} \pm 2.58\, S.E._{\bar{x}}$$

$$= 150 \pm 2.58 \times 2$$

$$= 150 \pm 5.16$$

$$= 155.16 \; or \; 144.84$$

Thus, $144.84 < \mu < 155.16$

Example 2.14: A random sample of 900 workers in a steel plant showed an average height of 67 inches with a standard deviation of 5 inches.

(a) Establish a 95% confidence interval estimate of the mean height of all the workers at the steel plant.

(b) Establish a 99% confidence interval estimate of the mean height of all the workers at the steel plant.

Solution: You are given: $n = 900, \bar{x} = 67, s = 5$

$$\text{S. E.}\,(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{5}{\sqrt{900}} = 0.167 \quad \text{[For large sample, } s = ó\text{]}$$

At 95% Confidence level the value of $Z_{á/2} = 1.96$

At 99% Confidence level the value of $Z_{á/2} = 2.58$

(a) 95% confidence interval for μ is :

$$\bar{x} \pm 1.96, S.E._{\bar{x}}$$

Putting the values, you will get

$$67 \pm 1.96 \times (0.167)$$

$$= 67 \pm 0.327 = 67.327 \text{ to } 66.673$$

Thus, $66.673 < \mu < 67.327$

(b) 99% confidence interval for μ are:

$$\bar{x} \pm 2.58. S.E._{\bar{x}}$$

Putting the values, you will get

$$= 67 \pm 2.58\ (0.167)$$

$$= 67 \pm 0.43$$

$$= 67.43\ to\ 66.57$$

Thus, $66.57 < \mu < 67.43$

2. Confidence Interval or Limits for population Proportion P

Though the sampling distribution associated with proportions is the binomial distribution but, the normal distribution can be used as an approximation provided that the sample is large $(i.e., n \geq 30)$ and both np and $nq > 5$ (when n is the size of the sample p is the proportion of success and $q = 1 - p$).

(i) $(1 - á)$ 100% Confidence limits for P are given by:

$$p \pm Z_{á/2}.S.E.(p)$$

or $$p \pm Z_{á/2}.\sqrt{\frac{PQ}{n}}$$ When P is known.

or $$p \pm Z_{á/2}.\sqrt{\frac{pq}{n}}$$ When P is not known

(ii) $(1 - á)$ 100% confidence interval for P is given by:

$$p - Z_{á/2}.\sqrt{\frac{pq}{n}} < P < p + Z_{á/2}.\sqrt{\frac{pq}{n}}$$

95% confidence limits for P are:

$$p \pm 1.96.\sqrt{\frac{pq}{n}}$$

99% confidence limits for P are:

$$p \pm 2.58.\sqrt{\frac{pq}{n}}$$

Procedure: The construction of confidence limits or interval for population proportion involves the following steps:

(i) Compute p or take p

(ii) Compute the $S.E.(p)$ by using the following formula:

$$S.E.(p) = \sqrt{\frac{PQ}{n}}$$ When P is known.

$$S.E.(p) = \sqrt{\frac{pq}{n}}$$ When P is not known.

(iii) Select the desired confidence level and corresponding to that level you will find the value of $Z_{á/2}$

(iv) Substitute the values of p, $S.E.(p)$ and $Z_{á/2}$, in the above stated formula.

Note:

1. If the population proportion (p) is not known, then sample proportion (p) is used for large samples.
2. When no reference to the confidence level is given then always take $Z_{á/2} = 3$ for 99.73% confidence level.

Example 2.15: Out of 1,200 tosses of a coin it gave 480 heads and 720 tails. Find the 95 percent confidence interval for the heads.

Solution: You are given: $n = 1200, \text{Total heads } (np) = 480$

$$p = \text{Sample proportion of heads} = \frac{480}{1200} = 0.4$$

Also the population proportion of heads $= p = 0.50$

$$Q = 1 - P = 1 - 0.50 = 0.50$$

$$S.E._{(p)} = \sqrt{\frac{PQ}{n}} \qquad [\text{For large sample } p = P]$$

$$= \sqrt{\frac{0.5 \times 0.5}{1200}} = 0.0144$$

For 95% confidence level the value of $Z_{á/2} = 1.96$

95% Confidence interval for P is given by:

$$p \pm 1.96\, S.E._p$$

Putting the values, you will get

$$= 0.4 \pm 1.96 \times 0.0144$$

$$= 0.4 \pm 0.028$$

$$= 0.372 \; to \; 0.428$$

Thus, $0.372 < P < 0.428$

Example 2.16: A random sample of 600 pineapples was taken from a large consignment and 75 of them were found to be bad. Estimate the proportion of bad apples in the consignment and obtain the standard error of the estimate. Assign the limits within which the percentage of bad pineapples in the consignment lies.

Solution: You are given: $n = 600$ No. of bad pineaplles (np) $= 75$

Samples proportion,

$$p = \frac{75}{600} = 0.125 = 12.5\%$$

$$q = 1 - 0.125 = 0.875$$

$$S.E._{(p)} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.125 \times 0.875}{600}} = 0.013$$

Since the level of confidence is not specified then you may assume it as 99.73%

For a 99.73% confidence level, the value of $Z_{á/2} = 3$

99.73% confidence limits for P are given by

$$p \pm 3 \times S.E._{\bar{x}}$$

Putting the values, you will get

$$= 0.125 \pm 3 \times 0.013$$

$$= 0.125 \pm 0.039$$

$$= 0.164 \pm 0.086$$

Hence, the required percentage lies between 16.4% and 8.6%

Confidence interval or limits for population proportion P when the sample is drawn without replacement from a finite population: In this case $(1 - á)$ 100% confidence interval or limits are given by:

$$p \pm Z_{á/2}.\sqrt{\frac{pq}{n}}.\sqrt{\frac{N-n}{N-1}}$$

Where, $\sqrt{\frac{N-n}{N-1}}$ = Finite Population Correction Factor

Note: If N is sufficiently large as compared to the sample size n the finite population correction factor may be ignored.

Example 2.17: Out of 20000 customers' ledger accounts, a sample of 600 accounts was taken to test the accuracy of posting and balancing, wherein 45 mistakes were found. Assign limits within which the number of defective cases can be expected at a 95% level.

Solution: You are given: $n = 600, N = 20{,}000,$ Number of mistakes in the sample ledger accounts (np) $= 45$

Sample Proportion

$$p = \frac{np}{n} = \frac{45}{600} = 0.075$$

$$q = 1 - p = 1 - 0.075 - 0.925$$

Since N is sufficiently large compared to the sample size n, the finite population correction factor $\sqrt{\frac{N-n}{N-1}}$ may be ignored. Hence, assuming it as a sample from a finite (large) population, the standard error of p is given by

$$S.E.(p) = \sqrt{\frac{pq}{n}}$$

$$= \sqrt{\frac{0.075 \times 0.925}{600}} = \sqrt{0.0001156}$$

$$= 0.011 \text{ (approx)}$$

For 95% confidence level the value of $Z_{á/2} = 1.96$

95% confidence limits for population P are given by:

$$p \pm 1.96\, S.E._{\bar{x}}$$

Putting the values, you get

$$= 0.075 \pm 1.96 \times 0.011$$

$$= 0.075 \pm 0.022 = (0.053, 0.097)$$

Hence, the number of defective cases in a lot of 20,000 is expected to lie between $20{,}000 \times 0.053\ and\ 20{,}000 \times 0.097\ i.e.\ 1060\ and\ 1940$.

Note: If the finite population correction factor is not ignored then; 95% confidence limits for P are:

$$p \pm 1.96. \sqrt{\frac{pq}{n}} . \sqrt{\frac{N-n}{N-1}}$$

$$= 0.075 \pm 1.96 \sqrt{\frac{0.075 \times 0.925}{600}} \times \sqrt{\frac{20{,}000 - 600}{20{,}000 - 1}}$$

$$= 0.075 \pm 1.96 \times 0.0108$$

$$= 0.075 \pm 0.021168$$

$$= (0.0538, 0.096168)$$

Hence, the required number of defective cases in the lot lies between 20,000 (0.0538, 0.096168) *i.e.*, 1076 and 1924.

(3) Confidence Interval or Limits for Populations Standard Deviation: The determination of the confidence interval or limits for population S.D. σ in the case of large sample ($n \geq 30$) requires the use of normal distribution.

(i) $(1 - á)$ 100% confidence limits for σ are given by

$$s \pm Z_{á/2}.S.E._s$$

or

$$s \pm Z_{á/2}.\frac{ó}{\sqrt{2n}}$$ When ó is known

or

$$s \pm Z_{á/2}.\frac{s}{\sqrt{2n}}$$ When ó is not known.

(ii) $(1 - á)$ 100% confidence interval for ó is given by:

$$s - Z_{á/2}.\frac{s}{\sqrt{2n}} < ó < s + Z_{á/2}.\frac{s}{\sqrt{2n}}$$

95% confidence limits for ó are:

$$s \pm 1.96\,.\frac{s}{\sqrt{2n}}$$ [For large sample, $s = ó$]

99% confidence limits for ó are:

$$s \pm 2.5 \pm \frac{s}{\sqrt{2n}}$$

Procedure: The construction of confidence limits for ó involves the following steps:

(i) Compute s or take s

(ii) Compute $S.E.(s)$ by using the following formula.

$$S.E.(s) = \frac{ó}{\sqrt{2n}}$$

or

$$S.E.(s) = \frac{s}{\sqrt{2n}}$$

(iii) Select the desired confidence level and corresponding to that confidence level, the value of $Z_{á/2}$

(iv) Substitute the values of $s, Z_{á/2}$ and n in the above stated formula

(4) Determination of Sample Size for estimating μ *or* P:

As you have calculated the confidence intervals based on the assumption that the sample size n is known. Now, you will be able to calculate the confidence interval when sample size is not known.

(a) Sample Size for Estimating a Population Mean: In order to determine the sample size for estimating population mean the following three factors must be known.

(i) The desired confidence level and the corresponding values of Z.

(ii) The permissible sampling error E.

(iii) The standard deviation ó or an estimate of ó $(i.e., \bar{s})$

After having known the above-mentioned factors, the sample size n is given by:

$$n = \left(\frac{Z.ó}{E}\right)^2$$

Note:

1. The values of Z and E are predetermined.

2. The population $S.D.$ (ó) may be actual or estimated.

Example 3.18: A cigarette manufacturer wishes to use a random sample to estimate the average nicotine content. The sampling error should not be more than one milligram above or below the true mean, with a 99 percent confidence level. The population standard deviation is 4 milligrams. What sample size should the company use in order to satisfy these requirement?

Solution: You are given: $E = 1, Z_{á/2} = 2.58$ for 99% confidence level and ó $= 4$.

Sample size formula is:

$$n = \frac{Z^2.ó^2}{E^2}$$

Substituting the values, you will get

$$n = \frac{(2.58)^2\ (4)^2}{1^2} = 106.50\ or\ 107$$

Hence, the required sample size $n = 107$ which the company should use for their requirements.

(b) Sample size for Estimating a Population proportion: In order to determine the sample size for estimating population proportion, the following three factors must be known.

(i) the desired level of confidence and the corresponding value of Z.

(ii) the permissible sampling error E.

(iii) the actual or estimated true proportion of success P.

The sample size n is given by:

$$n = \frac{Z^2 \times PQ}{E^2} \quad \text{Where, } Q = 1 - P$$

Note:

1. The values of Z and E are predetermined.

2. The value of the population proportion P may be actual or estimated.

Example 2.19: A firm wishes to determine with a maximum allowable error of 0.05 and a 98 percent level of confidence for the proportion of consumer who prefer its product. How large a sample will be required in order to make such an estimate if the preliminary sales reports indicate that 25 percent of all the consumers prefer the firm's product?

Solution: You are given:

$E = 0.05, P = 0.25, Q = 1 - 0.25 = 0.75, Z = 2.33$ for 98% confidence level.

Sample size formula is

$$n = \frac{Z^2 \times PQ}{E^2}$$

Substituting the values, you will get

$$n = \frac{(2.33)^2}{(0.05)^2}\ (0.25)(0.75)$$

$$= \frac{5.4289}{0.0025}\ (0.1875) = \frac{1.0179}{0.0025} = 407.16\ or\ 408$$

Hence, the required sample size $n = 408$.

2.6.2 INTERVAL ESTIMATION FOR SMALL SAMPLES ($n < 30$)

The determination of confidence intervals in case of small sized sample ($n < 30$) is studied under two headings:

(1) Confidence interval or limits for population mean($n < 30$)**:** When the samples size is small ($i.e., n < 30$) and ó (the population S.D.) is unknown the desired confidence interval or limits for population mean ì can be found by making use of t-distribution. In case of small samples, t-values are used in place of Z-values.

(i) $(1 - ó)$ 100% confidence limits for population mean ì are given by:

$\overline{x} \pm t_{á/2}.\frac{\hat{s}}{\sqrt{n}}$ Where, $\hat{s}$ = modified sample $S.D. = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$ or $\hat{s} = \sqrt{\frac{n}{n-1}}s^2$

(ii) $(1 - á)$ 100% confidence interval for ì is given by:

$$\overline{x} \pm t_{á/2}.\frac{\hat{s}}{\sqrt{n}} < ì < \overline{x} \pm t_{á/2}.\frac{\hat{s}}{\sqrt{n}}$$

95% confidence limits for ì are given by

$$\overline{x} \pm t_{0.025}.\frac{\hat{s}}{\sqrt{n}}$$

99% confidence limits for ì are given by

$$\overline{x} \pm t_{0.005}.\frac{\hat{s}}{\sqrt{n}}$$

Procedure: The Construction of the confidence interval or limits in case of small sample ($n < 30$) involves the following steps:

(i) Compute $\overline{x}$ or take $\overline{x}$.

(ii) Compute modified sample S.D. using the following formula.

$$\hat{s} = \sqrt{\frac{\sum(x-\overline{x})^2}{n-1}}$$

or

$$\hat{s} = \sqrt{\frac{n}{n-1}.s^2} \quad \text{When, } s \text{ is given.}$$

(iii) Compute the degree of freedom $(d.f.)$ using the formula:

$$d.f. = v = n - 1$$

(iv) Select the desired confidence level and corresponding to that specified level of confidence and for given degrees of freedom, you should note the value of the $t_{á/2}$ from the t-table.

(v) Substitute the values of $\bar{x}, \hat{s}$ and $t_{á/2}$ in the above stated formula.

Example 2.20: A random sample of size 16 has a mean of 50 with standard deviation of 3. Obtain 98 percent confidence limits of the mean of the population.

Solution: You are given: $n = 16, \bar{x} = 50, s = 3 \Longrightarrow s^2 = 9$

$$\hat{s} = \sqrt{\frac{n}{n-1}}.s^2 = \sqrt{\frac{16}{16-1}} \times 9 = 3.098$$

Degrees of freedom $= v = n = 16 - 1 = 15$

For 98% confidence level $á = 0.02 \; so \; that \; \frac{á}{2} = \frac{0.02}{2} = 0.01$

Using t-table the value of $t_{.01}$ for 15 $d.f. = 2.602.$

98% confidence limits for ì are given by

$$\bar{x} \pm t_{.01}.\frac{\hat{s}}{\sqrt{n}}$$

Putting the values, you will get

$$= 50 \pm 2.602 \times \frac{3.098}{\sqrt{16}}$$

$$= 50 \pm 2.015$$

$$= 52.015 \; to \; 47.985$$

Example 2.21: A random sample of 16 items from a normal population showed a mean of 53 and the sum of squares of deviations from this mean is equal to 150. Obtain 95% and 99% confidence limits for the mean of the population.

Solution: You are given: $n = 16, \bar{x} = 53, \sum(x - x)^2 = 150$

$$\hat{s} = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

$$= \frac{150}{16 - 1} = \sqrt{\frac{150}{15}} = \sqrt{10} = 3.162$$

Degrees of freedom $= v = n - 1 = 16 - 1 = 15$

For a 95% confidence level, á = 0.05 so $\frac{á}{2} = \frac{0.05}{2} = 0.025$

For a 99% confidence level, á = 0.01 so $\frac{á}{2} = \frac{0.01}{2} = 0.005$

The table value of $t_{.025}$ for 15 $d.f. = 2.131$

The table vale of $t_{.005}$ for 15 $d.f. = 2.947$

(a) 95% confidence limits for population mean ì are:

$$\bar{X} \pm t_{0.025}.\frac{\hat{s}}{\sqrt{n}}$$

Putting the values, you will get

$$= 53 \pm 2.31 \times \frac{3.162}{\sqrt{16}}$$

$$= 53 \pm 2.131 \times \frac{3.162}{4}$$

$$= 53 \pm 1.684$$

$$= 51.316 \; to \; 54.684$$

Thus, $51.316 < ì < 54.684$

(b) 99% confidence limits for population mean ì are

$$\overline{x} \pm t_{0.005} . \frac{\hat{s}}{\sqrt{n}}$$

Putting the values, you get

$$= 53 \pm 2.947 \times \frac{3.162}{\sqrt{16}}$$

$$= 53 \pm 2.947 \times \frac{3.162}{4}$$

$$= 53 \pm 2.947 \times 0.7905$$

$$= 53 \pm 2.33$$

$$= 55.33 \; and \; 50.67$$

Thus, $50.67 < ì < 55.33.$

(2) Confidence Interval or Limits for Population Variance (when $n < 30$). The determination of confidence interval or limits for population variance $ó^2$ requires the use of χ^2 (Chi-square) distribution. Here χ^2 values are used in place of t-values.

$(1 - á)$ 100% confidence interval for population variance $ó^2$ is given by:

$$\frac{(n-1)\hat{s}^2}{\div^2_{á/2}} < ó^2 < \frac{(n-1)\hat{s}^2}{\div^2_{1-á/2}}$$

In particular, 95% confidence interval for the population variance $ó^2$ is

$$\frac{(n-1)\hat{s}^2}{\div^2_{0.025}} < ó^2 < \frac{(n-1)\hat{s}^2}{\div^2_{0.975}}$$

Similarly, 99% confidence interval for the population variance $ó^2$ is

$$\frac{(n-1)\hat{s}^2}{\div^2_{0.005}} < ó^2 < \frac{(n-1)\hat{s}^2}{\div^2_{0.995}}$$

Procedure: The construction of the confidence interval for the variance $ó^2$ involves the following steps:

(i) Calculate modified sample variance $\hat{s}^2$ by using the formula

$$\hat{s}^2 = \frac{n}{n-1} s^2 \frac{\sum(x-\overline{x})^2}{n-1}$$

(ii) Select the desired confidence level and corresponding to that specified level of confidence, you should note the value of the confidence coefficient $\div^2_{á/2}$ and $\div^2_{1-á/2}$ from the $\div^2$ table for certain degrees of freedom

(iii) Construct the confidence interval for $ó^2$ by putting the values of $\hat{s}^2$,$\div^2_{á/2}$ and $\div^2_{1-á/2}$ in the above stated formula.

Example 2.22: A random sample of size 15 selected from a normal population has a standard deviation $s = 2.5$. Construct a 95 percent confidence interval for variance σ^2 and standard deviation σ.

Solution: You are given: $n = 15, s = 2.5 \Longrightarrow s^2 = 6.25$

$$\hat{s}^2 = \left(\frac{n}{n-1}\right) s^2$$

$$= \frac{15}{15-1} \times 6.25 = 6.696$$

For a 95% confidence level,

á $= 0.05$ so $\frac{á}{2} = 0.025$ and $1 - á = 1 - 0.025 = 0.975$.

Degrees of freedom $(v) = n - 1 = 15 - 1 = 14$

The table value of $\div^2_{0.025}$ for 14 $d.f. = 26.1$

The table value of $\div^2_{0.0975}$ for 14 $d.f. = 5.63$

(a) 95% confidence interval for $ó^2$ is

$$\frac{(n-1)\hat{s}^2}{\div^2_{0.025}} < ó^2 < \frac{(n-1)\hat{s}^2}{\div^2_{0.975}}$$

Putting the values, you will get

$$\frac{(15-1) \times 6.696}{26.1} < ó^2 < \frac{(15-1) \times 6.696}{5.63}$$

or $3.59 < ó^2 < 16.65$

(b) 95% confidence interval for ó is:

$$\sqrt{3.59} < ó < \sqrt{16.65}$$

or $1.89 < ó < 4.08$

2.7 SUMMARY

Sometimes, you are not able to draw any conclusion about the population or you are not able to convey what the population results are in statistical terms. In that case, there is a need to estimate a population parameter based on sample statistics. This is known as the theory of estimation. To estimate the population parameter sample statistics are used. And such types of estimates must have the characteristics of unbiased, consistent, efficient, and sufficient.

2.8 GLOSSARY

Estimator: To estimate composite parameters, you use various sample data which are sample data like sample mean,$\bar{X}$ sample medium m, sample variability, etc., which are unknown composite parameters like composite mean, μ, population Variance ó^2 estimate the overall variability, etc. They are called estimators.

2.9 CHECK YOUR PROGRESS

Fill in the blanks

1. A single value of a statistic that is used to estimate the unknown population parameter is called a ……. estimate.
2. When a single independent random sample is drawn from an unknown population is c

2.10 ANSWER TO CHECK YOUR PROGRESS

1. Point

2. Single sampling

2.11 TERMINAL QUESTIONS

1. Measurements of sample of masses were determined to be 8.3, 10.6, 9.7, 8.8, 10.2 and 9.4 kilograms (kg). Determine unbiased and efficient estimates of (a) the population mean and (b) the population variance and (c) compare the sample standard deviation and estimated population S.D.

2. A random sample of 9 individuals has the following heights in inches: 45, 47, 50, 52, 48, 47, 49, 53 and 51. Find the unbiased and efficient estimate of (a) true mean and (b) true variance.

3. A sample of 10 television tubes produced by a company showed a mean life of 1200 hr. and a standard deviation of 10 hr. Find the unbiased and efficient estimates of the (a) population mean and (b) population variance.

4. A random sample of 144 observations yields sample mean $\bar{x} = 160$ and sample variance $s^2 = 100$. Compute a 95% confidence interval for population mean.

5. From a random sample of 64 farms are found to have a mean area of 45 hectares with a standard deviation of 12. What are the 95% and 99% confidence limits for the mean area?

6 From a random sample of 100 farms are found to have a mean area of 250 hectares with a standard deviation of 50. Compute 99% confidence interval of the mean area. How does the width of the confidence interval change if the size of the sample were increased to 400?

7 A random sample 300 households in a city revealed that 123 of these house had the holy book Ramayana. Find a 95% percent confidence interval for the proportion of households in the city with Ramayana.

8. A random sample of 500 houses in a city disclosed that 125 of these houses had color TV sets. Find a 98 percent confidence interval for the proportion of houses in the city with colour TV Sets. (Table value of Z for 98% confidence level is 2.33).

9. In a market survey for the introduction of a new product given in a town a sample of 400 persons was drawn. When they were approached for sale, 80 of them purchased the product. Find a 95% confidence limits for the purchase of persons who would buy the product in the town.

10. Out of 10,000 customers ledger accounts a sample of 400 accounts was selected to judge the accuracy of posting and balancing. It contained 40 mistakes. Assign limits within which the number of defective cases could be expected at 95 percent level.

11. A sample of 100 items gives a standard deviation 25. Set up the limits for the population standard deviation at 95% level of confidence.

12. A sample of 100 items gives a standard deviation of 4700. Set up the limits for the population standard deviation at 99% confidence level of confidence.

13. A firm wishes to estimate with an error of not more than 0.03 and a level of confidence of 98% for the proportion of consumers that prefers its brand of household detergent. Sales reports indicate that about 0.20 of all consumers prefer the firm's brand. What is the requisite sample size?

14. Mr. X wants to determine the average time to complete a certain job. The past records show that population standard deviation is 10 days. Determine the sample size so that Mr. X may be 95% confident that the sample average remains ± 2 days of the average.

15. In measuring reaction time, a psychologist estimates that the standard deviation is 0.05 seconds. How large a sample of measurements must be taken in order to be 95% confident that the error of his estimate will not be exceeded 0.01 seconds.

16. A sample of 9 cigarettes of a certain brand was observed for nicotine content. It showed the average nicotine of 25 milligrams and a standard deviation of 2.8 milligrams. Construct a 99 percent confidence interval for the true average nicotine content of this particular brand of cigarettes.

17. A random sample of 15 ladies from a colony shows that their monthly expenditure on cosmetics is Rs. 120 with a standard deviation of Rs. 40. Construct 95 percent confidence interval for the true monthly average expenditure on cosmetics by all the ladies in the colony.

18. A sample of 5 individuals had the following heights in inches 63.3, 63.7, 63.6, 63.2 and 3.8. Construct confidence intervals for population variance at 95%.

ANSWERS

1. [(a) 9.5, (b) 0.736 and (c) $\hat{s}=\sigma=0.86$, $s=0.78$]

2.[(a) 49.11 and (b) 6.91]

3.[(a) $\mu=1200$ hrs. and (b) $\hat{s}^2=11$

4.[$158.37 < \mu < 161.03$]

5.[(a) 47.94, 42.06, (b) 48.87, 41.63]

6. [(a) $237.1 < \mu < 262.9$(b) reduced to half]

7.[$0.355 < P < 0.465$]

8.[$0.205 < P < 0.295$]

9.[0.1608, 0.2392]

10.[(a) $0.071 < P < 0.129$ (b) $710 < x < 1290$]

11.[21.55, 28.46]

12.[5032.4, 4367.60]

13.[$n = 965$]

14.[$n = 96$]

15.[$n = 96$]

16.[$21.67 < \mu < 28.33$]

17.[$97.01 < \mu < 142.99$]

18.[$0.0240 < \sigma < 0.5537$]

2.12 SUGGESTED READINGS

1 Roy Ramendu, 'Principles of Statistics' Prayag Pustak Bhawan, Allahabad.

2. Gupta S. P. & Gupta M. P., 'Business Statistics' Sultan Chand & Sons, New Delhi.

3. Shukla S. M. & Sahai S. P., 'Advanced Statistics' Sahitya Bhawan Publications, Agra.

4. Goon, Gupta and Dasgupta, 'Basic Statistics' World Press Limited – Calcutta.

5. Fundamentals of Business Statistics – Sanchethi and Kapoor.

6. Srivastava, Shenoy and Guptha, 'Quantitative Methods in Management'

UNIT: 3 APPROACHES TO PROBABILITY

Structure

OBJECTIVES

After studying this unit, you will be able to understand:

- The meaning of random experiments and cite examples thereof;
- The role of chance in such random experiments;
- Sample space corresponding to an experiment;
- Difference between various types of events, such as equally likely, mutually exclusive, exhaustive, independent and dependent events;
- probability of occurrence of an event; and
- Use permutations and combinations in solving problems of probability.

3.1 INTRODUCTION

In day-to-day life, you see that before the commencement of a cricket match two captains go for a toss. Tossing a coin is an activity, and getting either a "Head" or a "Tail" are two possible outcomes. (Assuming that the coin does not stand on the edge). If you throw a die (of course fair die) the possible outcomes of this activity could be any one of its faces having numerals, namely 1,2,3,4,5 and 6 at the top face.

An activity that yields a result or an outcome is called an experiment. Normally, there are varieties of outcomes of an experiment and it is a matter of chance as to which one of these occurs when an experiment is performed. In this unit, you propose to study various experiments and their outcomes.

Further, you often used phrases such as "It may rain today", or "India may win the match" or "I may be selected for this post." These phrases involve an element of uncertainty. How can you measure this uncertainty? A measure of this uncertainty is provided by a branch of Mathematics, called the theory of probability. Probability Theory is designed to measure the degree of uncertainty regarding the happening of a given event. The dictionary meaning of probability is ' likely though not certain to occur. Thus, when a coin is tossed, a head is likely to occur but may not occur. Similarly, when a die is thrown, it may or may not show the number 6.

3.2 RANDOM EXPERIMENT

Let us consider the following activities:

(i) Toss a coin and note the outcomes. There are two possible outcomes, either a head (H) or a tail (T).

(ii) In throwing a fair die, there are six possible outcomes, that is, any one of the six faces 1,2,3,4,5 and 6 may come on top.

(iii) Toss two coins simultaneously and note down the possible outcomes. There are four possible outcomes, HH,HT,TH,TT.

(iv) Throw two dice and there are 36 possible outcomes which are represented as below:

1,1	1,2	1,3	1,4	1,5	1,6
2,1	2,2	2,3	2,4	2,5	2,6
3,1	3,2	3,3	3,4	3,5	3,6
4,1	4,2	4,3	4,4	4,5	4,6
5,1	5,2	5,3	5,4	5,5	5,6
6,1	6,2	6,3	6,4	6,5	6,6

Each of the above mentioned activities fulfill the following two conditions:

(a) The activity can be repeated number of times under identical conditions.

(b) Outcome of an activity is not predictable beforehand, since the chance play a role and each outcome has the same chance of being selection. Thus, due to the chance playing a role, an activity is

(i) repeated under identical conditions, and

(ii) whose outcome is not predictable beforehand is called a random experiment.

Example 3.1: Is drawing a card from well shuffled deck of cards, a random experiment?

Solution:

(a) The experiment can be repeated, as the deck of cards can be shuffled every time before drawing a card.

(b) Any of the 52 cards can be drawn and hence the outcome is not predictable beforehand. Hence, this is a random experiment.

Example 3.2: Selecting a chair from 100 chairs without preference is a random experiment Justify.

Solution:

(a) The experiment can be repeated under identical conditions.

(b) As the selection of the chair is without preference, every chair has equal chances of selection. Hence, the outcome is not predictable beforehand. Thus, it is a random experiment.
Can you think of any other activities which are not random in nature?

Let us consider some activities which are not random experiments.

(i) Birth of Avni: Obviously this activity, that is, the birth of an individual is not repeatable and hence is not a random experiment.

(ii) Multiplying 4 and 8 on a calculator.
Although this activity can be repeated under identical conditions, the outcome is always 32. Hence, the activity is not a random experiment.

3.3 SAMPLE SPACE

You throw a die once, what are possible outcomes? Clearly, a die can fall with any of its faces at the top. The number on each of the faces is, therefore, a possible outcome.

You write the set S of all possible outcomes as

$$S = \{1, 2, 3, 4, 5, 6\}$$

Again, if you toss a coin, the possible outcomes for this experiment are either a head or a tail.

You write the set S of all possible outcomes as

$$S = \{H, T\}.$$

The set S associated with an experiment satisfying the following properties:

(i) each element of S denotes a possible outcome of the experiment.

(ii) any trial results in an outcome that corresponds to one and only one element of the set S is called the sample space of the experiment and the elements are called sample points. Sample space is generally denoted by S.

Example 3.3: Write the sample space in two tosses of a coin.

Solution: Let H denote a head and T denote a tail in the experiment of tossing a coin.

$$S = \{ (H, H), (H, T), (T, H), (T, T) \}.$$

Note: If two coins are tossed simultaneously then the sample space S can be written as

$$S = \{ H H, H T, T H, T T\}.$$

Example 3.4: Consider an experiment of rolling a fair die and then tossing a coin. Write the sample space.

Solution: In rolling a die possible outcomes are 1, 2, 3, 4, 5 and 6. On tossing a coin, the possible outcomes are either a head or a tail. Let H (head) = 0 and T (tail) = 1.

S = {(1, 0), (1, 1), (2, 0), (2, 1),(3, 0), (3, 1),(4, 0), (4, 1), (5, 0), (5, 1), (6, 0), (6, 1)} n(S) = 6 x2=12

Example 3.5: Suppose you take all the different families with exactly 3 children. The experiment consists in asking them the sex (or genders) of the first, second and third child. Write down the sample space.

Solution: Let us write 'B' for boy and 'G' for girl. The sample space is

S = {BBB, BBG, BGB, BGG, GBB, GBG, GGB, GGG}

The advantage of writing the sample space in the above form is that a question such as "Was the second child a girl" ? or "How many families have first child a boy ?" and so forth can be answered immediately.

n (S) = 2 × 2 × 2 = 8

3.4 DEFINITION OF VARIOUS TERMS

3.4.1 Event:

Let us consider the example of tossing a coin. In this experiment, you may be interested in 'getting a head'. Then the outcome 'head' is an event.

In an experiment of throwing a die, our interest may be in, 'getting an even number'. Then the outcomes 2, 4 or 6 constitute the event. You have seen that an experiment which, though repeated under identical conditions, does not give unique results but may result in any one of the several possible outcomes, which constitute the sample space.

Some outcomes of the sample space satisfy a specified description, which you call an 'event'. You often use the capital letters A, B, C etc. to represent the events.

Example 3.6: Let E denote the experiment of tossing three coins at a time. List all possible outcomes and the events that

(i) the number of heads exceeds the number of tails.

(ii) getting two heads.

Solution: The sample space S is

S = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

= w1 , w 2 , w 3 , w 4 , w 5 , w6 , w 7 , w8

(say)

If E1 is the event that the number of heads exceeds the number of tails, and E 2 the event getting two heads. Then

E1 w1, w2, w3, w5
and
E 2 w2, w3, w5

3.4.2 Equally Likely Events

Outcomes of a trial are said to be equally likely if taking into consideration all the relevant evidences there is no reason to expect one in preference to the other.

Examples:

(i) In tossing an unbiased coin, getting head or tail are equally likely events.
(ii) In throwing a fair die, all the six faces are equally likely to come.
(iii) In drawing a card from a well shuffled deck of 52 cards, all the 52 cards are equally likely to come.

3.4.3 Mutually Exclusive Events

Events are said to be mutually exclusive if the happening of any one of them preludes the happening of all others, i.e., if no two or more of them can happen simultaneously in the same trial.

Examples:

(i) In throwing a die, all 6 faces numbered 1 to 6 are mutually exclusive. If any one of these faces comes at the top, the possibility of others, in the same trial, is ruled out.
(ii) When two coins are tossed, the event that both should come up with tails and the event that there must be at least one head are said to be mutually exclusive.
Mathematically, events are said to be mutually exclusive if their intersection is a null set (i.e., empty)

3.4.4 Exhaustive Events

If you have a collection of events with the property that no matter what the outcome of the experiment, one of the events in the collection must occur, then you say that the events in the collection are exhaustive events.

For example, when a die is rolled, the event of getting an even number and the event of getting an odd number are exhaustive events. Or when two coins are tossed, the event that at least one head will come up and the event come up with at least one tail are called exhaustive events.

Mathematically, a collection of events is said to be exhaustive if the union of these events is the complete sample space.

3.4.5 Independent and Dependent Events

A set of events is said to be independent if the happening of any one of the events does not affect the happening of others. If, on the other hand, the happening of any one of the events influences the happening of the other, the events are said to be dependent.

Examples:

(i) In tossing an unbiased coin the event of getting a head in the first toss is independent of getting a head in the second, third and subsequent throws.

(ii) If you draw a card from a pack of well shuffled cards and replace it before drawing the second card, the result of the second draw is independent of the first draw. But however, if the first card drawn is not replaced then the second card is dependent on the first draw (in the sense that it cannot be the card drawn the first

3.5 EVENTS AND THEIR PROBABILITY

In the previous section, you have learnt whether an activity is a random experiment or not. The study of probability always refers to random experiments. Hence, from now onwards, the word experiment will be used for a random experiment only. In the preceding section, you have defined different types of events such as equally likely, mutually exclusive, exhaustive, independent and dependent events and cited examples of the above-mentioned events.

Here you are interested in the chance that a particular event will occur, when an experiment is performed. Let us consider some examples.

What are the chances of getting a ' Head' in tossing an unbiased coin ? There are only two equally likely outcomes, namely head and tail. In our day to day language, you say that the coin has chance 1 in 2 of showing up a head. In technical language, you say that the probability of getting a head is 1/2.

Similarly, in the experiment of rolling a die, there are six equally likely outcomes 1, 2,3,4,5 or 6. The face with number '1' (say) has chance 1 in 6 of appearing on the top. Thus, you say that the probability of getting 1 is 1/6.

In the above experiment, suppose you are interested in finding the probability of getting even number on the top, when a die is rolled. Clearly, the numbers possible are 2, 4 and 6 and the chance of getting an even number is 3 in 6. Thus, you say that the probability of getting an even number is 3/6 or 1/2

In total, if an experiment with 'n' exhaustive, mutually exclusive and equally likely outcomes, m outcomes are favourable to the happening of an event A, the probability 'p' of happening of A is given by

p=Number of favourable outcomes/Total number of possible outcomes

or

m/n……..(i)

Since the number of cases favourable to the non-happening of the event A are n-m , the probability 'q' that 'A' will not happen is given by

$$q=\frac{n-m}{n}= 1-\frac{m}{n}$$

$$=1-p \text{ [see eq. (i)] or } p+q=1.$$

Obviously, p as well as q are non-negative and cannot exceed unity.

$$\text{i.e., } 0 \leq p \leq 1, 0 \leq q \leq 1$$

Thus, the probability of occurrence of an event lies between 0 and 1[including 0 and 1].

Notes:

1. Probability 'p' of the happening of an event is known as the probability of success and the probability 'q' of the non-happening of the event as the probability of failure.
2. Probability of an impossible event is 0 and that of a sure event is 1 if P (A) = 1, the event A is certainly going to happen and if P (A) = 0, the event is certainly not going to happen.
3. The number (m) of favourable outcomes to an event cannot be greater than the total number of outcomes (n).

Example 3.7: A die is rolled once. Find the probability of getting a 5.

Solution: There are six possible ways in which a die can fall, out of these only one is favourable to the event.

$$P(5)=1/6$$

Example 3.8: A coin is tossed once. What is the probability of the coin coming up with head?

Solution: The coin can come up either 'head' (H) or a tail (T). Thus, the total possible out- comes are two and one is favourable to the event.

$$P(H)=1/2$$

Example 3.9: A die is rolled once. What is the probability of getting a prime number ?

Solution: There are six possible outcomes in a single throw of a die. Out of these; 2, 3 and 5 are the favourable cases.

$$P(\text{Prime Number}) = 3/6 \text{ or } 1/2.$$

Example 3.10: A die is rolled once. What is the probability of the number '7' coming up ? What is the probability of a number 'less than 7'

coming up ?

Solution: There are six possible outcomes in a single throw of a die and there is no face of the die with mark 7.

P (number 7)=0/6.

As every face of a die is marked with a number less than 7.Therefore, p (£7)= 6/6=1

Example 3.11: In a simultaneous toss of two coins, find the probability of (i) getting 2 heads (ii) exactly 1 head.

Solution: Here, the possible outcomes are HH, HT, TH, TT. i.e., Total number of possible outcomes = 4.

(i) Number of outcomes favourable to the event (2 heads) = 1 (i.e., HH). Therefore, P (2 heads)= 1/4.

(ii) Now the event consisting of exactly one head has two favourable cases, namely HT and TH . Therefore, P (one head)= 2/4=1/2.

Example 3.12: In a single throw of two dice, what is the probability that the sum is 9?

Solution: The number of possible outcomes is $6 \times 6 = 36$. You write them as given below:

1,1	1,2	1,3	1,4	1,5	1,6
2,1	2,2	2,3	2,4	2,5	2,6
3,1	3,2	3,3	3,4	3,5	3,6
4,1	4,2	4,3	4,4	4,5	4,6
5,1	5,2	5,3	5,4	5,5	5,6
6,1	6,2	6,3	6,4	6,5	6,6

Now, how do you get a total of 9. You have:

3 + 6 = 9
4 + 5 = 9
5 + 4 = 9
6 + 3 = 9

In other words, the outcomes (3, 6), (4, 5), (5, 4) and (6, 3) are favourable to the said event, i.e., the number of favourable outcomes is 4.

Hence, P (a total of 9)=4/36 or 1/9.

Example 3.13: From a bag containing 10 red, 4 blue and 6 black balls, a ball is drawn at random. What is the probability of drawing (i) a red ball? (ii) a blue ball ? (iii) not a black ball ?

Solution: There are 20 balls in all. So, the total number of possible outcomes is 20. (Random drawing of balls ensure equally likely outcomes)

(i) Number of red balls = 10
Therefore, p (a red ball)= 10/20 or 1/2
(ii) Number of blue balls = 4
Therefore, P (a blue ball)=4/20 or 1/5
(iii) Number of balls which are not black = 10 + 4=14
Therefore, P (not a black ball)=14/20 or 7/10

Example 3.14: A card is drawn at random from a well shuffled deck of 52 cards. If A is the event of getting a queen and B is the event of getting a card bearing a number greater than 4 but less than 10, find P(A) and P (B).

Solution: Well shuffled pack of cards ensures equally likely outcomes.
Therefore, the total number of possible outcomes is 52.
(i) There are 4 queens in a pack of cards.
P (A)=4/52 or 1/13.
(ii) The cards bearing a number greater than 4 but less than 10 are 5,6, 7,8 and 9. Each card bearing any of the above number is of 4 suits diamond, spade, club or heart. Thus, the number of favourable outcomes = 5 × 4 = 20
P(B)= 20/52 or 5/13.

Example 3.15: What is the chance that a leap year, selected at random, will contain 53 Sundays?

Solution: A leap year consists of 366 days consisting of 52 weeks and 2 extra days. These two extra days can occur in the following possible ways.
(i) Sunday and Monday
(ii) Monday and Tuesday
(iv) Tuesday and Wednesday
(v) Wednesday and Thursday
(vi) Thursday and Friday
(vi) Friday and Saturday
(vii) Saturday and Sunday

Out of the above seven possibilities, two outcomes, e.g., (i) and (vii), are favourable to the event.
Therefore, P (53 Sundays)= 2/7.

3.6 PROBABILITY USING PERMUTATIONS AND COMBINATIONS

In the previous section, you calculated the probability of an event by listing down all the possible outcomes and the outcomes favourable to the event. It is possible when the number of outcomes is small, otherwise; it becomes difficult and time consuming process. In general, you do not require the actual listing of the outcomes, but require only the total number of possible outcomes and the

number of outcomes favourable to the event. In many cases, these can be found by applying the knowledge of permutations and combinations, which you have already studied.

Example 3.16: A bag contains 3 red, 6 white and 7 blue balls. What is the probability that two balls drawn are white and blue ?

Solution: Total number of balls = 3 + 6 + 7 = 16
Now, out of 16 balls, 2 can be drawn in $^{16}C_2$ ways.
Therefore, Exhaustive number of cases = $^{16}C_2$= (16x15)/2=120
Out of 6 white balls, 1 ball can be drawn in $^{6}C_1$ ways and out of 7 blue balls, one can be drawn in $^{7}C_1$ ways. Since each of the former case is associated with each of the later case, therefore total number of favourable cases are $^{6}C_1X^{7}C_1$ = 42.
Thus, the required probability = 42/120 or 7/20.

Example 3.17: Find the probability of getting both red balls, when from a bag containing 5 red and 4 black balls, two balls are drawn,
(i) with replacement.
(ii) without replacement.

Solution: (i) Total number of balls in the bag in both the draws = 5 + 4 = 9
Hence, by fundamental principle of counting, the total number of possible outcomes
= 9 × 9 = 81.
Similarly, the number of favourable cases = 5 × 5 = 25.
Hence, probability (both red balls) =25/81.

(ii) Total number of possible outcomes is equal to the number of ways of selecting 2 balls out of 9 balls = $^{9}C_2$.
Number of favourable cases is equal to the number of ways of selecting 2 balls out of 5 red balls =$^{5}C_2$
Hence, P (both red balls) = $^{5}C_2/^{9}C_2$= 5/18.

Example 3.18: Six cards are drawn at random from a pack of 52 cards. What is the probability that 3 will be red and 3 black?

Solution: Six cards can be drawn from the pack of 52 cards in $^{52}C_6$ ways.
i.e., Total number of possible outcomes = $^{52}C_6$
3 red cards can be drawn in $^{26}C_3$ ways and
3 black cards can be drawn in $^{26}C_3$ ways.
Thus, total number of favourable cases = $^{26}C_3$ X $^{26}C_3$
Hence, the required probability == ($^{26}C_3$ X $^{26}C_3$)/ $^{52}C_6$ =13000/39151.

Example 3.19: Four persons are chosen at random from a group of 3 men, 2 women and 4 children. Show that the chance that exactly two of them will be children is 10/21.

Solution: Total number of persons in the group = 3 + 2 + 4 = 9. Four persons are chosen at random. If two of the chosen persons are children, then the remaining two can be chosen from 5 persons (3 men + 2 women).
Number of ways in which 2 children can be selected from 4 children $= {}^4C_2 = 6$
Number of ways in which remaining of the two persons can be selected from 5 persons $= {}^5C_2 = 10$
Total number of ways in which 4 persons can be selected out of 9 persons $= {}^9C_4 = 126$
Hence, the required probability $= ({}^4C_2 X {}^5C_2)/ {}^9C_4 = 10/21$.
Hence proved the chance that exactly two of them will be children is 10/21.

Example 3.20: Three cards are drawn from a well-shuffled pack of 52 cards. Find the probability that they are a king, a queen and a jack.

Solution: From a pack of 52 cards, 3 cards can be drawn in ${}^{52}C_3$ ways, all being equally likely. Therefore, exhaustive number of cases $= {}^{52}C_3$
A pack of cards contains 4 kings, 4 queens and 4 jacks. A king, a queen and a Jack can each be drawn in ${}^4C_1$ ways and since each way of drawing a king can be associated with each of the ways of drawing a queen and a jack, the total number of favourable cases $= {}^4C_1 X {}^4C_1 X {}^4C_1$
Thus, required probability $= ({}^4C_1 X {}^4C_1 X {}^4C_1)/ {}^{52}C_3 = 16/5525$.

Example 3.21: From 25 tickets, marked with the first 25 numerals, one is drawn at random. Find the probability that it is a multiple of 5.

Solution: Numbers (out of the first 25 numerals) which are multiples of 5 are 5,10, 15, 20 and 25, i.e., 5 in all. Hence, required favourable cases are = 5. Therefore, required probability =5/25 or 1/5.

3.7 SUMMARY

An activity that yields a result or an outcome is called an experiment. An activity repeated number of times under identical conditions and outcome of activity is not predictable is called Random Experiment. The set of possible outcomes of a random experiment is called sample space and elements of the set are called sample points. Some outcomes of the sample space satisfy a specified description, which is called an Event. Events are said to be equally likely, when you have no preference for one rather than the other. If happening of an event prevents the happening of another event, then they are called Mutually Exclusive Events. The total number of possible outcomes in any trial is known as Exhaustive Events. A set of events is said to be Independent events, if the

happening of any one of the events does not affect the happening of other events, otherwise they are called dependent events.

3.8 GLOSSARY

Event: Some outcomes of the sample space satisfy a specified description, which you call an 'event'

Equally likely Events: if taking into consideration all the relevant evidence there is no reason to expect one in preference to the other.

3.9 CHECK YOUR PROGRESS

State whether the following statements are True or False.

4 (i) A and B are mutually exclusive.

5 (ii) A and B are mutually exclusive and exhaustive. (iii) A and C are mutually exclusive.
(iv) C and D are mutually exclusive and exhaustive

3.10 ANSWER TO CHECK YOUR PROGRESS

(i) True (ii) True (iii) False (iv) True

3.11 TERMINAL QUESTIONS

1 Selecting a student from a school without preference is a random experiment. Justify.

2 Adding two numbers on a calculator is not a random experiment. Justify.

3 Write the sample space of tossing three coins at a time.

4 Write the sample space of tossing a coin and a die.

5 Two dice are thrown simultaneously, and you are interested in getting six on top of each of the dice. Are the two events mutually exclusive or not?

6 Two dice are thrown simultaneously. The events A, B, C, and D are as below:
A: Getting an even number on the first die.
B: Getting an odd number on the first die.
C: Getting the sum of the number on the dice < 7.
D: Getting the sum of the number on the dice > 7.

7. A ball is drawn at random from a box containing 6 red balls, 4 white balls and 5 blue balls. There will be how many sample points in its sample space?

8. In a single rolling with two dice, write the sample space and its elements.

9. Suppose you take all the different families with exactly 2 children. The experiment consists of asking them the sex of the first and second child. Write down the sample.

10. A die is rolled once. Find the probability of getting 3.

11. A coin is tossed once. What is the probability of getting the tail?

12. What is the probability of the die coming up with a number greater than 3?

13. In a simultaneous toss of two coins, find the probability of getting ' at least' one tail.

14. From a bag containing 15 red and 10 blue balls, a ball is drawn at random'. What is the probability of drawing (i) a red ball ? (ii) a blue ball?

15. If two dice are thrown, what is the probability that the sum is
(i) 6 ? (ii) 8? (iii) 10? (iv) 12?

16. If two dice are thrown, what is the probability that the sum of the numbers on the two faces is divisible by 3 or by 4 ?
17. If two dice are thrown, what is the probability that the sum of the numbers on the two faces is greater than 10?
18. What is the probability of getting a red card from a well shuffled deck of 52 cards?
19. If a card is selected from a well shuffled deck of 52 cards, what is the probability of drawing
    (i) a spade ? (ii) a king ? (iii) a king of spade ?
20. A pair of dice are thrown. Find the probability of getting
    (i) a sum as a prime number
    (ii) a doublet, i.e., the same number on both dice
    (iii) a multiple of 2 on one die and a multiple of 3 on the other.
21 Three coins are tossed simultaneously. Find the probability of getting
    (i) no head (ii) at least one head (iii) all heads.
22. A bag contains 3 red, 6 white and 7 blue balls. What is the probability that two balls drawn at random are both white?
23. A bag contains 5 red and 8 blue balls. What is the probability that two balls drawn are red and blue?
24. A bag contains 20 white and 30 black balls. Find the probability of getting 2 white balls, when two balls are drawn at random (a) with replacement (b) without replacement.
25. Three cards are drawn from a well-shuffled pack of 52 cards. Find the the probability that all the three cards are jacks.
26. Two cards are drawn from a well-shuffled pack of 52 cards. Show that the probability of drawing both aces is 1/221.
27. In a group of 10 outstanding students in a school, there are 6 boys and 4 girls. Three students are to be selected out of these at random for a debate competition. Find the probability that (i) one is boy and two are girls. (ii) all are boys. (iii) all are girls.
28. Out of 21 tickets marked with numbers from 1 to 21, three are drawn at random. Find the probability that the numbers on them are in A.P.
29. Two cards are drawn at random from 8 cards numbered 1 to 8. What is the probability that the sum of the numbers is odd, if the cards are drawn together?
30. A team of 5 players is to be selected from a group of 6 boys and 8 girls. If the selection is made randomly, find the probability that there are 2 boys and 3 girls in the team.
31. An integer is chosen at random from the first 200 positive integers. Find the probability that the integer is divisible by 6 or 8.

ANSWERS OF TERMINAL QUESTIONS

1. Both properties are satisfied.
2. The outcome is predictable.
3. S:HHH,HHT,HTH,HTT,THH,THT,TTH,TTT
4. H1,H2,H3,H4,H5,H6,T1,T2,T3,T4,T4,T6
5. No.
6. 15
7. 1,1 1,2 1,3 1,4 1,5 1,6
   2,1 2,2 2,3 2,4 2,5 2,6

	3,1	3,2	3,3	3,4	3,5	3,6		
	4,1	4,2	4,3	4,4	4,5	4,6		
	5,1	5,2	5,3	5,4	5,5	5,6		
	6,1	6,2	6,3	6,4	6,5	6,6		
8.	{MM, MF, FM, FF}							
9	1/6							
10.	1/2							
11.	1/2							
12.	3/4							
13.	(i)	3/5	(ii)	2/5				
14.	(i)	3/5	(ii)	5/36	(iii)	1/12	(iv)	1/36
15	5/9							
16.	1/12							
17.	1/2							
18.	(i)	1/4	(ii)	1/13	(iii)	1/52		
19.	(i)	5/12	(ii)	1/6	(iii)	11/36		
20.	(i)	1/8	(ii)	7/8	(iii)	1/8		
21.	1/8							
22.	20/39							
23.	(a) 4/25		(b)	38/245				
24.	1/5525							
25.	Proved.							
26.	(i)	3/10	(ii)	1/6	(iii)	1/30		
27.	10/133							
28.	4/7							
29.	60/143							
30.	1/4.							

3.12 SUGGESTED READINGS

1. Roy Ramendu, ‘Principles of Statistics’ Prayag Pustak Bhawan, Allahabad.
2. Gupta S. P. & Gupta M. P., ‘Business Statistics’ Sultan Chand & Sons, New Delhi.
3. Shukla S. M. & Sahai S. P., ‘Advanced Statistics’ Sahitya Bhawan Publications, Agra.
4. Goon, Gupta and Dasgupta, ‘Basic Statistics’ World Press Limited – Calcutta.
5. Fundamentals of Business Statistics – Sanchethi and Kappor.
6. Srivastava, Shenoy and Guptha, ‘Quantitative Methods in Management’

UNIT:4 THEOREMS OF PROBABILITY

Structure

OBJECTIVES

After studying this unit, you will be able to understand:

- Addition theorems on probability.
- Multiplication theorem on probability.
- Use of Bernoulli's Theorem in probability.
- To solve problems involving Bayes' Theorem

4.1 INTRODUCTION

After reading the preceding unit, you can understand various aspects and terms of probability. In these two statements "It may rain today" and "He can reach today", the words may and can convey the same meaning, i.e. the events are not certain to take place. In other words, there is involved an element of certainty or chance in these two cases. A numerical measure of uncertainty is provided by the theory of probability. The aim of probability theory is to provide a measure of uncertainty. The theory of probability owes its origin to the study of chance like games of cards, tossing of coins, dice, etc. But, in modern times, it has great importance is decision-making problems.

4.2 ADDITION THEOREM

You should understand the addition theorem of probability in the following ways:

4.2.1Addition Theorem for Mutually Exclusive Events

The addition theorem states that if A and B are two mutually exclusive events, then the probability of occurrence of either A or B is the sum of the individual probabilities of A and B.

Symbolically,

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

Generalization: The theorem can be extended to three or more mutually exclusive events. If A, B and C are three mutually exclusive events, then

$$\mathbf{P(A+ B+C) = P(A) + P(B) + P(C)}$$

Example 4.1: A card is drawn from a pack of 52 cards. What is the probability of getting either a king or queen?

Solution: There are 4 kings and 4 queens in a pack of 52 cards.

The probability of drawing a king card is $P(K) = \frac{4}{52}$

and the probability of drawing a queen card is $P(Q) = \frac{4}{52}$

Since, both the events are mutually exclusive, the probability that the card drawn either a king or queen is

$$P(K \text{ or } Q) = P(K) + P(Q)$$
$$= \frac{4}{52} + \frac{4}{52} = \frac{8}{52} = \frac{2}{13}$$

Example 4.2: An investment consultant predicts that the odds against the price of a certain stock will go up during the next week are 2 : 1 and the odds in favour of the price remaining the same are 1 : 3. What is the probability that the price of the stock will go down during the next week?

Solution: Let A denote the event 'stock price will go up', and B be the event 'stock price will remain same'

Then $P(A) = \frac{1}{3}$, and $P(B) = \frac{1}{4}$

P (stock price will either go up or remain the same) = $P(A \cup B)$

$$=P(A)+P(B) = \frac{1}{3} + \frac{1}{4} = \frac{7}{12}$$

Now, P(stock price will go down) = 1 - P $(A \cup B)$

= 1-$\frac{7}{12} = \frac{5}{12}$

Example 4.3: Among 3 events A, B and C only one event can take place. The odds against A are 3 : 2, and against B are 4 : 3. Find the odds against events C.

Solution: The probability of happening of A event is P(A) = $\frac{2}{5}$

The probability of happening of A event is P(B) = $\frac{3}{7}$

Since events are mutually exclusive,

P(A or B or C) = P(A) + P(B) + P(C) = 1

By substituting values,

$1 = \frac{2}{5} + \frac{3}{7} + P(C)$

$1 - \left(\frac{2}{5} + \frac{3}{7}\right) = P(C)$

$P(C) = \frac{6}{35}$

$\frac{6}{35}$ probability implies 6 in favour out of 35 chances.

Or odds against event C are 35-6 = 29

Thus, odds against event C are = 29 : 6

Example 4.4: A card is drawn at random from a pack of cards. Find the probability that the drawn card is either a club or an ace of diamond.

Solution: The probability of drawing a card of club P(A) = $\frac{13}{52}$

The probability of drawing an ace of diamond club P(B) = $\frac{1}{52}$

Since the events are mutually exclusive, the probability of the drawn card being a club or an ace of diamond is:

P(A or B) = P(A) + P(B)

= $\frac{13}{52} + \frac{1}{52} = \frac{14}{52} = \frac{7}{26}$

Example 4.5: In a single throw of 2 dice, determine the probability of getting a total 7 or 9.

Solution: In a throw of 2 dice, there are 6 x 6 = 36 possible outcomes as follows:

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

A total of 7 can occur in the following 6 ways:
(6, 1) (5, 2) (4, 3) (3, 4) (2, 5) (1, 6)
A total of 9 can occur in the following 4 ways:
(6, 3) (5, 4) (4, 5) (3, 6)

The probability of getting a total of 7 is $P(A) = \frac{6}{36}$

The probability of getting a total of 9 is $P(B) = \frac{4}{36}$

Since the events are mutually exclusive, the probability of getting either a total of 7 or 9 is:

$$P(A \text{ or } B) = P(A) + P(B)$$
$$= \frac{6}{36} + \frac{4}{36} = \frac{10}{36} = \frac{5}{18}$$

Example 4.6: An urn contains 4 red, 5 black, 3 yellow and 11 green balls. A ball is drawn at random. Find the probability that it is (i) either red, black or a yellow ball (ii) either a red, black, yellow or green.

Solution: Total number of balls= 4R + 5B + 3Y + 11G = 23

Probability of getting a red ball $P(A) = \frac{4}{23}$

Probability of getting a black ball $P(B) = \frac{5}{23}$

Probability of getting a yellow ball $P(C) = \frac{3}{23}$

Probability of getting a red ball $P(D) = \frac{11}{23}$

(i) Since the events are mutually exclusive, the probability of the drawn ball being R, B and Y is

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C)$$
$$= \frac{4}{23} + \frac{5}{23} + \frac{3}{23} = \frac{12}{23}$$

(ii) The probability of the drawn ball being R, B, Y or G is

$$P(A \text{ or } B \text{ or } C \text{ or } D) = P(A) + P(B) + P(C) + P(D)$$
$$= \frac{4}{23} + \frac{5}{23} + \frac{3}{23} + \frac{11}{23} = \frac{23}{23} = 1$$

Example 4.7: If a pair of dice is thrown, find the probability that (i) the sum is neither 7 nor 11 (ii) neither a doublet nor a total of 9 will appear.

Solution: There are 36 possible outcomes, you write them as follows:

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

(i) A total of 7 can occur in the following 6 ways:

(6, 1) (5, 2) (4, 3) (3, 4) (2, 5) (1, 6)

A total of 11 can occur in the following 2 ways: (6, 5) and (5, 6)

The probability of getting a total of 7 is P(A) = $\frac{6}{36}$

The probability of getting a total of 11 is P(B) = $\frac{2}{36}$

Since the events are mutually exclusive, the probability of getting either a total of 7 or 11 is:

$$P(A \text{ or } B) = \frac{6}{36} + \frac{2}{36} = \frac{8}{36} = \frac{2}{9}$$

The probability that the sum is neither 7 or 11 is

$$P(\text{neither 7 or 11}) = 1 - P(\text{either 7 or 11})$$

$$= 1 - \frac{2}{9} = \frac{7}{9}$$

(ii) A doublet can come in 6 ways:

(1, 1) (2, 2) (3, 3) (4, 4) (5, 5) (6, 6)

A total of 9 can come in 4 ways:

(6, 3) (5, 4) (4, 5) (3, 6)

The probability of getting a double P(A) = $\frac{6}{36}$

The probability of getting a total of 9 is P(B) = $\frac{4}{36}$

Since, the events are mutually exclusive, the probability of getting a doublet or a total 9 is:

$$P(A \text{ or } B) = \frac{6}{36} + \frac{4}{36} = \frac{10}{36} = \frac{5}{18}$$

The probability that neither a doublet nor a total of 9 will appear is:

$$P(\text{Neither a doublet nor 11}) = 1 - \frac{5}{18} = \frac{13}{18}$$

Example 4.8: There are 11 red and 14 white balls in the bag. Two balls are drawn. Find the probability that both of them are of the same colour.

Solution: Total number of ways in which 2 balls can be drawn out of 25 balls = ${}^{25}C_2$

Total number of ways in which 2 red balls can be drawn out of 11 red balls = ${}^{11}C_2$

Total number of ways in which 2 white balls can be drawn out of 14 white balls = ${}^{14}C_2$

There are two cases:

(i) Both balls are red,

The probability of getting two red balls = $\frac{{}^{11}C_2}{{}^{25}C_2}$

(ii) Both balls are white,

The probability of getting two white balls = $\frac{{}^{14}C_2}{{}^{25}C_2}$

Since the (i) and (ii) cases are mutually exclusive, therefore,

P (Both balls are of the same colour)

= P (Either 2R or 2W)

= P(2R) + P(2W)

= $\frac{{}^{11}C_2}{{}^{25}C_2} + \frac{{}^{14}C_2}{{}^{25}C_2}$

= $\frac{11}{60} + \frac{91}{300} = \frac{55+91}{300} = \frac{146}{300} = \frac{73}{150}$

Example 4.9: A and B are mutually exclusive events for which P(A) = 0.3, P(B) = p and P(A+B) = 0.5. Find the value of p.

Solution: Since, A and B are mutually exclusive events, then

P(A + B) = P(A) + P(B)

Substituting the values, you get

0.5 = 0.3 + p

p = 0.5 – 0.3 = 0.2

4.2.2 Addition Theorem for Not Mutually Exclusive Events

The addition theorem discussed above is not applicable when the events are not mutually exclusive. For example, if the probability of drawing a card of spade is 13/52 and that of drawing a card of king is 4/52, you cannot calculate the probability of drawing a card of either spade or king by adding the two probabilities because the events are not mutually exclusive. The card could very well be a spade card as well as a king. When the events are not mutually exclusive, the addition theorem has to be modified.

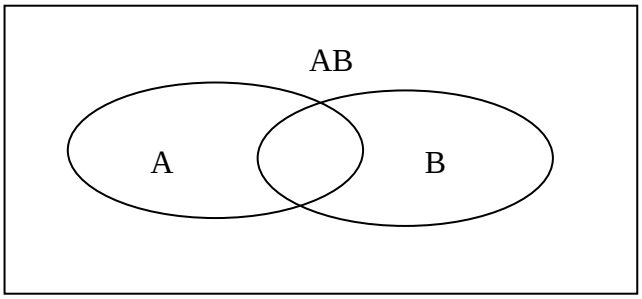
Modified Addition Theorem states that if A and B are not mutually exclusive events, the probability of the occurrence of either A or B or both is equal to the probability that event A occurs, plus the probability that event B occurs minus the probability that events common to both A and B occur.

Symbolically,

P(A or B or Both) = P(A) + P(B) – P(AB)

In this formula, you subtract P(A and B), namely the probability of the events which are counted twice from P(A) + P(B). The theorem is thus modified in such a way as to render A and B mutually exclusive.

The following figure illustrates this point:



Generalization: The theorem can be extended to three more events. If A, B and C are not mutually exclusive events, then

P (Either A or B or C) = P(A) + P(B) + P(C) – P(AB) – P(AC) – P(BC) + P(ABC)

The following examples illustrate the application of the modified addition theorem:

Example 4.10: A card is drawn at random from a well shuffled pack of cards. What is the probability that it is either a spade or a king?

Solution: The probability of drawing a spade P(A) = $\frac{13}{52}$

The probability of drawing a king P(B) = $\frac{4}{52}$

Because one of the kings can belong to a spade, therefore the events are not mutually exclusive.

The probability of drawing a king of spade P(AB) = $\frac{1}{52}$

So, the probability of drawing a spade or king is:

P (A or B or Both) = P(A) + P(B) – P(AB)

= $\frac{13}{52}+\frac{4}{52}-\frac{1}{52}$

= $\frac{16}{52}=\frac{4}{13}$

Example 4.11: A bag contains 30 balls numbered from 1 to 30. One ball is drawn at random. Find the probability that the number of ball is multiple of 5 or 6.

Solution: The probability of the ball being multiple of 5 is:

(5, 10, 15, 20, 25, 3); $P(A) = \frac{6}{30}$

The probability of the ball being multiple of 6 is

(6, 12, 18, 24, 30); $P(B) = \frac{5}{30}$

Since, 30 is a multiple of 5 as well as 6, therefore the events are not mutually exclusive.

$P(A \text{ and } B) = \frac{1}{30}$ (common multiple 30)

So, the probability of getting a ball being multiple of 5 or 6 is:

$$P(A \text{ or } B) = P(A) + P(B) - P(AB)$$
$$= \frac{6}{30} + \frac{5}{30} - \frac{1}{30} = \frac{10}{30} = \frac{1}{3}$$

Example 4.12: One number is drawn from numbers 1 to 150. Find the probability that it is either divisible by 3 or 5.

Solution: The probability that the number being divisible by 3 is

(3, 6, 9, …… 147, 150) $P(A) = \frac{50}{150}$

The probability that the number being divisible by 5 is

(5, 10, 15, …… 145, 150) $P(A) = \frac{30}{150}$

Since, the numbers (15, 30, 45 …….. 135, 150) = 10 are common to both, therefore, the events are not mutually exclusive.

$P(A \text{ and } B) = \frac{10}{150}$ (Common multiple)

So, the probability of getting either divisible by 3 of 5 is:

$$P(A \text{ or } B) = P(A) + P(B) - P(AB)$$
$$= \frac{50}{150} + \frac{30}{150} - \frac{10}{150} = \frac{70}{150} = \frac{7}{15}$$

Example 4.13: A card is drawn at random from a standard pack of cards. What is the probability that (i) it is either a king or queen (ii) it is either a king or a black card?

Solution: (i) The probability of drawing a king card $P(K) = \frac{4}{52}$

The probability of drawing a queen card $P(Q) = \frac{4}{52}$

Since both the events are mutually exclusive, the probability that the card drawn is either a king or queen is

$$P(K \text{ or } Q) = P(K) + P(Q)$$
$$= \frac{4}{52} + \frac{4}{52} = \frac{8}{52} = \frac{4}{26} = \frac{2}{13}$$

(ii) The probability of drawing a king card P(K) = $\frac{4}{52}$

The probability of drawing a black card P(B) = $\frac{26}{52}$

Since black kings are common to both the events are not mutually exclusive.

P (Black Kings) = $\frac{2}{52}$

Thus, the probability that the card drawn is either a king of a black card is

P(a king or black) = P(a king) + P(a black card) - P(a black king)

$= \frac{4}{52} + \frac{26}{52} - \frac{2}{52} = \frac{28}{52}$

Example 4.14: A chartered accountant applied for a job in two firms X and Y. He estimated that the probability of him being selected in a firm X is $\frac{7}{10}$ and being rejected Y is $\frac{5}{10}$ and the probability that he will be selected in both the firm is $\frac{4}{10}$. What is the probability that he will be selected in one of the firm?

Solution: P (Chartered accountant is selected in firm X) = $\frac{7}{10}$

P (he is selected in firm Y) = 1 - P (he is being rejected in form Y)

$= 1 - \frac{5}{10} = \frac{5}{10}$

P (he is selected in both X and Y firms) = $\frac{4}{10}$

P (he will be selected in one of the firm) = P(X) + P(Y) – P(X and Y)

$= \frac{7}{10} + \frac{5}{10} - \frac{4}{10} = \frac{8}{10} = \frac{4}{5}$

4.3 MULTIPLICATION THEOREM

Now, you will study the multiplication theorem of probability under two headings:

4.3.1 Multiplication Theorem for Independent Events

The multiplication theorem states that if A and B are two independent events, then the probability of the simultaneous occurrence of A and B is equal to the product of their individual probabilities. Symbolically,

P(AB) = P(A) x P(B)

Generalization: The theorem can be extended to three or more independent events. If A,B and C are three independent events, then

P(ABC) = P(A) x P(B) x P(C)

Example 4.15: A coin is tossed 3 times. What is the probability of getting all the 3 heads?

Solution: Probability of head in the first toss P(A) = $\frac{1}{2}$

Probability of head in the second toss P(B) = $\frac{1}{2}$

Probability of head in the third toss P(C) = $\frac{1}{2}$

Since the events are independent the probability of getting all heads in three tosses is;

$$P(ABC) = P(A) \times P(B) \times P(C)$$
$$= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$$

Example 4.16: From a pack of 52 cards, two cards are drawn at random one after another with replacement. What is the probability that both cards are kings?

Solution: The probability of drawing a king P(A) = $\frac{4}{52}$

The probability of drawing again a king after replacement P(B) = $\frac{4}{52}$

Since the events are independent, the probability of drawing two kings is:

$$P(AB) = P(A) \times P(B)$$
$$= \frac{4}{52} \times \frac{4}{52} = \frac{1}{169}$$

Example 4.17: A bag containing 5 white and 3 black balls. Two balls are drawn at random one after another with replacement. Find the probability that both the balls drawn are black.

Solution: Probability of drawing black ball in the first draw = P(A) = $\frac{3}{8}$

Probability of drawing a black ball in the second draw P(A) = $\frac{3}{8}$

Since the events are independent, the probability that both the balls are black:

$$P(2 \text{ Black}) = P(1\text{st Black}) \times P(2\text{nd Black})$$
$$= \frac{3}{8} \times \frac{3}{8} = \frac{9}{64}$$

Example 4.18: An electronic device is made of three components A, B, and C. The probability of failure of component A is 0.01, that of B is 0.02 and that of C is 0.05 in some fixed period. Find the probability that the device will work satisfactorily during the period of time assuming that the three components work independently of one another.

Solution: Let the three failure components are denoted by $\overline{A}$, $\overline{B}$ and $\overline{C}$ respectively.

$$P(\overline{A}) = 0.01,\ P(\overline{B}) = 0.02,\ P(\overline{C}) = 0.05$$

The probability that these components do not fail

$P(A) = 1 - P(\overline{A}) = 1 - 0.01 = 0.99$

$P(A) = 1 - P(\overline{B}) = 1 - 0.02 = 0.98$

$P(C) = 1 - P(\overline{C}) = 1 - 0.05 = 0.95$

The probability that the device will work satisfactory is

$P(ABC) = P(A).\ P(B).\ P(C)$

$= 0.99 \times 0.98 \times 0.95$

$= 0.92162 = .092$ (approx.)

4.3.1.1 Probability of happening of at least one event in case of n independent events

If you are given n independent $A_1, A_2, A_3 \ldots A_n$ with a respective probability of happenings as $p_1, p_2, p_3 \ldots p_n$, then the probability of happening of at least one of independent events $A_1, A_2, A_3 \ldots A_n$ is given by:

P (happening of at least one of the events) = 1 - P (happening of none of the events)

$= 1 - [(1- p_1) . (1- p_2) . (1- p_3). (1- p_n)]$

Example 4.19: A problem in statistics is given to three students A, B, and C whose chances of solving it are 1/2, 1/3, and 1/4. What is the probability that the problem will be solved?

Solution: Probability that A will solve the problem = $P(A) = \frac{1}{2}$

Probability that B will solve the problem = $P(B) = \frac{1}{3}$

Probability that C will solve the problem = $P(C) = \frac{1}{4}$

The probability that A will not solve the problem = $P(\overline{A}) = 1-\frac{1}{2} = \frac{1}{2}$

The probability that A will not solve the problem = $P(\overline{B}) = 1-\frac{1}{3} = \frac{2}{3}$

The probability that A will not solve the problem = $P(\overline{C}) = 1-\frac{1}{4} = \frac{3}{4}$

Since, all the events are independent, so

P (that none will solve the problem) $= P(\overline{A}) . P(\overline{B}) . P(\overline{C})$

$= \frac{1}{2} \times \frac{2}{3} \times \frac{3}{4} = \frac{1}{4}$

P (that problem will be solved) = 1-P (that none will solve)

$= 1- \frac{1}{4} = \frac{3}{4}$

Example 4.20: A candidate (Mr. X) is interviewed for 3 posts. For the first post, there are 3 candidates, for the second post, there are 4 and for the third, there are 2. What are the chances of Mr. X being getting selected?

Solution: Probability of selection for 1st post = P(A) = $\frac{1}{3}$

Probability of selection for 2nd post = P(B) = $\frac{1}{4}$

Probability of selection for 3rd post = P(C) = $\frac{1}{2}$

Probability of not selecting on 1st post = P($\overline{A}$) = 1-$\frac{1}{3}$ = $\frac{2}{3}$

Probability of not selecting on 2nd post = P($\overline{B}$) = 1-$\frac{1}{4}$ = $\frac{3}{4}$

Probability of not selecting on 3rd post = P($\overline{C}$) = 1-$\frac{1}{2}$ = $\frac{1}{2}$

Since, the events are independent, the probability that Mr. X is not selected for three posts is :

$$P(\overline{A}\ \overline{B}\ \overline{C}) = P(\overline{A}) \cdot P(\overline{B}) \cdot P(\overline{C})$$

$$= \frac{2}{3} \times \frac{3}{4} \times \frac{1}{2} = \frac{1}{4}$$

Probability of selection for at least 1 post

= 1 - P (not selected at all)

= $1 - \frac{1}{4} = \frac{3}{4}$

Example 4.21: Find the probability of throwing 6 at least once in six throws with a single die.

Solution: The probability of throwing 6 at least 6 at least once = 1 – Probability that 6 is not thrown at all

Probability that 6 is not thrown in the 1st throw = $\frac{5}{6}$

Probability that 6 is not thrown in the 2nd throw = $\frac{5}{6}$

Probability that 6 is not thrown in the 3rd throw = $\frac{5}{6}$

Probability that 6 is not thrown in the 4th throw = $\frac{5}{6}$

Probability that 6 is not thrown in the 5th throw = $\frac{5}{6}$

Probability that 6 is not thrown in the 6th throw = $\frac{5}{6}$

Since, the events are independent the probability that 6 is not thrown in any throw

$$= P(\overline{I}) \cdot P(\overline{II}) \cdot P(\overline{III}) \cdot P(\overline{IV}) \cdot P(\overline{V}) \cdot P(\overline{VI})$$

$$= \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} = \left(\frac{5}{6}\right)^6$$

Hence, the probability of throwing 6 at least once

$$= \quad 1 - \left(\frac{5}{6}\right)^6$$

Example 4.22: Let p be the probability that a man aged *x* year dies in a year. Find the probability that out of *n* men A_1, A_2, A_3… A_n each aged *x*, A_1 will die and be the first to die.

Solution: The probability that a man aged *x* year dies in a year = 1 - p
The probability that a man aged *x* year does not die in a year = 1 – p
The probability that out of n men none dies in that year:

$$= \quad (1 - p)\ (1 - p)\ (1 - p) \ldots \text{ n times} = (1 - p)^n$$

The probability that at least one man dies in that year

$$= \quad 1 - P \text{ (none dies in that year)}$$

$$= \quad [1 - (1 - p)^n]$$

Also the probability that out of n men, A_1 will die is $\frac{1}{n}$.

Thus, required probability = $\frac{1}{n}\ [1 - (1 - p)^n]$

Example 4.23: A and B are two independent witnesses. The probability that A will speak truth is x and the probability that B will speak the truth is y. A and B agree with the statement. Find the probability that the statement is true.

Solution: Given, P(A) = x, P(B) = y
P($\bar{A}$) = 1 – x, P($\bar{B}$) = 1 – y,
A and B both agree when (i) either of them speaking the truth or (ii) making false statements.

The probability that both A and B speaks truth = P(A) . P(B) = xy
The probability that both A and B makes false statements
= P($\bar{A}$) . P($\bar{B}$)
= (1 - x) (1 - y)
Thus, the total number of case agreeing both = xy + (1 - x) (1 - y)

$$P \text{ (the statement is true)} = \frac{\text{No. of cases speaking the truth}}{\text{Total no. of cases}}$$

$$= \frac{xy}{xy + (1 - x)\ (1 - y)}$$

4.3.1.2 Conditional Probability

The multiplication theorem discussed above is not applicable in case of dependent events. Dependent events are those in which the occurrence of one event affects the probability of other events. The probability of the events B given that A has occurred is called the conditional probability of B. It is denoted by P(B/A). Similarly, the conditional of A given that B has occurred is denoted by P(A/B).

Definition of Conditional Probability

If A and B are two dependent events, then the conditional probability of B given A is defined and given by:

$$\mathbf{P(B/A)} = \frac{P(AB)}{P(A)} \qquad \text{provided } P(A) > 0$$

Similarly, the conditional probability of A given B is defined and given by:

$$\mathbf{P(A/B)} = \frac{P(AB)}{P(B)} \qquad \text{provided } P(B) > 0$$

4.3.2 Multiplication Theorem for Dependent Events or Multiplication Theorem in Case of Conditional Probability

When the events are not independent, i.e., they are dependent events, then the multiplication theorem has to be modified. The Modified Multiplication Theorem sates that of A and B are two dependent events, then the probability of their simultaneous occurrence is equal to the probability of one event multiplied by the conditional probability of the other.

Symbolically,

$$P(AB) = P(A) \cdot P(B/A)$$

or $$P(AB) = P(B) \cdot P(A/B)$$

Where, P(B/A) = Conditional Probability of B given A

P(A/B) = Conditional Probability of A given B

Example 4.24: A bag contains 10 white and 5 black balls. Two balls are drawn at random one after the other without replacement. Find the probability that both balls drawn are black.

Solution: The probability of drawing a black ball in the first attempt is:

$$P(A) = \frac{5}{10+5} = \frac{5}{15}$$

The probability of drawing the second black ball given that the first drawn is black and not replaced is:

$$P(B/A) = \frac{4}{10+4} = \frac{4}{14}$$

Since, the events are dependent, so the probability that both balls drawn are black is:

$$P(AB) = P(A) \cdot P(B/A)$$

$$= \frac{5}{15} \text{ x } \frac{4}{14} = \frac{2}{21}$$

Example 4.25: Find the probability of drawing a king, a queen and a knave in that order from a pack of cards in three consecutive draws, the cards drawn not being replaced.

Solution: The probability of drawing a king = $P(A) = \frac{4}{52}$

The probability of drawing a queen after a king has been drawn

$$P(B/A) = \frac{4}{51}$$

The probability of drawing a knave after a king and a queen has been drawn

$$P(C/AB) = \frac{4}{50}$$

Since, the events are dependent, the required probability of drawing a king, a queen and ace in that order is:

$$P(ABC) = \frac{4}{52} \times \frac{4}{51} \times \frac{4}{50} = \frac{8}{16{,}575}$$

Example 4.26: Four cards are drawn without replacement. What is the probability that they are all aces?

Solution: Probability of drawing an ace in the first attempt = $\frac{4}{52}$

Probability of drawing 2nd ace after the I ace has been drawn = $\frac{3}{51}$

Probability of drawing 3rd ace after the I and II aces have been drawn = $\frac{2}{50}$

Probability of drawing 4th ace after the I, II and III aces have been drawn

$$= \frac{1}{49}$$

Since, the events are dependent, the required probability is:

$$P\text{ (1st Ace x 2nd Ace x 3rd Ace x 4th Ace)} = \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50} \times \frac{1}{49} = \frac{1}{270725}$$

Example 4.27: A bag contains 5 white and 8 red balls. Two successive drawings of 3 balls are made such that (i) the balls are replaced before the second trial, and (ii) the balls are not replaced before the second trial. Find the probability that the first drawing will give 3 white and the second 3 red balls in each case.

Solution: **(i) When balls are replaced**

Total balls in a bag = 8 +5 = 13

3 balls can be drawn out of 13 balls in ${}^{13}C_3$ ways.

3 white balls can be drawn out of 5 white balls in ${}^{5}C_3$ ways.

Probability of 3 white balls = P(3W) = $\frac{{}^{5}C_3}{{}^{13}C_3}$

Since, the balls are replaced after the first drawn so again there are 13 balls in the bag 3 red balls can be drawn out of 8 red balls in ${}^{8}C_3$ ways.

Probability of 3 red balls = P(3R) = $\frac{{}^{8}C_3}{{}^{13}C_3}$

Since, the events are independent, the required probability is:

$$P(3W \text{ and } 3R) = P(3W) \times P(3R)$$

$$= \frac{^5C_3}{^{13}C_3} \text{ x } \frac{^8C_3}{^{13}C_3} = \frac{10}{286} \text{ x } \frac{56}{286} = \frac{140}{20449}$$

(ii) When the balls are not replaced before second draw

Total balls in a bag = 8 + 5 = 13

3 balls can be drawn out of 13 balls in $^{13}C_3$ ways.

3 white balls can be drawn out of 5 white balls in $^5C_3$ ways.

The probability of drawing 3 white balls = P(3W) = $\frac{^5C_3}{^{13}C_3}$

After the first drawn, balls left are 10.3 balls can be drawn out of 10 balls in $^{10}C_3$ ways.

3 red balls can be drawn out of 8 balls in $^8C_3$ ways.

Probability of drawing 3 red balls = $\frac{^8C_3}{^{10}C_3}$

Since, both the events are dependent, the required probability is:

$$\text{P(3W and 3R)} = \frac{^5C_3}{^{13}C_3} \text{ x } \frac{^8C_3}{^{10}C_3} = \frac{5}{143} \text{ x } \frac{7}{15} = \frac{7}{429}$$

Example 4.28: A bag contains 5 white and 3 red balls and four balls are successively drawn out and not replaced. What is the chance that (i) white and red balls appear alternatively and (ii) red and white balls appear alternatively?

Solution: (i) The probability of drawing a white ball = $\frac{5}{8}$

The probability of drawing a red ball = $\frac{3}{7}$

The probability of drawing a white ball = $\frac{4}{6}$

The probability of drawing a red ball = $\frac{2}{5}$

Since, the events are dependent, therefore the required probability is:

$$\text{P(1W 1R 1W 1R)} = \frac{5}{8} \text{ x } \frac{3}{7} \text{ x } \frac{4}{6} \text{ x } \frac{2}{5} = \frac{1}{14}$$

(ii) The probability of drawing a red ball = $\frac{3}{8}$

The probability of drawing a white ball = $\frac{5}{7}$

The probability of drawing a red ball = $\frac{2}{6}$

The probability of drawing a white ball = $\frac{4}{5}$

Since, the events are dependent, the required probability is:

$$P(1R\ 1W\ 1R\ 1W) = \frac{3}{8} \times \frac{5}{7} \times \frac{2}{6} \times \frac{4}{5} = \frac{1}{14}$$

4.4 COMBINED USE OF ADDITION AND MULTIPLICATION THEOREMS

Under probability, there are certain problems where both addition and multiplication theorems are used simultaneously. In such cases, you first apply multiplication theorem and then ultimately you apply addition theorem.

Example 4.29: A speaks truth in 80% cases, B in 90% cases. In what percentage of cases are they likely to contradict each other in stating the same fact?

Solution: Let P(A) and P(B) denote the probability that A and B speak the truth. Then,

$$P(A) = \frac{80}{100} = \frac{4}{5}, \qquad P(\bar{A}) = 1 - P(A) = 1 - \frac{4}{5} = \frac{1}{5}$$

$$P(B) = \frac{90}{100} = \frac{9}{10}, \qquad P(\bar{B}) = 1 - P(B) = 1 - \frac{9}{10} = \frac{1}{10}$$

They will contradict each other only when one of them speaks the truth and the other speaks a lie.

Thus, there are two possibilities:

(i) A speaks the truth and B tells a lie

(ii) B speaks the truth and A tells a lie.

Since, the events are independent, so by using multiplication theorem, you have

(i) Probability in the 1st case = $\frac{4}{5} \times \frac{1}{10} = \frac{4}{50}$

(ii) Probability in the 2nd case = $\frac{9}{10} \times \frac{1}{5} = \frac{9}{50}$

Since the cases are mutually exclusive, so by using addition theorem, you have

Required Probability = $\frac{4}{50} + \frac{9}{50} = \frac{13}{50} = 26\%$

Example 4.30: A bag contains 5 white and 4 black balls. A ball is drawn from this bag and it replaced and then second draw of a ball is made. What is the probability that two balls are of different colours (i.e., one is white and one is black)?

Solution: There are two possibilities:

(i) 1st ball drawn is white and the second drawn in black.

(ii) 1st ball drawn is black and the second drawn in white.

Since, the events are independent, so by using multiplication theorem, you have

(i) Probability in the 1st case = $\frac{5}{9} \times \frac{4}{9} = \frac{20}{81}$

(ii) Probability in the 2nd case = $\frac{4}{9} \times \frac{5}{9} = \frac{20}{81}$

Since, these possibilities are mutually exclusive, so by using addition theorem, you have

Required Probability = $\frac{20}{81} + \frac{20}{81} = \frac{40}{81}$

Example 4.31: A bag contains 5 white and 3 red balls and four balls are successively drawn out and not replaced. What is the chance that they are alternatively of different colours?

Solution: 4 balls of alternative colours can be white, red, white, red or red, white, red, white.

Beginning with White Ball:

The probability of drawing a white ball = $\frac{5}{8}$

The probability of drawing a red ball = $\frac{3}{7}$

The probability of drawing a white ball = $\frac{4}{6}$

The probability of drawing a red ball = $\frac{2}{5}$

Since, the events are dependent, so by using multiplication theorem, you have

P(1W 1R 1W 1R) = $\frac{5}{8} \times \frac{3}{7} \times \frac{4}{6} \times \frac{2}{5} = \frac{1}{14}$

Beginning with Red Ball:

The probability of drawing a red ball = $\frac{3}{8}$

The probability of drawing a white ball = $\frac{5}{7}$

The probability of drawing a red ball = $\frac{2}{6}$

The probability of drawing a white ball = $\frac{4}{5}$

Since, the events are dependent, so by using multiplication theorem, you have

P(1R 1W 1R 1W) = $\frac{3}{8} \times \frac{5}{7} \times \frac{2}{6} \times \frac{4}{5} = \frac{1}{14}$

Since, (i) and (ii) cases are mutually exclusive, so by using addition theorem, you have

Required Probability = $\frac{1}{14} + \frac{1}{14} = \frac{2}{14} = \frac{1}{7}$

Example 4.32: A six-faced die is so biased that it is twice as likely to show an even number as an odd number when it is thrown twice. What is the probability that the sum of two numbers thrown is even?

Solution: Let p be the probability of getting an even number in a single throw of a die and q be that of an odd number.

Given: Even Number: Odd Number:: 2: 1

$$p = P(Even) = \frac{2}{3}, \; q = P(Odd) = \frac{1}{3}$$

There are two mutually exclusive cases in which the sum of two numbers may be even:

(i) Odd number in the first throw and again an odd number in the second throw.

(ii) An even number in the first throw and again an even number in the second throw.

Since the events are independent, so by using the multiplication theorem, you have

(i) Probability in the 1st case = $\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$

(ii) Probability in the 2nd case = $\frac{2}{3} \times \frac{2}{3} = \frac{4}{9}$

Since, these possibilities are mutually exclusive, so by using the addition theorem, you have:

Required Probability = $\frac{1}{9} + \frac{4}{9} = \frac{5}{9}$

Example 4.33: Three groups of workers contain 3 men and 1 woman, 2 men and 2 women, and 1 man and 3 women. One worker is selected at random from each group. What is probability that the group selected consist of 1 man and 2 women?

Solution: There are three possibilities in this case:

(i) 1 man is selected from the first group and 1 woman each from 2nd and 3rd group.

(ii) 1 man is selected from the 2nd group and 1 woman each from 1st and 3rd group.

(iii) 1 man is selected from the 3rd group and 1 woman each from 1st and 2nd group.

Since, the events are independent, so by using multiplication theorem, you have

(i) Probability in the 1st case = $\frac{3}{4} \times \frac{2}{4} \times \frac{3}{4} = \frac{18}{64}$

(ii) Probability in the 2nd case = $\frac{2}{4} \times \frac{1}{4} \times \frac{3}{4} = \frac{6}{64}$

(iii) Probability in the 3rd case = $\frac{1}{4} \times \frac{1}{4} \times \frac{2}{4} = \frac{2}{64}$

Since these possibilities are mutually exclusive, so by using the addition theorem, you have:

$$\text{Required Probability} = \frac{18}{64} + \frac{6}{64} + \frac{2}{64} = \frac{26}{64} = \frac{13}{32}$$

4.5 USE OF BERNOULLI'S THEOREM IN THEORY OF PROBABILITY

Bernoulli's theorem is very useful in working out various probability problems. This theorem states that if the probability of happening of an event in one trial or experiment is known, then the probability of its happening exactly, 1, 2, 3,…r times in n trials can be determined by using the formula:

$$P(r) = {}^{n}C_r\, p^r \,.\, q^{n-r} \qquad r = 1, 2, 3, \ldots n$$

where,

P(r) = Probability of r successes in n trials.

p = Probability of success or happening of an event in one trial.

q = Probability of failure or not happening of an event in one trial.

n = Total number of trials.

The following examples illustrate the applications of this theorem

Example 4.34: Three coins are tossed simultaneously. What is the probability that there will be exactly two heads?

Solution: Since you have to find the probability of exactly two heads, the use of the Bernoulli Theorem will be convenient. According to this theorem:

$$P(r) = {}^{n}C_r\, p^r \,.\, q^{n-r}$$

Given, n = 3, r = 2, p = probability of head in throw of one coin = $\frac{1}{2}$

$$P(2H) = {}^{3}C_2 \left(\frac{1}{2}\right)^2 . \left(\frac{1}{2}\right)^{3-2}$$

$$= \frac{3!}{(3-2)!\,2!} \text{ x } \frac{1}{8} = 3 \text{ x } \qquad \frac{1}{8} = \frac{3}{8}$$

Example 4.35: In an army battalion 3/5 of the soldiers are known to be married and the reminder 2/5 unmarried. Calculate the probability of getting exactly 4 married soldiers in a row of 5 soldiers.

Solution: Since, you have to fine the probability of exactly 4 married soldiers, the use of Bernoulli Theorem will be more convenient. According to this theorem,

$$P(r) = {}^{n}C_r\, p^r \,.\, q^{n-r}$$

Given, n = 5, r = 4, p = probability of married soldiers = $\frac{3}{5}$

$$q = 1 - p = 1 - \frac{3}{5} = \frac{2}{5}$$

$$P(\text{4 married soldiers}) = {}^5C_4 \left(\frac{3}{5}\right)^4 \cdot \left(\frac{2}{5}\right)^1$$

$$= \frac{5!}{4!\,1!} \cdot \frac{3 \times 3 \times 3 \times 3}{5 \times 5 \times 5 \times 5} \times \frac{2}{5} = \frac{162}{625}$$

Example 4.36: If there are three children in a family, find the probability that there is one girl in the family.

Solution: Given, n = 3, r = 1, p = probability of a girl child = $\frac{1}{2}$

$$q = 1 - \frac{1}{2} = \frac{1}{2}$$

$$P(r=1) = P(1G) = {}^3C_1 \left(\frac{1}{2}\right)^1 \cdot \left(\frac{1}{2}\right)^2 = \frac{3!}{2!\,1!} \times \frac{1}{8} = \frac{3}{8}$$

Example 4.37: The probability that India wins a cricket test match against England is given to be $\frac{1}{3}$. If India and England play three test matches, find the probability that (i) India will lose all three matches and (ii) India will win at least one test match.

Solution: Given, n = 3,

$$p = \text{Probability of winning the match} = \frac{1}{3}$$

$$q = 1 - \frac{1}{3} = \frac{2}{3}$$

(i) $P(\text{losing all matches}) = P(0) = {}^3C_0 \left(\frac{1}{3}\right)^0 \cdot \left(\frac{2}{3}\right)^3 = \frac{8}{27}$

(ii) P (at least win one test match) = 1 - P(does not win none)

$$= 1 - {}^3C_0 \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^3$$

$$= 1 - \frac{8}{27}$$

$$= \frac{19}{27}$$

4.6 BAYES' THEOREM

Bayes' Theorem is named after the British Mathematician Thomas Bayes and it was published in the year 1763. With the help of Bayes' Theorem, prior probabilities are revised in the light of some sample information and posterior probabilities are obtained. This theorem is also called the Theorem of Inverse Probability. Suppose in a factory, two machines A_1 and A_2 are manufacturing goods. Further suppose that machine A_1 and A_2 manufacture, respectively 70% and 30% of the total with 5% and 3% of total defective bolts. Suppose an item is selected from the total production and found to be defective. And if you want to find out the probability that

it was manufactured by machine A_1 or machine A_2, then this can be found by using Bayes' Theorem. Suppose, an urn contains 6 black and 4 white balls. Another urn contains 45 black and 6 white balls. A ball is drawn from one of the urns and found to be black. If you want to find out the probability that it came from 1st urn or 2nd urn. This can be found by using Bayes' Theorem.

Statement of Bayes' Theorem: If A_1 and A_2 are mutually exclusive and exhaustive events and B be an event which can occur in combination with A_1 and A_2, then the conditional probability for event A_1 and A_2 given the event B is given by:

$$P(A_1 / B) = \frac{P(A_1).P(B/A_1)}{P(A_1).P(B/A_1) + P(A_2).P(B/A_{2)}}$$

Similarly,

$$P(A_2 / B) = \frac{P(A_2).P(B/A_2)}{P(A_1).P(B/A_1) + P(A_2).P(B/A_{2)}}$$

Generalization: Bayes' Theorem can be extended to three or more events. If A_1, A_2 and A_3 are three mutually exclusive events and B is an event which can occur in combination with A_1, A_2 and A_3 then

$$P(A_1 / B) = \frac{P(A_1).P(B/A_1)}{P(A_1).P(B/A_1) + P(A_2).P(B/A_2) + P(A_3).P(B/A_3)}$$

$$P(A_2 / B) = \frac{P(A_2).P(B/A_2)}{P(A_1).P(B/A_1) + P(A_2).P(B/A_2) + P(A_3).P(B/A_3)}$$

$$P(A_3 / B) = \frac{P(A_3).P(B/A_3)}{P(A_1).P(B/A_1) + P(A_2).P(B/A_2) + P(A_3).P(B/A_3)}$$

Example 4.38: In a bolt factory machine A, B and C manufacture respectively 25%, 35% and 40% of the total. Of their output 5, 4, 2 per cent are defective bolts. A bolt is drawn at random from the product and is found to be defective. What is the probability that it was manufactured by machine C?

Solution: Let A, B and C be the events of drawing a bolt produced by machine A, B and C respectively and let D be the event that the bolt is defective.

You are given the information:

$P(A) = 25\% = \frac{25}{100} = 0.25$

$P(B) = 35\% = \frac{35}{100} = 0.35$

$P(C) = 40\% = \frac{40}{100} = 0.40$

The conditional probabilities are:

$P(D/A) = 5\% = \frac{5}{100} = 0.05$

$P(D/B) = 4\% = \frac{4}{100} = 0.04$

$P(D/C) = 2\% = \frac{2}{100} = 0.02$

Putting the given information in the table given below:

Events (1)	Prior Probabilities (2)	Conditional Probabilities (3)	Joint Probabilities Col. (2) x (3)
A	P(A) = 0.25	P(D/A) = 0.05	0.25 x 0.05
B	P(B) = 0.35	P(D/B) = 0.04	0.35 x 0.04
C	P(C) = 0.40	P(D/C) = 0.02	0.40 x 0.02

You have to calculate P(C/D), i.e., the probability that the defective item was produced by machine C.

$$P(C/D) = \frac{\text{Joint Probabilit y of the machine C}}{\text{Sum of Joint Probabilit y of three machines}}$$

$$= \frac{0.40 \text{ x } 0.02}{0.25 \text{ x } 0.05 + 0.35 \text{ x } 0.04 + 0.40 \text{ x } 0.02}$$

$$= \frac{0.008}{0.0125 + 0.014 + 0.008} = \frac{0.008}{0.0345}$$

$$= 0.2318 \text{ or } 23.18\ \%$$

Example 4.39: A manufacturing firm produces steel pipes in three plants with daily production volumes of 500, 1000 and 2000 units, respectively. According to past experience, the fraction of defective output produced by three plants: 0.005, 0.008, and 0.010, respectively. If a pipe is selected from a day's total production and found to be defective, find the probability that it was manufactured by the first plant.

Solution: Let E_1, E_2 and E_3 be the events of selecting steel pipes by plants I, II and III and let D be the event that the pipe is defective.

You are given the information: Conditional probabilities are:

$P(E_1) = \frac{500}{500+1000+2000} = \frac{1}{7}$ $P(D/E_1) = 0.005$

$P(E_2) = \frac{100}{500+1000+2000} = \frac{2}{7}$ $P(D/E_2) = 0.008$

$P(E_3) = \frac{2000}{500+1000+2000} = \frac{4}{7}$ $P(D/E_3) = 0.010$

Putting the given information in the table given below:

Events (1)	Prior Probabilities (2)	Conditional Probabilities (3)	Joint Probabilities Col. (2) x (3)
E_1	$P(E_1) = 1/7$	$P(D/E_1) = 0.005$	1/7 x 0.005
E_2	$P(E_2) = 2/7$	$P(D/E_2) = 0.008$	2/7 x 0.008
E_3	$P(E_3) = 4/7$	$P(D/E_3) = 0.010$	4/7 x 0.010

You have to calculate $P(E_1/D)$, i.e., the probability that the defective pipe was produced by the Ist plant.

$$P(E_1/D) = \frac{\text{Joint Probabilit y of the Ist Plant}}{\text{Sum of Joint Probabilit y of three plants}}$$

$$= \frac{\frac{1}{7} \text{ x } 0.005}{\frac{1}{7} \text{ x } 0.005 + \frac{2}{7} \text{ x } 0.008 + \frac{4}{7} \text{ x } 0.010}$$

$$= \frac{0.005}{0.005 + 0.016 + 0.040} = \frac{0.005}{0.061} = \frac{5}{61}$$

Example 4.40: A, B and C are three candidates for the post of director in a company. Their respective chances of selection are in the ratio of 4:5:3. The probability that A, if selected will introduce internet trading in the company is 0.30. Similarly, the probability of B and C are 0.50 and 0.6 respectively. Find the probability that the company will introduce internet trading. Also, find the probability that Director B introduce internet trading in the company.

Solution: Let A_1, A_2 and A_3 denote the events that the person's A, B and C respectively are selected as Director of the Company and let E be the event of introducing internet trading in the company. Then you are given:

$P(A_1) = \frac{4}{4+5+3} = \frac{4}{12}$ $P(A_2) = \frac{5}{4+5+3} = \frac{5}{12}$

$P(A_3) = \frac{3}{4+5+3} = \frac{3}{12}$

$P(E/A_1) = 0.30$ $P(E/A_2) = 0.50$ $P(E/A_3) = 0.60$

Putting the given information in the table given below:

Events (1)	**Prior Probabilities (2)**	**Conditional Probabilities (3)**	**Joint Probabilities Col. (2) x (3)**
A_1	$P(A_1) = \frac{4}{12}$	$P(E/A_1) = 0.30$	$\frac{4}{12}$ x 0.30
A_2	$P(A_2) = \frac{5}{12}$	$P(E/A_2) = 0.50$	$\frac{5}{12}$ x 0.50
A_3	$P(A_3) = \frac{3}{12}$	$P(E/A_3) = 0.60$	$\frac{3}{12}$ x 0.60

(i) Now, P (Internet trading is introduce in the company)

$P(E) = P(A_1 E \text{ or } A_2 E \text{ or } A_3 E) = P(A_1 E) + P(A_2 E) + P(A_3 E)$

$= P(A_1) .P(E/A_1) + P(A_2) .P(E/A_2) + P(A_3) .P(E/A_3)$

$= \frac{4}{12} \text{ x } 0.30 + \frac{5}{12} \text{ x } 0.50 + \frac{3}{12} \text{ x } 0.60 = \frac{55}{120} = \frac{11}{24}$

(ii) You have to find $P(A_2/E)$, i.e., internet trading is introducing by Director B By Bayes' Theroem, you have

P(Director B introduces internet trading)

$$P(A_2/E) = \frac{\text{Joint probability of the 2nd}}{\text{Sum of the joint probability}}$$

$$= \frac{\frac{5}{12} \text{ x } 0.50}{\frac{4}{12} \text{ x } 0.30 + \frac{5}{12} \text{ x } 0.50 + \frac{3}{12} \text{ x } 0.60} = \frac{\frac{5}{12} \text{ x } \frac{1}{2}}{\frac{11}{24}} = \frac{5}{11}$$

4.7 SUMMARY

In total, the addition theorem on probability is used for mutually exclusive events and not mutually exclusive events. The multiplication theorem on probability is used for independent and dependent events whether they are one or more. It is also used in case of conditional probability of events. Further, the Bernoulli's theorem is very useful in working out various probability of happening of an event in one trial or experiment and Bayes' Theorem is used to solve the problems of probability using prior probabilities and posterior probabilities.

4.3 GLOSSARY

Mutually Exclusive Events: Two events are mutually exclusive if they cannot occur at the same time.

4.4 CHECK YOUR PROGRESS

1……. are those in which the probability of occurrence of one event affects the probability of the other events.
2……..=P(A)x P(B)
3. Bayes' Theorem is named after the British Mathematician…….

4.5 ANSWER TO CHECK YOUR PROGRESS

1. Independent events 2. P(AB) 3. Thomas Bayes

4.6 TERMINAL QUESTIONS

1. A card is drawn from a pack of 52 cards. What is the probability of getting either a heart or queen of spade?

2. In a class of 25 students with role numbers 1 to 25, a student is picked up at random to answer the question. Find the probability that roll number of the students is either a multiple 5 or 7.

3. A bag contains 3 red, 6 white, 4 blue and 7 yellow balls. A ball is drawn. What is the probability that the ball will be either white or yellow?

4. In a single throw of three dice, find the probability of getting a total of 17 or 18.

5. In a single throw of 2 dice, find the probability of getting a total of 9 or 11.

6. What is the probability of drawing a heart or a king card from a pack of cards?

7. A bag contains 50 balls numbered from 1 to 50. One ball is drawn at random. Find the probability that a drawn ball is a multiple of 5 or 7.

8. The probability that a contractor will get a plumbing contract is 2/3 and probability that he will not get an electric contract is 5/9. If the probability of getting at least one contract is 4/5, what is probability that he will get both?

9. A student applies for a job in two firms X and Y. The probability of his being selected in a firm X is 0.7 and being rejected in the firms Y is 0.5. The probability that his application is being rejected in both the firms is 0.6. What is the probability that he will selected in one of the firms?

10. The result of an examination given to a class on 3 papers, A, B and C are given. It is estimated that 40% failed 30% failed in paper B, 25% failed in paper C, 15 % failed in paper A and B both , 12 % failed in paper B and C both, 10% failed in paper A and C both and 3% failed in all the papers. What is the probability of randomly selected candidate passing in at least one of three papers?

11. Find the probability of getting 3 tails in 3 tosses of a coin.

12. Three airplanes fly from Bombay to London. Odds in favour of their arriving safety are 2:1, 3:1, and 4:1. Find the probability that they all arrive safely.

13. A husband and a wife appear in an interview for 2 vacancies for the same post. The probability of selection of husband is 4/5 and that if wife is 3/4. Find the probability that (i) both of them will be selected (ii) none of them selected and (iii) only wife will be selected.

14. The odds in favour of passing driving test by Mohan are 3:5 and odds in favour of passing the same test by Ram is 3:2. What is the probability that both will pass the test?

15. A university has to appoint examiners to evaluate paper in Statistics. Out of a panel of 40 examiners, 10 are women, 30 of them knowing Hindi and 5 of them are Ph.D. Find the probability of selecting a Hindi knowing Ph.D. women teacher to evaluate the papers.

16.A problem in Statistics is given to four students. Their chances of solving it are 1/2, 1/3, 1/4 and 1/5, respectively. What is the probability that the problem is solved?

17.A and B decide to meet at Durga Temple between 5 to 7 p.m. with the condition that no one would wait for the other for more than 30 minutes. What is the probability that they meet?

18.The probability that a boy will get a scholarship 0.90 and that a girl will get is 0.80. What is the probability that at least one of them will get the scholarship?

19.Find the probability of throwing 6 at least once in three tosses of a die.

20.A bag contains 6 white and 4 black balls. Two balls are drawn at random one after another without replacement. Find the probability that both drawn balls are white.

21.A bag contains 7 red, 5 white and 4 blue balls. Three balls are drawn successively. Find the probability that these are drawn, in order of red, white and blue if the drawn ball is not replaced.

22.Find the probability of drawing a king and an ace in this order from a pack of cards in two successive draws assuming that first card drawn is not replaced.

23.A box contains 8 tickets bearing the following numbers: 1, 2, 3, 4, 5, 6, 8 and 10. One ticket is drawn at random and kept side. Then a second ticket is drawn. Find the probability that both the tickets show even numbers.

24. A bag contains 5 white and 4 black balls and 4 balls are successively drawn out and not replaced. What is the chance that white and black balls appear alternatively?

25. The odds that A speaks truth is 3:2 and the odds that B speaks the truth is 5 : 3. In what percentage of case are they like to contradict each other?

26. A box contains 10 white and 5 black balls and 4 balls are successively drawn and not replaced. Find the probability that they are alternatively of different colours.

27. There are two bags. One bag contains 4 white and 2 black balls. The second bag contains 5 white and 4 black balls. Two balls are transferred from first bag to second bag. Then one ball is taken from the second bag. Find the probability that it is white ball.

28. A purse contains 2 silver and 4 copper coins. A second purse contains 4 silver and 3 cooper coins. If a coin is picked at random from one of the two purses, what is the probability that it is a silver coin?

29. The odds that A speaks the truth are 3:2 and the odds that B does so are 5:3. In what percentage cases are they likely to contradict each other in stating the same fact?

30.The chance that a ship arrives safely at a port is $\frac{9}{10}$. Find the probability that out of 5 ships expected exactly 4 will arrive safely

31.What is the probability of getting exactly 3 heads in five throws of a single coin?

32.If three coins are tossed simultaneously then what is the probability that they will fall alike?

33.Eight coins are tossed simultaneously. What is the probability that they will fall 6 heads and 2 tails up?

34. Find the probability of having at least one head in 5 throws with a coin.

35.and machine II produces 70% of the items. Further, 5% of the items produce by the machine I were defective and only 1% produced by machine II were defective. If a defective item is drawn at random, what is the probability that it was produced by machine I?

36.There are 4 boys and 2 girls in Room No. I and 5 boys and 3 girls in Room No. II. A girl from one of two rooms laughed loudly. What is the probability that the girl who laughed loudly was from Room No. II?

37.A purse contains three one rupee coins and four 50 paise coins. Another purse contains four one-rupee coins and five 50 paise coins. A one-rupee coin has been taken out from one of the purses. Find out the probability that it is from the first purse.

38.There are two identical boxes containing respectively 4 white and 3 red balls, 3 white and 7 red balls. A box is chosen at random and a ball is drawn from it. If the ball drawn is white, what is the probability that it is from first box?

39. A manufacturing firm produces sheet pipes in three plants with daily production volume of 250, 350 and 400 units respectively. According to past experience, it is known that fraction of defective outputs produce by plants are respectively 0.05, 0.04 and 0.02. If a pipe is selected from a day's total production and found to be defective, find out the probability that it came from 1st machine**.**

ANSWERS OF TERMINAL QUESTIONS

1. [14/52]
2. [8/25]
3. [13/20]
4. [1/54]
5. [1/54]
6. [4/13]
7. [8/25]
8. [14/15]
9. [0.8]
10. [0.39]
11. [1/8]
12. [2/5]
13. [3/5, 1/20, 3/20]
14. [9/40]

15. [3/128]
16. [4/5]
17. [7/16]
18. [0.98]
19. [91/216]
20. [1/3]
21. [1/24]
22. [4/663]
23. [5/14]
24. [5/63]
25. [$\frac{19}{40}$]
26. [$\frac{10}{91}$]
27. [$\frac{19}{33}$]
28. [$\frac{19}{21}$]
29. [47.5%]
30. [$\frac{59049}{1{,}00{,}000}$]
31. [$\frac{5}{16}$]
32. [$\frac{1}{4}$]
33. [$\frac{28}{256}$]
34. [$\frac{31}{32}$]
35. [15/22]
36. [9/17]
37. [27/55]
38. [40/61]
39. [25/69]

4.7 SUGGESTED READINGS

1. Roy Ramendu, 'Principles of Statistics' Prayag Pustak Bhawan, Allahabad.
2. Gupta S. P. & Gupta M. P., 'Business Statistics' Sultan Chand & Sons, New Delhi.
3. Shukla S. M. & Sahai S. P., 'Advanced Statistics' Sahitya Bhawan Publications, Agra.
4. Goon, Gupta and Dasgupta, 'Basic Statistics' World Press Limited – Calcutta.
5. Fundamentals of Business Statistics – Sanchethi and Kappor.
6. Srivastava, Shenoy and Guptha, 'Quantitative Methods in Management'

UNIT:5-BINOMIAL AND POISSON DISTRIBUTIONS

Structure

OBJECTIVES

After studying this unit, you will be able to understand:

i. Observed frequency distribution;
ii. Theoretical or probability distribution;
iii. Binomial probability distribution;
iv. Poisson distribution; and
v. Their applications.

5.1 INTRODUCTION

In statistics, different types of distributions are studied. They are broadly classified into two categories such as observed frequency distribution and theoretical or probability distribution. The first is derived based on the actual observations or experiments whereas, the last one is derived not based on are not obtained by actual observations or experiments.

5.2 OBSERVED FREQUENCY DISTRIBUTION

Observed frequency distribution refers to those frequency distributions which are obtained by actual observations or experiments. For example, the observed distribution of the marks obtained by 70 students of a class is as follows:

Marks :	0-10	10-20	20-30	30-40	40-50
Nos. of Students:	5	15	20	25	5

The observed frequency distributions are generally analyzed by using various statistical devices like average, dispersion, skewness, etc.

5.3 THEORETICAL OR PROBABILITY DISTRIBUTION

Theoretical frequency distribution refers to those distributions which are not obtained by actual observations or experiments but are mathematically deduced under certain assumptions. Theoretical frequency distributions are also called probability distribution or expected frequency distribution. For example, if four coins are tossed 160 times and the probability of getting a head is considered a success, then on the basis of theory of probability, the expected frequency distribution will be as follows:

No. of Success (X)	**Probability (p)**	**Expected Frequency**
0	1/16	160 x 1/16 = 10
1	4/16	160 x 4/16 = 40
2	6/16	160 x 6/16 = 60
3	4/16	160 x 4/16 = 40
4	1/16	160 x 1/16 = 10
	$\sum p = 1$	160

Thus, the theoretical frequency distribution are not based on actual observations but are mathematically deduced under certain assumptions.

5.4 USES OF THEORETICAL FREQUENCY DISTRIBUTION

The uses of theoretical distributions are as follows:

i. Theoretical distribution are useful in analyzing the nature of given distribution under certain assumptions.
ii. The expected frequencies obtained from the theoretical frequency distribution are useful for making logical decision.
iii. Theoretical frequency distribution helps in comparing actual and expected frequencies and then determining whether the difference between the two is significant or due to sampling fluctuations.
iv. Theoretical distribution helps in making predictions, projections and forecasting.

v. Theoretical distributions are useful in solving many business and other problems. Poisson distribution is useful in making important decision regarding quality control. Normal distribution helps in determining the stock of ready-made garments of different sizes.

vi. In such cases, where the actual experiments are not possible or in case of high cost involved in the collection of actual observation, theoretical frequency distribution can be substituted in place of observed frequency distributions.

5.5 TYPES OF THEORETICAL AND PROBABILITY DISTRIBUTIONS

The main types of theoretical distributions are as follows:

A. Discrete Probability Distribution
   (i) Binomial Distributions
   (ii) Poisson Distribution

B. Continuous Probability Distribution (Normal Distribution)

In this unit, you will study only Binomial and Poisson Distribution. The normal distribution will be discussed in the next unit.

5.6 BINOMIAL DISTRIBUTION

The binomial distribution is a discrete probability distribution. This distribution was discovered by Swiss Mathematician James Bernoulli. It is used in such situations where an experiment results in two possibilities – success and failure. A binomial distribution is a discrete probability distribution that expresses the probability of one set of two alternatives: success (*p*) and failure (*q*).

5.6.1 Definition of Binomial Distribution

Binomial distribution is defined and given by the following probability function:

$$P(X = x) = {}^{n}C_x \, q^{n-x} \, . \, p^x$$

Where, p = probability of success, q = probability of failure = 1 - p, n = number of trials, $P(X = x)$ = probability of x successes in n trials.

By substituting the different values of X in the above probability functions of the Binomial distribution, you can obtain the probability of 0, 1, 2, …., n successes as follows:

Number of Success (X)	**Probability of Success P(X-*x*)**
0	${}^{n}C_0 \, q^{n-0} \, . \, p^0 = q^n$
1	${}^{n}C_1 \, q^{n-1} \, . \, p^1 = q^{n-1} . \, p^1$
2	${}^{n}C_2 \, q^{n-2} \, . \, p^2 = \frac{n(n-1)}{2 \text{ x } 1} \, q^{n-2} . \, p^2$
X	${}^{n}C_x \, q^{n-x} \, . \, p^x$

. . . *N*	. . . ${}^{n}C_{n}\, q^{n-n} \,.\, p^{n} = p^{n}$

5.6.2 Condition or Assumptions to apply Binomial Distribution

Binomial distribution can be used only under the following conditions:

(1) **Finite Number of Trials:** Under binomial distribution, an experiment is performed under identical conditions for a finite or fixed number of trials, *i. e*, number of trials is finite.

(2) **Mutually Exclusive Outcomes:** Each trial must result in two mutually exclusive outcomes – success or failure. For example, if a coin is tossed, then either the head (H) may turn up or the tail (T) may turn up.

(3) **The probability of success in each trial is constant:** In each trial, the probability of success, denoted by a *p* remains constant. In other words, the probability of success in different trials does not change. For example, in tossing a coin, the probability of getting a head in each toss remains the same, *i.e.*, p = *P(H)* = 1/2.

(4) **Trials are independent:** In binomial distribution, statistical independent among trials is assumed, *i.e.*, the outcome of any trial does not affect the outcomes of the subsequent trials.

5.6.3 Properties of Binomial Distribution

The following are the important properties or characteristics of binomial distribution:

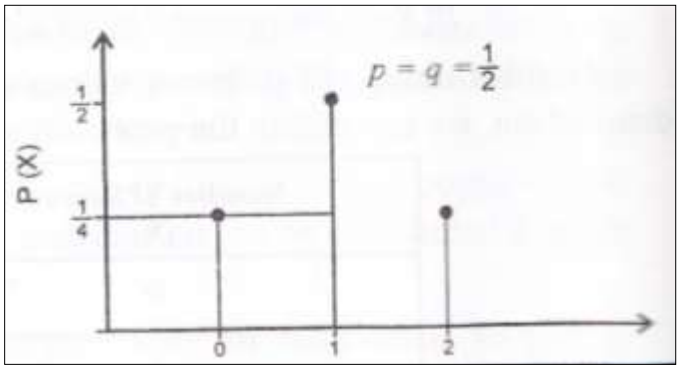
(1) **Theoretical Frequency Distribution:** The binomial distribution is a theoretical frequency distribution which is based on Binomial Theorem of algebra. With the help of this distribution, you can obtain the theoretical frequencies by multiplying the probability of success by the total number (N).

(2) **Discrete Probability Distribution:** The binomial distribution is a discrete probability distribution in which the number of successes 0, 1, 2, 3, …, *n* are given in whole numbers and not in fractions.

(3) **Line Graph:** The binomial distribution can be presented graphically by means of a line graph. The number of successes (X) is taken on the X-axis and the probability of successes (*p*) taken on the Y-axis. The following line graph is based on tossing of a coin twice:

Number of Heads (X)	**Probability P(X =*x*)**
0	${}^{2}C_{0}\left(\frac{1}{2}\right)^{2} = \frac{1}{4}$
1	${}^{2}C_{1}\left(\frac{1}{2}\right)^{1}\left(\frac{1}{2}\right)^{1} = \frac{1}{2}$

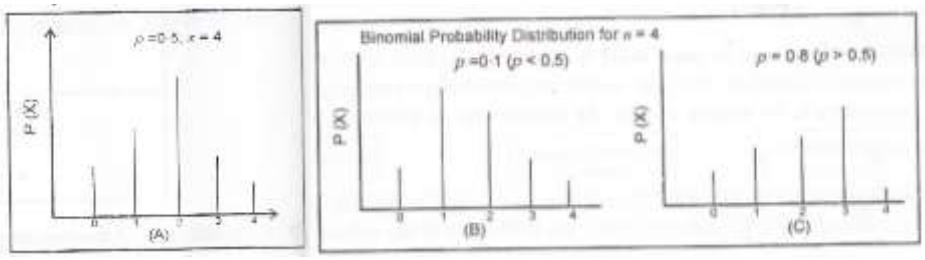
	$^2C_2\left(\frac{1}{2}\right)^2 = \frac{1}{4}$



(4) **Shape of Binomial Distribution:** The shape of Binomial distribution depends on the values of *p* and *q*.

(i) If $p = q = \frac{1}{2}$, then the binomial distribution is symmetrical (see the figure A).

(ii) When $p \neq q \neq \frac{1}{2}$, the binomial distribution is skewed. *i.e.*, asymmetrical. It is positively skewed when $p < q$ *i. e.* $\left(p < \frac{1}{2}\right)$ and negatively skewed when $p > q$ *i. e.* $\left(p > \frac{1}{2}\right)$. See figures (B) and (C) given below:



(5) **Main Parameters:** The binomial distribution has two parameters *n* and *p*. The entire distribution can be known from these two parameters.

(6) **Constants of Binomial Distribution:** The constants of Binomial distribution can be obtained by using the formula.

Mean = $\left(\overline{X}\right) = np$ Moment Coeff. of skewness = $\sqrt{\beta_1} = \frac{q-p}{\sqrt{npq}}$

Variance = $\sigma^2 = npq$

S.D. = $\sigma = \sqrt{npq}$ Moment Coeff. of Kurtosis = $\beta_2 = 3 + \frac{1-6pq}{npq}$

(7) **Uses:** It is also useful in those fields where the outcome is classified into success and failure. In other words, it is useful in coin experiments, dice throwing, manufacturing of items by a company, etc.

5.7 APPLICATION OF BINOMIAL DISTRIBUTION

Now, you will study the applications of binomial distribution under the following ways:

(A) Application of binomial Distribution Formula

When you have given the probability of occurrence of an event relating to a problem, *i.e.,* t he value of *p* and *q* are given, then you can find the probability of the happening of the event exactly *x* times out of *n* trials by using the formula:

$[P(X = x) = {}^{n}C_{x} q^{n-x} p^{x}]$.

Example 5.1: A fair coin is tossed thrice. Find the probability of getting:

(i) exactly 2 Heads
(ii) at least 2 Heads
(iii) at the most 2 Heads

Solution: Let *p* = probability of getting head when a coin is tossed = $\frac{1}{2}$

q = the probability of tail = $\frac{1}{2}$

and *n* = 3, P(X= x) = ${}^{n}C_{x} q^{n-x}.p^{x}$

(i) $P(2H) = {}^{3}C_{2}\left(\frac{1}{2}\right)^{1}\left(\frac{1}{2}\right)^{2} = 3 \text{ x } \frac{1}{2} \text{ x } \frac{1}{4} = \frac{3}{8}$

(ii) P (at least 2 Heads) = P(2H) + P(3H)

$= {}^{3}C_{2}\left(\frac{1}{2}\right)^{1}\left(\frac{1}{2}\right)^{2} + {}^{3}C_{3}\left(\frac{1}{2}\right)^{0}\left(\frac{1}{2}\right)^{3}$

$= 3 \text{ x } \frac{1}{8} + 1 \text{ x } \frac{1}{8} = \frac{4}{8} = \frac{1}{2}$

(iii) P (at most 2 Heads) = P(0H) + P(1H) + P(2H)

= 1 – P(3H)

$= 1 - {}^{3}C_{3}\left(\frac{1}{2}\right)^{0}\left(\frac{1}{2}\right)^{3}$

$= 1 – 1 \text{ x } \frac{1}{8} = 1 - \frac{1}{8} = \frac{7}{8}$

Example 5.2: Four coins are tossed simultaneously. What is the probability of getting *(i)* No head *(ii)* No tail and *(iii)* Two heads only?

Solution: Let *p* = probability of getting head when a coin is thrown = $\frac{1}{2}$

Because, *q* = the probability of tail = 1 – *p* = 1 - $\frac{1}{2} = \frac{1}{2}$

and *n* = 4$P(X = x) = {}^{n}C_{x} q^{n-x}.p^{x}$

(i) $P(0H) = {}^4C_0\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^0 = 1 \text{ x } \frac{1}{16} = \frac{1}{16}$

(ii) $P(0T) = P(4H) = {}^4C_4\left(\frac{1}{2}\right)^0 \cdot \left(\frac{1}{2}\right)^4 = 1 \text{ x } \frac{1}{16} = \frac{1}{16}$

(iii) $P(2H) = {}^4C_2 \cdot \left(\frac{1}{2}\right)^2 \cdot \left(\frac{1}{2}\right)^2 = 6 \text{ x } \frac{1}{16} = \frac{6}{16} = \frac{3}{8}$

Example 5.3: The probability of a bomb hitting a target is 1/5. Two bombs are enough to destroy a bridge. If six bombs are fired at the bridge, find the probability that the bridge is destroyed.

Solution: Let p = probability of bomb hitting a target, q = probability of not hitting the target.

Here, $p = \frac{1}{5}$ $\quad \therefore\ q = \frac{4}{5}$ $\quad (\because q = 1 - p)$

Also, $n = 6$ $\quad P(X = x) = {}^nC_x q^{n-x}.p^x$

The bridge will be destroyed is two or more of 6 bombs hit it.

$\therefore$ Required probability $= P(2) + P(3) + P(4) + P(5) + P(6)$

$= 1 - [P(0) + P(1)]$

$$= 1 - \left[{}^6C_0\left(\frac{4}{5}\right)^6\left(\frac{1}{5}\right)^0 + {}^6C_1\left(\frac{4}{5}\right)^5\left(\frac{1}{5}\right)^1 \right]$$

$$= 1 - \left[1 \text{x} \left(\frac{4}{5}\right)^6 + 6 \text{ x } \frac{(4)^5}{(5)^6} \right] = 1 - \frac{4^6 + 6 \text{ x } 4^5}{5^6}$$

$$= 1 - \frac{10240}{15625} = \frac{15625 - 10240}{15625} = \frac{5385}{15625} = 0.345$$

Example 5.4: The incidence of occupational disease in an industry is such that the workers have 20% chances of suffering from it. What is the probability that out of six workers chosen at random, four or more will suffer from disease?

Solution: Let p = probability of man suffering from disease.

$\therefore \quad p = \frac{20}{100} = \frac{1}{5}$

$\therefore \quad q = 1 - \frac{1}{5} = \frac{4}{5}$

Also $n = 6$

$\therefore \quad P(X = x) = {}^nC_x q^{n-x}.p^x$

Required probability = P(4) + P(5) + P(6)

$$= {}^6C_4\left(\frac{4}{5}\right)^2\left(\frac{1}{5}\right)^4 + {}^6C_5\left(\frac{4}{5}\right)^1\left(\frac{1}{5}\right)^5 + {}^6C_6\left(\frac{4}{5}\right)^0\left(\frac{1}{5}\right)^6$$

$$= 15 \text{ x } \frac{16}{15625} + 6 \text{ x } \frac{4}{15625} + \frac{1}{15625}$$

$$= \frac{240+24+1}{15625} = \frac{265}{15625} = \frac{53}{3125} = 0.01696$$

Example 5.5: Out of 1,000 families with 4 children each, what percentage would you expect to have *(i)* at least one boy *(ii)* at the most 2 girls? Assume equal probabilities for boys and girls.

Solution: Let p = probability for a boy = $\frac{1}{2}$

q = probability for a girl = $\frac{1}{2}$ $\quad n = 4, N = 1000$

(i) **At least one boy:**

P (at least one boy) = P(1B) + P(2B) + P(3B) + P(4B)

= 1 - P (0B)

$$= 1 - {}^4C_0\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^0 = 1 - \frac{1}{16} = \frac{15}{16}$$

Percentage of families with at least one boys = $\frac{15}{16}$ x 100 = 93.75%

(ii) **At least 2 girls:**

P (at least two girls) = P(0G)+P(1G)+P(2G) = P(4B)+P(3B)+ P(2B)

$$= {}^4C_4\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^4 + {}^4C_3\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^3 + {}^4C_2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2$$

$$= \frac{1}{16} + \frac{4}{16} + \frac{6}{16} = \frac{11}{16}$$

Percentage of such families = $\frac{11}{16}$ x 100 = 68.75%

Example 5.6: A pair of dice is thrown 7 times. If getting a total of 7 is considered a success, find the probability of getting *(i)* no success *(ii)* 6 successes and *(iii)* at least 6 successes.

Solution: Let p = probability of getting a total of 7 = $\frac{6}{36} = \frac{1}{6}$

$\therefore\ q = 1 - p = 1 - \frac{1}{6} = \frac{5}{6}\ n = 7,$

$P(X = x) = {}^nC_x q^{n-x}.p^x$

(i) P(0 Success) = $${}^7C_0\left(\frac{5}{6}\right)^7 \cdot \left(\frac{1}{6}\right)^0 = \left(\frac{5}{6}\right)^7$$

(ii) P(6 Successes) = $${}^7C_6\left(\frac{5}{6}\right)^1 \cdot \left(\frac{1}{6}\right)^6 = 35 \text{ x } \left(\frac{1}{6}\right)^7$$

(iii) P(at least 6 successes) = P(6) + P(7)

$$= {}^7C_6\left(\frac{5}{6}\right)^1 \cdot \left(\frac{1}{6}\right)^6 + {}^7C_7\left(\frac{5}{6}\right)^0 \cdot \left(\frac{1}{6}\right)^7$$

$$= 35 \cdot \left(\frac{1}{6}\right)^7 + \left(\frac{1}{6}\right)^7 = 36 \cdot \left(\frac{1}{6}\right)^7 .$$

(B) To Find n, p and q from $\overline{X}$ and σ

When you have the mean ($\overline{X}$ **)** and variance (σ^2) or S.D. (σ) of the binomial distribution, then you can find out *n*, *p* and *q*. The following examples will illustrate the procedure:

Example 5.7: The mean of a binomial distribution is 20 and standard deviation is 4. Find *n*, p and q.

Solution: In a binomial distribution, Mean = np

S.D. = $\sqrt{npq}$

$\overline{X} = np = 20$ …*(i)*

$\sigma = \sqrt{npq} = 4$ …*(ii)*

Squaring both sides,

$\Rightarrow \quad \sigma^2 = npq = 16$ …*(iii)*

Dividing *(iii)* by *(i)*

$$\frac{npq}{np} = \frac{16}{20}$$

$$\Rightarrow \quad q = \frac{16}{20} = \frac{4}{5}$$

$$\therefore \quad p = 1 - q = 1 - \frac{4}{5} = \frac{1}{5}$$

Putting the value of p in *(i)*

$$n \times \frac{1}{5} = 20$$

$$\Rightarrow \quad n = 100$$

Hence, $n = 100,\ p = \frac{1}{5},\ q = \frac{4}{5}$

Example 5.8: Obtain the mean and standard deviation of a binomial distribution for which $P(X = 3) = 16\ P(X = 7)$ and $n = 10$.

Solution: $P(X = 3) = {}^{10}C_3\ q^{10-3}\ p^3 = {}^{10}C_3\ q^7\ p^3$

$P(X = 7) = {}^{10}C_7\ q^{10-7}\ p^7 = {}^{10}C_7\ q^3\ p^7$

As per the question,

$${}^{10}C_3\ q^7\ p^3 = 16\ {}^{10}C_7\ q^3\ p^7$$

$$\Rightarrow \quad q^7\ p^3 = 16\, q^3\ p^7 \qquad [\because {}^{10}C_3 = {}^{10}C_7\ ]$$

$$\Rightarrow \quad q^4 = 16\ p^4 \quad \Rightarrow (q)^4 = (2p)^4 \quad \Rightarrow q = 2p$$

In a binomial distribution

$$p + q = 1 \quad \Rightarrow p + 2p = 1 \quad \Rightarrow p = 1/3$$

$$\therefore \quad q = 1 - p = 1 - \frac{1}{3} = \frac{2}{3}$$

$$\therefore \quad \text{Mean} = np = \frac{10}{3}$$

$$\text{SD} = \sqrt{npq} = \sqrt{\frac{10}{3} \text{ x } \frac{2}{3}} = \frac{\sqrt{20}}{3}$$

Example 5.9: Find the probability of 5 successes in a binomial distribution whose mean and variance are 6 and 2, respectively.

Solution: In a binomial distribution, you have

$$\text{Mean} = np = 6 \quad \ldots(i)$$

$$\text{Variance} = npq = 2 \quad \ldots(ii)$$

Dividing *(ii)* by *(i)*

$$\frac{npq}{np} = \frac{2}{6}$$

$$\therefore \quad q = \frac{1}{3}$$

$$p = 1 - q = 1 - \frac{1}{3} = \frac{2}{3}$$

Substituting the value of p in *(i)*

$$n \text{ x } \frac{2}{3} = 6$$

$$\therefore \quad n = 9$$

Here, $n = 9$, $p = \frac{2}{3}$, $q = \frac{1}{3}$

Now, $$P(X = 5) = {}^{9}C_5 \left(\frac{1}{3}\right)^4 \cdot \left(\frac{2}{3}\right)^5 = 126 \text{ x } \frac{32}{19683} = 0.2048$$

(C) To Find $\overline{X}$ and σ when n, p and q are given

Example 5.10: If the probability of a defective bolt is 0.1, find *(i)* the mean and *(ii)* standard deviation for the distribution of defective bolts in a total of 500. Also find the coefficient of skewness and kurtosis.

Solution: Given, $p = 0.1 \therefore q = 1 - 0.1 = 0.9$, $n = 500$

(i) $$\text{Mean} = np = 500 \text{ x } 0.1 = 500 \text{ x } \frac{1}{10} = 50$$

(ii) $$\text{S.D.} = \sigma = \sqrt{npq} = \sqrt{500 \text{ x } 0.1 \text{ x } 0.9} = 6.70$$

(iii) Coefficient of Skewness $$\left(\sqrt{\beta_1}\right) = \frac{q - p}{\sqrt{npq}} = \frac{0.9 - 0.1}{\sqrt{500 \times 0.1 \times 0.9}} = \frac{0.8}{6.70} = 0.119$$

(iv) Coefficient of Kurtosis (β_2)=3 + $\frac{1-6pq}{npq}$ = 3 + $\frac{1-6(0.1)(0.9)}{500\times 0.1\times 0.9}$ = 3.010

Example 5.11: Find the mean and standard deviation of the number of heads in 100 tosses of a fair coin.

Solution: Given, n = 100 , $P(H) = p = \frac{1}{2}$, $q = \frac{1}{2}$

$\therefore$ Mean = np = 100 x $\frac{1}{2}$ = 50

$$S.D. = \sqrt{npq} = \sqrt{100\times\frac{1}{2}\times\frac{1}{2}} = \sqrt{25} = 5$$

(D) Fitting of Binomial Distribution

The following procedure is adopted while fitting a binomial distribution to the observed data:

(i) Determine the values of p and q from the given information.

(ii) Note the value of n and N, where n is the number of trials in an experiments and N is the total number of trials of all the experiments.

(iii) Find the probability of all possible number of successes coming out of a given experiment.

(iv) Multiply these probabilities by N and the result will be the required expected frequencies.

Example 5.12: Four coins were tossed 160 times and the following results were obtained:

No. of heads:	0	1	2	3	4
Frequency:	17	52	54	31	6

Fit a binomial distribution under the assumption that the coins were unbiased.

Solution: Under the assumption that the coins are unbiased; the probability of head (p) and tail (q) are $\frac{1}{2}$ and $\frac{1}{2}$.

In this case, n = 4, N = 160

The probability of 0, 1, 2, 3, 4 heads will be given by

$$P(X = x) = {}^nC_x q^{n-x}.p^x$$

In order to obtain the expected frequencies, you will have to multiply each probability by N.

The expected frequencies will be obtained as follows:

Number of heads (*n*)	**Expected Frequency** $N\times {}^nC_x\ q^{n-x}\ .\ p^x$
0	$160\times {}^4C_0\left(\frac{1}{2}\right)^4.\left(\frac{1}{2}\right)^0 = 10$

1	$160\times\ {}^4C_1\left(\frac{1}{2}\right)^3.\left(\frac{1}{2}\right)^1=40$
2	$160\times\ {}^4C_2\left(\frac{1}{2}\right)^2.\left(\frac{1}{2}\right)^2=60$
3	$160\times\ {}^4C_3\left(\frac{1}{2}\right)^1.\left(\frac{1}{2}\right)^3=40$
4	$160\times\ {}^4C_4\left(\frac{1}{2}\right)^0.\left(\frac{1}{2}\right)^4=10$

Example 5.13: A survey of 800 families with four children each revealed the following distribution:

No. of Boys:	0	1	2	3	4
No. of Families:	42	178	290	226	64

Fit a Binomial Distribution under the hypothesis that male and female births are equally probable.

Solution: Under the assumption that male and female births are equally probable, the probability of male birth, $p=\frac{1}{2}$

$\therefore \quad q = 1 - \frac{1}{2} = \frac{1}{2}$

In this case, $n = 4$, $N = 800$

The probability of 0, 1, 2, 3, 4 boy's will be given by

$$P(X=x) = {}^nC_x q^{n-x}.p^x$$

The expected frequencies will be obtained multiplying *P(X)* with *N, i.e., N* x *P(X).*

These are given as follows:

Number of Boys (*n*)	**Expected Frequency** $N\times\ {}^nC_x\ q^{n-x}\ .\ p^x$
0	$800\times\ {}^4C_0\left(\frac{1}{2}\right)^4.\left(\frac{1}{2}\right)^0=800\times\frac{1}{16}=50$
1	$800\times\ {}^4C_1\left(\frac{1}{2}\right)^3.\left(\frac{1}{2}\right)^1=800\times\frac{4}{16}=200$
2	$800\times\ {}^4C_2\left(\frac{1}{2}\right)^2.\left(\frac{1}{2}\right)^2=800\times\frac{6}{16}=300$
3	$800\times\ {}^4C_3\left(\frac{1}{2}\right)^1.\left(\frac{1}{2}\right)^3=800\times\frac{4}{16}=200$

4	$800 \times {}^4C_4 \left(\frac{1}{2}\right)^0 . \left(\frac{1}{2}\right)^4 = 800 \times \frac{1}{16} = 50$

5.8 POISSON DISTRIBUTION

Poisson distribution is a discrete probability distribution and it is widely used in statistical work. This distribution was developed by a French Mathematician Dr. Simon Denis Poisson in 1837 and the distribution is named after him. The Poisson distribution is used in those situations where the probability of the happening of an event is very small, *i.e.*, the event rarely occurs. For example, the probability of defective items in a manufacturing company is very small, the probability of occurring earthquake in a year is very small, the probability of the accidents on a road is very small, etc. All these are examples of such events where the probability of occurrence is very small.

5.8.1 Poisson distribution as Limiting Form of Binomial Distribution

Poisson distribution is derived as a limiting form of binomial under certain conditions:
(1) *n*, the number of trials is infinitely large, *i.e.*, $n \to \infty$.
(2) *p*, the probability of success is very small and *q*, the probability of failure is very large, *i.e.*, $p \to 0$, $q \to 1$.
(3) The average number of successes (np) is equal to a positive finite quantity (m) *i.e.*, $np = m$, where, m is the parameter of the distribution.

5.8.2 Definition of Poisson distribution

From binomial equation:

$$ {}^nC_x = \frac{n(n-1)(n-2)...(n-x+1)}{x!} p^x q^{n-x} $$

$Since, np = m \Rightarrow p = \frac{m}{n}$, Therefore, q= 1-$\frac{m}{n}$

$$ = \frac{n(n-1)(n-2)...(n-x+1)}{n^x . x!} n^x p^x q^{n-x} $$

$$ = \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)...\left(1-\frac{x-1}{n}\right) q^{n-x} \cdot \frac{(np)^x}{x!} $$

$$ = \left[\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)...\left(1-\frac{x-1}{n}\right)\left(1-\frac{m}{n}\right)^{n-x}\right] \cdot \frac{m^x}{x!} $$

When, $n \to \infty$ $$ = e^{-m} \cdot \frac{m^x}{x!} = \frac{e^{-m} m^x}{x!} $$

Poisson distribution is defined and given by the following probability function:

$$ P(X = x) = e^{-m} \cdot \frac{m^x}{x!} $$

Where, $P(X = x)$ = probability of obtaining x number of success
$m = np$ = parameter of the distribution
e = 2.7183 (base of natural logarithms)

By substituting the different values of *X* in the above probability function of the Poisson distribution, you can obtain the probability of 0, 1, 2….X successes as follows:

Number of Success (X)	**Probability P(X)**
0	$e^{-m} \cdot \frac{m^0}{0!} = e^{-m}$
1	$e^{-m} \cdot \frac{m^1}{1!} = me^{-m}$
2	$e^{-m} \cdot \frac{m^2}{2!} = \frac{m^2}{2} \cdot e^{-m}$
3	$e^{-m} \cdot \frac{m^3}{3!} = \frac{m^3}{3} \cdot e^{-m}$
X	$e^{-m} \cdot \frac{m^x}{x!}$

5.9 PROPERTIES OF POISSON DISTRIBUTION

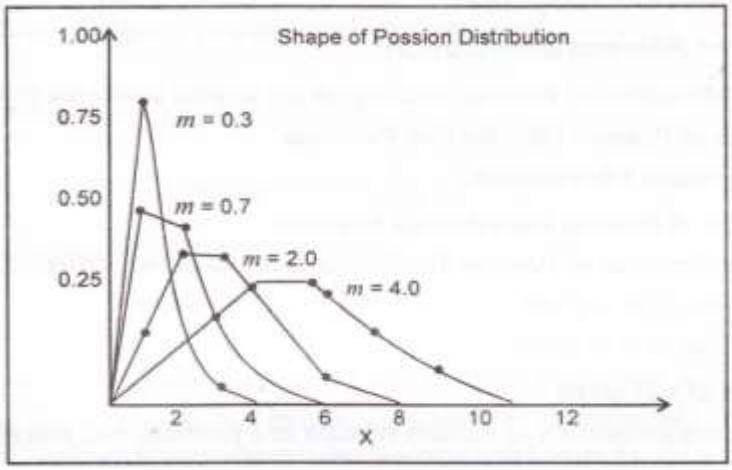
The following are the important properties of Poisson distribution:

(1) **Discrete Probability Distribution:** The Poisson distribution is a discrete probability distribution in which the numbers of successes are given in whole number such as 0, 1, 2…, etc.

(2) **Value of *p* and *q*:** The Poisson distribution is used in those situations where the probability of occurrence of an event is very small (*i.e.,* p→0) and the probability of the non-occurrence of the event is very large (*i.e.,* q→0) and the value of n is also infinitely large.

(3) **Main Parameter:** It has only one parameter *m* and its value is equal to *np, i.e., m = np.* The entire distribution can be known from this parameter.

(4) **Shape of Poisson distribution:** The Poisson distribution is always positively skewed but the skewness decreases as the value of *m* increases. The distribution shifts to the right and degree of skewness falls as *m* increases. It can be seen from the following figure:



(5) Constant of Poisson distribution: The constant of the Poisson distribution can be obtained from the following formula:

Mean = $\overline{X}$ = $m = np$ Moment coeff. of skewness = $\sqrt{\beta_1} = \frac{1}{\sqrt{m}}$

Variance = $\sigma^2 = m$

S.D. = $\sigma = \sqrt{m}$ Moment Coeff. of Kurtosis = $\beta_2 = 3 + 1/m$

(6) Equality of Mean and Variance: An important characteristic of the Poisson distribution is that its mean and variance are equal, *i.e.*,

$\overline{X} = \sigma^2$ or Mean = Variance,

5.10 IMPORTANCE OF POISSON DISTRIBUTION

Poisson distribution is widely used in the following areas:

i. It is used in statistical quality control to count the number of defects of an item.
ii. In Biology, to count the number of bacteria.
iii. In Insurance, problems to count the number of causalities.
iv. To count the number of typing errors per page in a typed material.
v. To count the number of incoming calls in a town.
vi. To count the number of defective blades in a lot of manufactured blades in a factory.
vii. To count the number of deaths at a particular crossing in a town as a result of road accident.
viii. To count the number of suicides committed by lovers point in a year.

In general, the Poisson distribution is useful in rare events where the probability of success (p) is very small and the value of *n* is very large.

5.11 APPLICATIONS OF POISSON DISTRIBUTION

The applications of Poisson distribution are studied as follows:

(A) Application of Poisson distribution Formula

You can study the applications of Poisson distribution formula in two different situations: (1) When the value of p is given and (2) When the value of m is given

(1) When the value of p is given

Example 5.14: It is given that 2% of the screws manufactured by a company are defective. Use Poisson distribution to find the probability that a packet of 100 screws contains: *(i)* no defective screws *(ii)* one defective and *(iii)* two or more defectives. [Given, $e^{-2} = 0.135$]

Solution: Let p = probability of a defective screw = 2% = 2/100

In the usual notation, you have given: $p = 2/100$, $n = 100$

$\therefore \quad m = np = 100 \times 2/100 = 2$

The Poisson distribution is given as:

$$P(X = x) = \frac{e^{-2}.2^{0}}{0!}$$

$$P(X = 0) = e^{-2} = 0.135 \qquad \text{[given: } e^{-2} = 0.135 \text{]}$$

(ii) P (One defective) $= P(X=1) = \dfrac{e^{-2}.2^{1}}{1!}$

$$= e^{-2} \times 2 = (0.135)(2) = 0.270$$

(3) P (Two or more defectives) $= 1 - [P(0) + P(1)]$

$$= 1 - [0.135 + 0.270] = 1 - 0.405 = 0.595$$

Example 5.15: A manufacturer of pins knows that on average 5% of his product is defective. He sells pins in a packet of 100 and guarantees that not more than 4 pins will be defective. What is the probability that a packet will meet the guaranteed quality?

Solution: Let p = probability of a defective pin= 5% = 5/100

Given : $n = 100$, $p = 5/100$

$\therefore \quad m = np = 100 \times 5/100 = 5$

The Poisson distribution is given as:

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

Required probability = P [packet will meet the guarantee]

= P [packet contains up to 4 defectives]

$$= P(0) + P(1) + P(2) + P(3) + P(4)$$

$$= e^{-5}\frac{5^{0}}{0!} + e^{-5}\frac{5^{1}}{1!} + e^{-5}\frac{5^{2}}{2!} + e^{-5}\frac{5^{3}}{3!} + e^{-5}\frac{5^{4}}{4!} +$$

$$= e^{-5}\left[1 + \frac{5}{1} + \frac{25}{2} + \frac{125}{6} + \frac{625}{24}\right]$$

$$= 0.0067 \times 65.374 = 0.438$$

(2) When the value of *m* is given

Example 5.16: Between the hours of 2 P.M. and 4 P.M., the average number of phone calls per minute coming into switch board of a company is 2.5. Find the probability that during one particular minute there will be *(i)* no phone call at all, *(ii)* exactly 3 calls, *(iii)* at least 2 calls. (Given: e^{-2}= 0.1353, $e^{-0.5}$= 0.6065).

Solution: This is a problem of Poisson distribution

$$P(X) = \frac{e^{-m}.m^{x}}{x!} \quad \text{where, X = 1, 2, 3....}$$

Given, Average number of phone calls = $\overline{X} = m = 2.5$

Poisson Distribution is given by

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

(i) P (No call) = P (X = 0) = $\frac{e^{-2.5}.(2.5)^{0}}{0!} = e^{-2.5}$

$= e^{-2}.e^{-0.5}$ (Given: e^{-2}= 0.1353, $e^{-0.5}$= 0.6065)

= 0.1353 x 0.6065 = 0.0821

Hence the probability that during one particular minute there will be no phone call at all is 0.0821.

(ii) P (Exactly 3 calls) = P (X=3) = $\frac{e^{-2.5}.(2.5)^{3}}{3!}$

$= \frac{(0.0821)(15.625)}{3\times 2\times 1} = 0.2138$

(ii) P (At least 2 calls) = 1 – [P (X = 0) + P(X = 1)]

= 1 – [$e^{-2.5}$ + (2.5) $e^{-2.5}$]

= 1 - $e^{-2.5}$ [1 + 2.5] = 1 – [(0.0821) (3.5)]

= 1 – 0.28735 = 0.71265

Example 5.17: It is known from the past experience that the average number of industrial accidents in a factory per month in a plant is 4. Find the probability that during a particular month, there will be less than 4 accidents. Use Poisson distribution to explain your answer (Given: e^{-4}= 0.0183).

Solution: Given, average no. of accidents = $\overline{X} = m = 4$.

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

$$P(0) = e^{-m}.\frac{m^0}{0!} = e^{-m} = e^{-4}$$

$$P(1) = e^{-m}.\frac{m^1}{1!} = me^{-m} = (4)\,e^{-4}$$

$$P(2) = e^{-m}.\frac{m^2}{2!} = \frac{m^2}{2!}.e^{-m} = \frac{(4)^2}{2!}\,e^{-4}$$

$$P(3) = e^{-m}.\frac{m^3}{3!} = \frac{m^3}{3!}.e^{-m} = \frac{(4)^3}{3!} \text{ x } e^{-4}$$

The probability that there will be less than 4 accidents

$$= P(0) + P(1) + P(2) + P(3)$$

$$= e^{-4}.\left[1 + 4 + \frac{4^2}{2!} + \frac{4^3}{3!}\right]$$

$$= 0.0183\,[1 + 4 + 8 + 10.67] \qquad [\because\ e^{-4} = 0.0183]$$

$$= 0.0183 \text{ x } 23.67 = 0.4332$$

$\therefore$ The probability of less than 4 accidents is 0.4332 or 43.32%

Example 5.18: Consider a Poisson probability distribution with the average number of 2 occurrences per time period:

(i) Write the appropriate Poisson probability function.

(ii) What is the average number of occurrence in 3 time periods?

(iii) Find the probability of 6 occurrences in 3 time periods.

Solution: Given, average no. of occurrences for 1 time period = $m = 2$.

(i) Poisson probability function = $P(X = x) = \dfrac{e^{-2}2^x}{x!}$

(ii) Average no. of occurrences for 3 time period = 2 x 3 = 6

(iii) $P[X = 6] = \dfrac{6^{-6}.(6)^6}{6!} = 0.1575$

5.12 FITTING OF POISSON DISTRIBUTION

The following procedure is adopted for fitting a Poisson distribution to the observed data:

(1) Firstly, compute mean ($\overline{X}$) from the observed frequency data by using the formula:

$$\overline{X} = \frac{\sum fX}{N}$$

You will use the value of this mean as the parameter of the Poisson distribution, *i.e.*, $\overline{X} = m$.

(2) The value of e^{-m} is obtained. If the value of e^{-m} is not given in the question, then the following formula is used to compute:

e^{-m} = Reciprocal [Antilog (m x 0.4343)]

(3) Then, compute the probability of 0, 1, 2, 3 or x success by using the Poisson distribution formula:

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

(4) The expected or theoretical frequencies are then obtained by multiply each probability with N, total frequencies. Thus,

No. of Successes ***(X)***	**Probability** ***P(X)***	**Expected Frequencies** ***fe(x)***
0	$P(0) = e^{-m}.\frac{m^{0}}{0!} = e^{-m}$	$N.P(0) = N\,e^{-m}$
1	$P(1) = e^{-m}.\frac{m^{1}}{1!} = e^{-m}.m$	$N.P(1) = N\,e^{-m}.m$
2	$P(2) = e^{-m}.\frac{m^{2}}{2!} = e^{-m}.\frac{m^{2}}{2!}$	$N.P(2) = N\,e^{-m}.\frac{m^{2}}{2!}$
⋮	⋮	⋮
X	$P(x) = \frac{e^{-m}.m^{x}}{x!}$	$N.P(x) = \frac{N.e^{-m}.m^{x}}{x!}$

Alternative Method:

The expected frequencies can also be calculated in an easy way as follows:

(i) First, calculate the $fe\,(0) = N\,.P(0) = N\,.e^{-m}$

(ii) Other expected frequencies can be calculated as follows:

$fe\,(0) = N\,.P(0) = N\,.e^{-m}$

$fe\,(1) = \frac{m}{1}\,.\,fe(0)$

$fe\,(2) = \frac{m}{2}\,.\,fe(1)$

$fe\,(3) = \frac{m}{3}\,.\,fe(2)$

$fe\,(4) = \frac{m}{4}\,.\,fe(3)$ and so on.

Example 5.20: Fit a Poisson distribution to the following data and calculate the theoretical frequencies:

Deaths:	0	1	2	3	4

Frequency:	109	65	22	3	1

Also find men and variance of the above distribution. (Given $e^{-0.61} = 0.5432$)

Solution: **Fitting of Poisson Distribution**

Deaths (*X*)	**Frequency (*f*)**	***Fx***
0	109	0
1	65	65
2	22	44
3	3	9
4	1	4
	$\Sigma f = 200$	$\Sigma f X = 122$

$$\overline{X} = \frac{\Sigma fX}{\Sigma f} = \frac{122}{200} = 0.61$$

$\therefore$ $m = 0.61$

Now, you can obtain the value of $e^{-0.61}$ either from the table or by using the formula

e^{-m} = Rec. [Antilog (m x 0.4343)]

Putting $m = 0.61$

$e^{-0.61}$ = Rec. [Antilog (0.61 x 0.4343)]

= Rec. [Antilog (026492)]

= Rec. [1.841] = 0.5432

Now, $P(0) = e^{-0.61} \cdot \frac{(0.61)^0}{0!}$

$= e^{-0.61} = 0.5432$

Calculation of the Expected Frequencies:

$fe\ (0) = N\ .P(0) = 200 \text{ x } (0.5432) = 108.64 \approx 109$

$fe\ (1) = fe(0) \text{ x } \frac{m}{1} = 108.64 \text{ x } \frac{0.61}{1} = 66.27 \approx 66$

$fe\ (2) = fe(1) \text{ x } \frac{m}{2} = 66.27 \text{ x } \frac{0.61}{2} = 20.21 \approx 20$

$fe\ (3) = fe(2) \text{ x } \frac{m}{3} = 20.21 \text{ x } \frac{0.61}{3} = 4.11 \approx 4$

$fe\ (4) = fe(3) \text{ x } \frac{m}{4} = 4.11 \text{ x } = 0.63 \approx 1$

Thus,

***X*:**	0	1	2	3	4
***fe*:**	109	66	20	4	1

Mean = $\overline{X}$ = Variance = σ^2 = 0.61

Example 5.20: After correcting the profits of the first 50 pages of a book, it is found that an average, there are 3 errors per 5 pages. Use Poisson distribution to estimate the number of pages with 0, 1, 2, 3, … errors in the whole book of 1000 pages. (You are given: $e^{-0.6}$= 0.5488).

Solution: The average number of mistake = m = 3/5 = 0.6

where, (Given) $P(0) = e^{-m} \cdot \frac{m^0}{0!} = e^{-0.6} \cdot \frac{0.6^0}{0!} = e^{-0.6} = 0.5488$

$$P(0) = \frac{e^{-0.6}.(0.6)^1}{1!} = \frac{0.5488 \times 0.6}{1} = 0.32928$$

$$P(0) = \frac{e^{-0.6}.(0.6)^2}{2!} = \frac{0.5488 \times 0.36}{2 \times 1} = 0.98784$$

$$P(0) = \frac{e^{-0.6}.(0.6)^3}{3!} = \frac{0.5488 \times 0.216}{3 \times 2 \times 1} = 0.0197568$$

$P[X > 3]$ = 1 – [$P(0)$ + $P(1)$ + $P(2)$ + $P(3)$]

$P[X > 3]$ = 1 – [0.5488 + 0.32928 + 0.098784 + 0.197568]

= 1 – [0.9966208]

= 0.0033792

Fitting of Poisson distribution

X	***P*(*X*)**	***fe*(*X*) = *N* . *P*(*X*)**
0	0.5488	1000 x 0.5488 = 548.8 = 549
1	0.32928	1000 x 0.32928 = 329.28 = 329
2	0.098784	1000 x 0.098784 = 98.74 = 98
3	0.0197568	1000 x 0.0197568 = 19.7568 = 20
more than 3	0.0033792	1000 x 0.0033792 = 3.37 = 3
		N = 1000

5.13 SUMMARY

Theoretical frequency distribution refers to those distributions that are not obtained by actual observations or experiments but are mathematically deduced under certain assumptions. Theoretical frequency distributions are also called Probability Distribution or Expected Frequency Distribution. The main types of theoretical distributions are (i) Binomial Distributions, (ii)Poisson Distribution, and (iii) Normal Distribution. The binomial distribution is a discrete probability distribution and it was discovered by a Swiss Mathematician James Bernoulli. It is used in such situations where an experiment results in two possibilities – success and failure. Further, the Poisson distribution is a discrete probability distribution and it was developed by a French Mathematician Dr. Simon Denis Poisson. The Poisson distribution is used in those situations where the probability of the happening of an event is very small, *i.e.,* the event rarely occurs.

5.14 GLOSSARY

THEORETICAL FREQUENCY DISTRIBUTION These distributions are not obtained from actual data or experiments but can be obtained by mathematical methods based on certain assumptions.

5.15 CHECK YOUR PROGRESS

1……..are called frequency distributions which are obtained from real data or experiments.
2……….. are also called probability distribution or expected frequency distribution.
3.. Binomial distribution was discovered by a Swiss Mathematician……
4.The …… is used in those situations where the probability of the happening of an event is very small, *i.e.,* the event rarely occurs.

5.16 ANSWERS TO CHECK YOUR PROGRESS

1 Observed frequency distribution
2. Theoretical frequency distributions
3. James Bernoulli
4. Poisson distribution

5.17 TERMINAL QUESTIONS

1. A fair coin is tossed six times. What is the probability of obtaining four or more heads?

2. A die is thrown 4 times. Getting a number greater than 2 is a success. Find the probability of
   (i) exactly 1 success
   (ii) less than 3 successes
   (iii) more than 3 successes.

3. In a binomial distribution consisting of 5 independent trials, probabilities of 1 and 2 successes are 0.4096 and 0.2048. Find the parameter '*p*' of the distribution.

4. The chances of suffering from cold by workers working in an ice factory during winter are 20%. What is the probability that of 5 workers, 4 or more will suffer from cold?

5. An experiment succeeds twice as many times as it fails. Find the chances that in 6 trials, there will be at least 5 successes.

6. The mean and standard deviation of a binomial distribution are 2 and 1, respectively. Calculate *n, p* and *q.*

7. Find the probability of 3 successes in a binomial distribution whose mean and variance are 2 and 3/2, respectively.

8. A discrete random variable X has a mean equal to 6 and a variance equal to 2. If it is assumed that the underlying distribution X is binomial, what is the probability that $5 \leq X \leq 7$?

9. If the probability of a defective bolt is 10 %. Find *(i)* mean *(ii)* standard deviation *(iii)* moment coefficient of skewness and *(iv)* moment coefficient of kurtosis for the distribution of defective bolts in a total of 400.

10. An unbiased coin is tossed ten times, find the mean and the standard deviation.

11. Five perfect dice are thrown together for 96 times. The number 4, 5 or 6 was actually thrown in the experiments are given below:

No. of dice showing 4, 5 or 6 :	0	1	2	3	4	5
Frequency:	2	8	22	35	24	5

Fit a Binomial distribution and calculate the expected frequencies.

12. Four coins were tossed 200 times. The number of tosses showing 0, 1, 2, 3 and 4 heads were observed as under:

Number of heads :	0	1	2	3	4
Numbers of tosses :	15	35	90	40	20

Fit a Binomial distribution to these observed results.

13. The screws produced by a certain machine were checked by examining samples of 128. The following table shows the distribution of 128 samples according to the number of defective items they contained:

No. of defectives in a sample of 128:	0	1	2	3	4	5	6	7	Total
No. of samples:	7	6	19	35	30	23	7	1	128

Fit a binomial distribution and find the expected frequencies if the chance of machine being defective is 1/2. Find the mean and variance of the fitted distribution.

14. 5 coins are tossed 128 times. What is the probability of getting 3 or more heads and find out the expected frequencies of 3 or more heads?

15. It is given that 3% of the bulbs manufactured by a company are defective. Using the Poisson distribution, find the probability that a sample of 100 bulbs will contain *(i)* no defective *(ii)* exactly one defective.
(Given: $e^{-3} = 0.04979$)

16. Find the probability that at most 5 defective bolts will be found in a box of 200 bulbs if it is known that 2 per cent of such bolts are expected to be defective (You may take the distribution to be Poisson). (Take $e^{-4} = 0.0183$)

17. Assuming one in 80 births is a case of twins, calculate the probability of 2 or more sets of twins on a day when 30 births occur.

18. A manufacturer of pins knows that on an average 2% of his product is defective. He sells pins in a box of 200 and guarantees that not more than 3 pins will be defective. What is the probability that a box will fail to meet the guaranteed quality? (Given: $e^{-4} = 0.0183$)

19. One-fifth percent of the blades produced by a blade manufacturer turn out to be defective. The blades are supplied in packets of
Use Poisson distribution to calculate the approximate number to packets containing no defective, one defective, and two defective blades respectively in a consignment of 1,00,000 packets. (Given: $e^{-0.02} = 0.9802$)

20. Suppose a manufacturing product has 4 defects per unit of product inspected. Using Poisson distribution, calculate the probability of finding a product with 2 defects. (Given: $e^{-4} = 0.0183$)

21. The number of accidents is a year attributes to taxi drivers in a city follows Poisson distribution with mean 3. Out of 1,000 taxi drivers, find the number of drivers with *(i)* no accidents in a year, and *(ii)* more than 3 accidents in a year.
(Given: $e^{-1} = 0.3679$, $e^{-2} = 0.1353$, $e^{-3} = 0.0498$)

22. A television company estimated that average demand for engineers for repairing TV sets on each day is 1.5. Assuming this as a Poisson distribution it appoints two engineers. Calculate the proportion of days in a year in which both engineers are unemployed and the proportion of days in which the some demand for engineers is refused.
(Given: $e^{-1} = 0.3678$, $e^{-0.5} = 0.6065$)

23. A telephone exchange receives on the average 4 calls per minute. Find the probability on the basis of Poisson distribution if *(i)* 2 or less calls per minute *(ii)* upto 4 calls per minute and *(iii)* more than 4 calls per minute.
(Given: $e^{-4} = 0.0183$)

24. The following mistakes per page were observed in a book:

No. of mistakes per page:	0	1	2	3	4
No. of pages:	211	90	19	5	0

Fit a Poisson distribution for the data

25. One hundred car radios are inspected as they come off the production line and number of defects per radio set are recorded as follows:

No. of defects:	0	1	2	3
No. of Radio sets:	79	18	2	1

Estimate the average number of defects per radio and expected frequencies of 0, 1, 2, 3.

26. Fit a Poisson distribution of the following data and calculate the theoretical frequencies:

Deaths:	0	1	2	3	4
Frequency:	122	60	15	2	1

(Given: $e^{-0.5} = 0.60657$)

27. Below are given the number of vacancies of judges occurring in a High Court over a period of 96 years:

No. of vacancies:	0	1	2	3
Frequency:	59	27	9	1

Fit a Poisson distribution and calculate the mean and variance of the above distribution.

28. What do you understand by theoretical frequency distribution?

29. Explain the properties of Binomial and Normal Distributions.

30. What is Binomial distribution? Discuss the conditions for application of the Binomial distribution.

31. What is Poisson distribution? Explain the characteristics of Poisson distribution.

32. Discuss the important properties of Binomial and Poisson distribution.

33. What is Poisson distribution? Give examples where it can be applied.

ANSWERS TO CHECK YOUR PROGRESS

1. [11/32]
2. [*(i)* 0.0988, *(ii)* 0.4074, *(iii)* 0.1975]
3. [$p = 0.2$]
4. [0.0067]
5. [$\frac{256}{729}$]
6. [n=4, p=1/2, q=1/2]
7. [0.2076]

8. [0.712]
9. [$\overline{X}$ = 40, σ^2 =36, $\sqrt{\beta_1}$ = 0.133, β_2 = 3.013]
10. [$\overline{X}$ = 5, σ = $\sqrt{2.5}$]
11. [3, 15, 30, 30, 15, 3]
12. [12.5, 50, 75, 50, 12.5]
13. [: 1, 7, 21, 35, 21, 7, 1; $\overline{X}$ = 3.5, σ^2 = 1.75]
14. [: (i) 16/32, (ii) 64]
15. [(i) 0.05, (ii) 0.15]
16. [0.784]
17. [0.055]
18. [0.567]
19. [98020, 1960.4, 19.604]
20. [0.146624]
21. [*(i)* 50 *(ii)* 353]
22. [*(i)* 0.2231, *(ii)* 0.1913]
23. [*(i)* 0.2379, *(ii)* 0.6283, *(iii)* 0.3717]
24. [209.40, 92.14, 20.27, 2.97, 0.33]
25. [*m*=0.25, 77.88, 19.47, 2.43, 0.21]
26. [121.3, 60.65, 15.16, 2.53, 0.32]
27. [58.22, 29.11, 7.278, 1.21, $\overline{X}$ = σ^2 = 0.5]

5.18 SUGGESTED READINGS

1. Roy Ramendu, 'Principles of Statistics' Prayag Pustak Bhawan, Allahabad.
2. Gupta S. P. & Gupta M. P., 'Business Statistics' Sultan Chand & Sons, New Delhi.
3. Shukla S. M. & Sahai S. P., 'Advanced Statistics' Sahitya Bhawan Publications, Agra.
4. Goon, Gupta and Dasgupta, 'Basic Statistics' World Press Limited – Calcutta.
5. Fundamentals of Business Statistics – Sanchethi and Kappor.
6. Srivastava, Shenoy and Guptha, 'Quantitative Methods in Management'.

PAPER CODE: MCM-02

BLOCK: 2

UNIT6: EXPONENTIAL, BETA AND NORMAL DISTRIBUTIONS

Structure

OBJECTIVES

After studying this unit, you will be able to understand:

i. Exponential Distribution;

ii. Beta Distribution;

iii. Normal Probability Distribution, and its applications.

6.1 INTRODUCTION

In probability theory and statistics, the exponential, beta and normal distributions are a family of continuous probability distributions. Normal distribution is one of the most important and widely used continuous probability distribution. It is mainly used to study the behaviour of continuous random variables like height, weight and intelligence of a group of students.

6.2 EXPONENTIAL DISTRIBUTION

The exponential distribution describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate. It is the continuous analogue of the geometric distribution.

6.2.1 Characteristics of Exponential Distribution

The exponential probability distribution has the following characteristics:

(a) Probability density function

The probability density function (PDF) of an exponential distribution is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Alternatively, this can be defined using the Heaviside step function, H(x).

$$f(x;\lambda) = \lambda e^{-\lambda x} H(x)$$

Here $\lambda > 0$ is the parameter of the distribution, often called the rate parameter. The distribution is supported on the interval $[0, \infty)$. If a random variable X has this distribution, we write $X \sim \text{Exp}(\lambda)$.

The exponential distribution exhibits infinite divisibility.

(b) Cumulative distribution function

The Cumulative distribution function is given by

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Alternatively, this can be defined using the Heaviside step function, $H(x)$.

$$F(x;\lambda) = (1 - e^{-\lambda x})H(x)$$

(c) Alternative parameterization

A commonly used alternative parameterization is to define the probability density function of an exponential distribution as

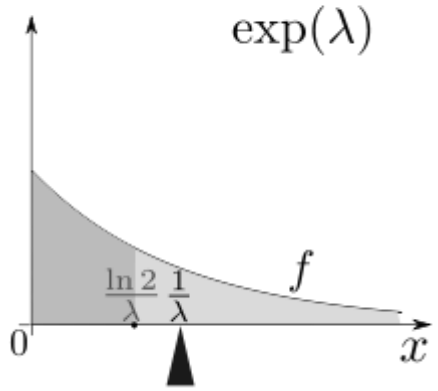
$$f(x;\beta) = \begin{cases} \frac{1}{\beta}e^{-x/\beta}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Where $\beta > 0$ is a scale parameter of the distribution and is the reciprocal of the rate parameter, λ, defined above. In this specification, β is a survival parameter in the sense that if a random variable X is the duration of time that a given biological or mechanical system manages to survive and $X \sim$ Exponential(β), then E[X] = β. That is to say, the expected duration of survival of the system is β units of time. The parameterization involving the "rate" parameter arises in the context of events arriving at a rate λ, when the time between events (which might be modelled using an exponential distribution) has a mean of $\beta = \lambda^{-1}$.

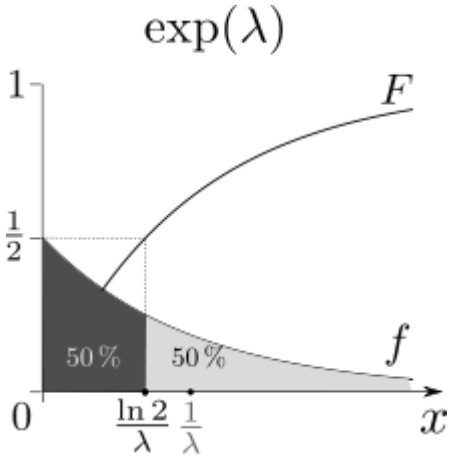
6.2.2 Properties of exponential distribution

The properties of the exponential distribution are as under:

(a) Mean, variance, moments and median



The mean is the probability mass centre, which is the first moment.



The median is the preimage $F^{-1}(\frac{1}{2})$.

The mean or expected value of an exponentially distributed random variable *X* with rate parameter λ is given by

$$\mathrm{E}[X] = \frac{1}{\lambda}.$$

In light of the examples given above, this makes sense: if you receive phone calls at an average rate of 2 per hour, then you can expect to wait half an hour for every call.

The variance of *X* is given by

$$\operatorname{Var}[X] = \frac{1}{\lambda^2}.$$

The moments of *X*, for n=1,2,..., are given by

$$\mathrm{E}[X^n] = \frac{n!}{\lambda^n}.$$

The median of *X* is given by

$$\mathrm{m}[X] = \frac{\ln 2}{\lambda} < \mathrm{E}[X],$$

Where, *ln* refers to the natural logarithm. Thus the absolute difference between the mean and median is

$$|\mathrm{E}[X] - \mathrm{m}[X]| = \frac{1 - \ln 2}{\lambda} < \frac{1}{\lambda} = \text{standard deviation},$$

in accordance with the median-mean inequality.

(b) Memory lessness

An important property of the exponential distribution is that it is memoryless. This means that if a random variable *T* is exponentially distributed, its conditional probability obeys

$$\Pr(T > s + t \mid T > s) = \Pr(T > t) \text{ for all } s, t \geq 0.$$

This says that the conditional probability that you have to wait, for example, more than another 10 seconds before the first arrival, given that the first arrival has not yet happened after 30 seconds, is equal to the initial probability that you have to wait more than 10 seconds for the first arrival. So, if you wait for 30 seconds and the first arrival didn't happen ($T > 30$), probability that you will have to wait another 10 seconds for the first arrival ($T > 30 + 10$) is the same as the initial probability that we need to wait more than 10 seconds for the first arrival ($T > 10$). The fact that $\Pr(T > 40 \mid T > 30) = \Pr(T > 10)$ does *not* mean that the events $T > 40$ and $T > 30$ are independent.

To summarise: "memory lessness" of the probability distribution of the waiting time *T* until the first arrival means

$$(\text{Right})\ \ \Pr(T > 40 \mid T > 30) = \Pr(T > 10).$$

It does *not* mean

$$(\text{Wrong})\ \Pr(T > 40 \mid T > 30) = \Pr(T > 40).$$

(That would be independence. These two events are *not* independent.)

The exponential distributions and the geometric distributions are the only memoryless probability distributions.

The exponential distribution is consequently also necessarily the only continuous probability distribution that has a constant Failure rate.

(c) Quartiles

The quartile function (inverse cumulative distribution function) for Exponential (λ) is

$$F^{-1}(p;\lambda) = \frac{-\ln(1-p)}{\lambda}, \qquad 0 \le p < 1$$

The quartiles are therefore: first quartile *ln*(4/3)/λ

median *ln*(2)/λ

third quartile ln(4)/λ

6.3 BETA DISTRIBUTION

The beta distribution is a continuous probability distribution which is defined on the interval (0, 1) parameterized by two positive shape parameters, typically denoted by α and β. The beta distribution can be suited to the statistical modelling of proportions in applications where values of proportions equal to 0 or 1 do not occur. One theoretical case where the beta distribution arises is as the distribution of the ratio formed by one random variable having a Gamma distribution divided by the sum of it and another independent random variable also having a Gamma distribution with the same scale parameter (but possibly different shape parameter).

The usual formulation of the beta distribution is also known as the beta distribution of the first kind, whereas beta distribution of the second kind is an alternative name for the beta prime distribution.

6.3.1 Characterization

The characteristics of the beta distribution are as follows:

(a) Probability density function

The probability density function of the beta distribution is:

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du}$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1}$$

$$= \frac{1}{\mathrm{B}(\alpha,\beta)}\,x^{\alpha-1}(1-x)^{\beta-1}$$

Where $\Gamma(z)$is the gamma function. The beta function, *B*, appears as a normalization constant to ensure that the total probability integrates to unity.

A random variable *X* that is Beta-distributed with shape *α* and *β* is denoted

$$X \sim \mathrm{Be}(\alpha,\beta)$$

(b) Cumulative distribution function

The cumulative distribution function is

$$F(x;\alpha,\beta) = \frac{\mathrm{B}_x(\alpha,\beta)}{\mathrm{B}(\alpha,\beta)} = I_x(\alpha,\beta)$$

where $\mathrm{B}_x(\alpha,\beta)$is the incomplete beta function and $I_x(\alpha,\beta)$is the regularized incomplete beta function.

6.3.2 Properties

The properties of the beta distribution are as under:

The mode of a Beta distributed random variable *X* with parameters α > 1 and β > 1 is:

$$\mathrm{mode} = \frac{\alpha-1}{\alpha+\beta-2}.$$

The expected value (mean) (μ) and variance (second central moment) of a Beta distribution random variable *X* with parameters α and β are:

$$\begin{aligned} \mu = \mathrm{E}(X) &= \frac{\alpha}{\alpha+\beta}, \\ \operatorname{var}(X) = \mathrm{E}(X-\mu)^2 &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \end{aligned}$$

In all cases $\operatorname{var}(X) < \frac{1}{4}$.

The skewness is

$$\frac{\mathrm{E}(X-\mu)^3}{[\mathrm{E}(X-\mu)^2]^{3/2}} = \frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}}.$$

The kurtosis is

$$\frac{\mathrm{E}(X-\mu)^4}{[\mathrm{E}(X-\mu)^2]^2} - 3 = \frac{6[\alpha^3 - \alpha^2(2\beta-1) + \beta^2(\beta+1) - 2\alpha\beta(\beta+2)]}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)},$$

or:

$$\frac{6[(\alpha-\beta)^2(\alpha+\beta+1) - \alpha\beta(\alpha+\beta+2)]}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)}.$$

In general, the kth raw moment is given by

$$\mathrm{E}(X^k) = \frac{\mathrm{B}(\alpha+k,\beta)}{\mathrm{B}(\alpha,\beta)} = \frac{(\alpha)_k}{(\alpha+\beta)_k}$$

where $(x)_k$ is a Pochhammer symbol representing the rising factorial. It can also be written in a recursive form as

$$\mathrm{E}(X^k) = \frac{\alpha+k-1}{\alpha+\beta+k-1}\,\mathrm{E}(X^{k-1}).$$

One can also show that

$$\mathrm{E}(\log X) = \psi(\alpha) - \psi(\alpha+\beta),$$

and

$$\mathrm{E}(X^{-1}) = \frac{\alpha+\beta-1}{\alpha-1}.$$

(c) Quantities of information

Given two beta distributed random variables, $X \sim \text{Beta}(\alpha, \beta)$ and $Y \sim \text{Beta}(\alpha', \beta')$, the differential entropy of X is

$$h(X) = \ln \mathrm{B}(\alpha,\beta) - (\alpha-1)\psi(\alpha) - (\beta-1)\psi(\beta) + (\alpha+\beta-2)\psi(\alpha+\beta),$$

Where $\psi$is the digamma function?

The cross-entropy is

$$H(X,Y) = \ln \mathrm{B}(\alpha',\beta') - (\alpha'-1)\psi(\alpha) - (\beta'-1)\psi(\beta) + (\alpha'+\beta'-2)\psi(\alpha+\beta).$$

It follows that the Kullback-Leibler divergence between these two beta distributions is

$$D_{\mathrm{KL}}(X,Y) = \ln \frac{\mathrm{B}(\alpha',\beta')}{\mathrm{B}(\alpha,\beta)} - (\alpha'-\alpha)\psi(\alpha) - (\beta'-\beta)\psi(\beta) + (\alpha'-\alpha+\beta'-\beta)\psi(\alpha+\beta).$$

6.4 NORMAL PROBABILITY DISTRIBUTION

Normal distribution was first discovered by the English Mathematician Abraham De Moivre in 1733. But it was later rediscovered and applied by Laplace and Karl Gauss. Normal distribution is also known as the Gaussian distribution after the name of Karl Gauss.

Moreover, normal distribution may be looked upon as the limiting form of binomial distribution under certain conditions:

(1) *n*, the number of trials is infinitely large, *i.e.*, $n \to \infty$

(2) Neither *p* nor *q* is very small.

Normal distribution is defined and given by the following probability function:

$$P\,(X = x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{X-\overline{X}}{\sigma}\right)^2} \quad -\infty < X < +\infty$$

Where $\overline{X}$ = Mean, σ = Standard deviation, *e* (base of natural logarithm) = 2.7183 and π= 3.1415

Normal distribution in its standard normal variate (SNV) form is given by:

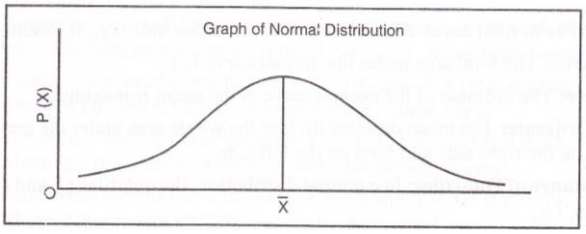
$$P\,(Z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}Z^2} \quad -\infty < Z < \infty \text{ where } Z = \frac{X - \overline{X}}{\sigma}$$

The mean of Z is zero, and the standard deviation of σ is 1.

In total, the Standard Normal Distribution is that normal distribution whose mean is zero and variance is unity.

6.4.1 Graph of Normal Distribution

The graph of the normal distribution is called the normal curve. The normal curve is the graphic presentation of normal distribution. The following figure illustrates the normal curve:



The shape of the normal curve depends on the values of the mean ($\overline{X}$) and standard deviation (σ). There will be different shapes of normal curves for different values of mean and standard deviation.

6.4.2 Assumption of Normal Distribution

The normal distribution is based on the following assumptions:

(i) **Independent Causes:** The forces affecting the event must be independent from each other.

(ii) **Condition of Symmetry:** The operation of causal forces must be such that the deviation from the mean on either side is equal in number and size.

(iii) **Multiple Causation:** The causal forces must be numerous and of approximately equal weight or importance.

6.4.3 Characteristics of Normal Distribution/Normal Curve

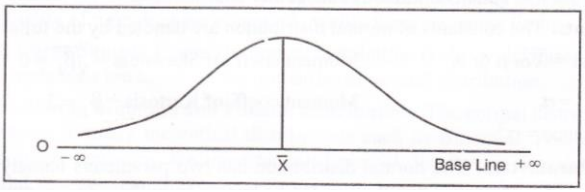
(i) **Perfectly Symmetrical and Bell-Shaped:** The normal curve is perfectly symmetrical and bell-shaped about the mean. This means that if we fold the curve along its vertical axis at the centre, the two halves would coincide or overlap completely.

(ii) **Unimodal Distribution:** It has only one mode, *i.e.*, it is a unimodal distribution.

(iii) **Equality of Mean, Median and Mode:** In a normal distribution, mean, median and mode are equal, *i.e.*,

$$\overline{X} = M = Z$$

(iv) **Asymptotic to the Base Line:** Normal curve is asymptotic to the base line on either sides, *i.e.*, it has a tendency to touch the base line but it never touches it. This is clear from the following figure:



(v) **Range:** The normal curve extends to infinity on either side, *i.e.*, it extends $-\infty$ to $+\infty$.

(vi) **Total Area:** The total area under the normal curve is 1.

(vii) **Ordinate:** The ordinate of the normal curve at the mean is maximum.

(viii) **Mean Ordinate:** The mean ordinate divides the whole area under the curve into two equal parts, *i.e.*, 50% on the right side and 50% on the left side.

(ix) **Equidistance of Quartiles:** In a normal distribution, the quartiles Q_1 and Q_3 are equidistant from the median, *i.e.*,

$$Q_1 - M = M - Q_1$$

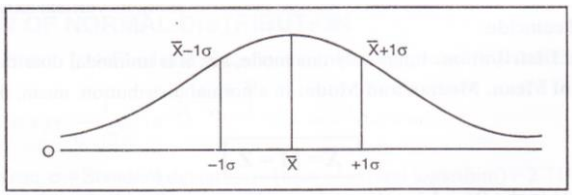
(x) Quartile Deviation: In a normal distribution, the quartile deviation is 2/3 times from the standard deviation, *i.e.,*

Q.D. =2/3 S.D.

(xi) Mean Deviation: In a normal distribution, the mean deviation is 4/5 times the standard deviation, *i.e.,*

M.D. =4/5 S.D.

(xii) Points of Inflexion: The normal curve has two points of inflexion (*i.e.,* the points where the curve changes its curvature) at $\overline{X} - 1\sigma$ and $\overline{X} + 1\sigma$. In other words, the points of inflexion occur at $\overline{X} \pm 1\sigma$, *i.e.,* at $\overline{X} - 1\sigma$ and $\overline{X} + 1\sigma$. This is clear from the figure given below:



(xiii) Continuous Probability Distribution: Normal distribution is a distribution of continuous variables. Therefore, it is called a continuous Probability Distribution.

(xiv) Constants: The constants of normal distribution are denoted by the following symbols:

Mean = $\overline{X}$ or μ or *m* Moment coeff. of Skewness = $\sqrt{\beta_1} = 0$

S.D. = σ Moment coeff. of Kurtosis = $\sqrt{\beta_2} = 3$

Variance = σ^2

(xv) Main Parameters: The normal distribution has two parameters, namely mean ($\overline{X}$) and standard deviation (σ). The entire distribution can be known from these two parameters.

(xvi) Areas Property: One of the most important properties of a normal curve is the area relationship property. The total area under the normal curve is 1. It has been found that:

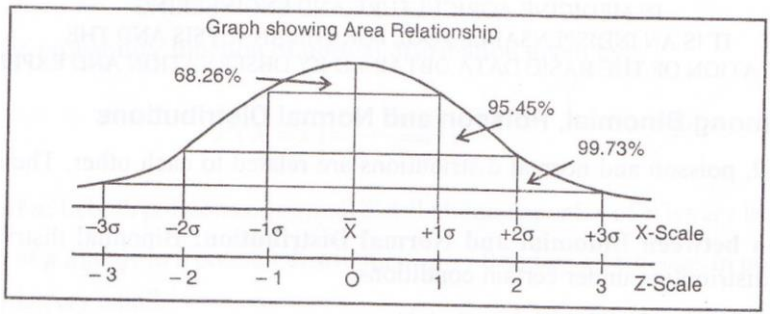
a. Area under the normal curve between $\overline{X} - 1\sigma$ and $\overline{X} + 1\sigma$ is 0.6826,

i.e., Mean $\pm 1\sigma$ covers 68.26% area under the normal curve.

b. Area under the normal curve between $\overline{X} - 2\sigma$ and $\overline{X} + 2\sigma$ is 0.9545, *i.e.*, Mean $\pm 2\sigma$ covers 95.45% area under the normal curve.

c. Area under the normal curve between $\overline{X} - 3\sigma$ and $\overline{X} + 3\sigma$ is 0.9973, *i.e.*, Mean $\pm 3\sigma$ covers 99.73% area under the normal curve.

The following figure illustrates the area property:



6.4.4 Importance of Normal Distribution

The normal distribution has great significance in statistical analysis. It is the basis of modern statistics. The following points highlight the importance and uses of normal distribution:

(i) **Study of Natural Phenomenon:** All-natural phenomenon possesses the characteristics of normal distribution, such as the length of leaves of a tree, the heights of adults, birth rates and death rates, etc. The normal distribution is widely used in the study of natural phenomenon.

(ii) **Basis of Sampling Theory:** The normal distribution is also of great importance in sampling theory. The normal distribution is the basis of sampling theory. With the help of normal distribution, one can test whether the samples drawn from the universe represent the universe satisfactorily or not.

(iii) **Statistical Quality Control:** It helps determine the tolerance or specification limits within which the product's quality lies. The variations in the quality of a product are acceptable within these tolerance limits.

(iv) **Useful for Large Sample Tests:** The normal distribution is also used in cases of large samples because large sample tests are based on the properties of the normal distribution.

(v) **Approximation to Binomial and Poisson distribution:** The normal distribution serves as a good approximation to many theoretical distributions such as Binomial, Poisson, etc. As the number of observations increases, the importance of normal distribution to solve the problems relating to Binomial, Poisson, etc., increases.

(vi) **Prof. Youden** has expressed the importance of normal distribution in the shape of a normal curve, which is shown below:

THE

NORMAL

LAW OF ERROR

STANDS OUT IN THE

EXPERIENCE OF MANKIND

AS ONE OF THE BROADEST

GENERALIZATIONS OF NATURAL

PHILOSOPHY. IT SERVES AS THE

GUIDING INSTRUMENT IN RESEARCH

IN THE PHYSICAL AND SOCIAL SCIENCES AND

IN MEDICINE, AGRICULTURAL AND ENGINEERING

IT IS AN INDISPENSABLE TOOL FOR THE ANALYSIS AND THE

INTERPRETATION OF THE BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT.

6.4.5 Relationship among Binomial (BD), Poisson (PD) and Normal Distributions (ND)

The binomial, Poisson and normal distribution are related to each other. The relationship is shown below:

(A) **Relation between Binomial and Normal Distribution:** Binomial distribution tends to become a normal distribution under certain conditions:

(i) n approaches to infinity, *i.e.*, $n \to \infty$

(ii) Neither p nor q is very small.

(B) **Relation between Poisson and Normal Distributions:** Poisson distribution tends to become normal distribution if its parameter 'm' becomes very large, *i.e.*, if $m \to \infty$, than PD tends to ND.

6.4.6 Difference between Normal and Binomial Distributions

The following are the main differences between normal and binomial distributions:

(i) **Nature:** Binomial distribution is a discrete probability distribution whereas normal distribution is a continuous probability distribution.

(ii) **Probability Functions:** The probability function of the binomial distribution is given by:

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

The probability function of the normal distribution is given by:

$$P\,(X = x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{X-\overline{X}}{\sigma}\right)^2}$$

(iii) **Value of *n*:** In a binomial distribution, the value of *n, i.e.*, the number of trials, is finite, whereas in a normal distribution, *n* approaches infinity, *i.e*, $n \to \infty$

(iv) **Parameters:** The binomial distribution has two parameters, *n* and *p*, whereas the normal distribution also has two parameters, namely $\overline{X}$ and σ.

(v) **Shape:** The binomial distribution can be symmetrical and asymmetrical. It depends on the values of *p* and *q*. On the other hand, the normal distribution is always perfectly symmetrical.

6.5 MEASURING AREA UNDER THE NORMAL CURVE

The following steps are to be followed to measure the area under the normal curve:

(1) Firstly, the given value of the normal variate is transformed into standard units by substitution. The formula of Z-transformation is given by:

$$Z = \frac{X - \overline{X}}{\sigma}$$

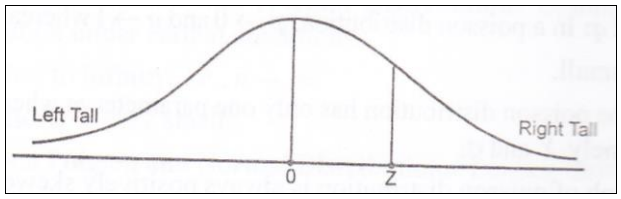
For example, if $\overline{X}$ = 30, σ = 5 and X = 35, then the standard normal variate corresponding to 35 will be:

$$Z = \frac{35 - 30}{5} = 1$$

Thus, the Z-transformation of 35 will be 1.

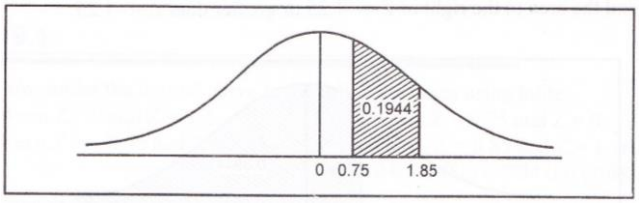
(2) Then the area is obtained from the area tables (given at the end of the Unit) for any particular value of Z.

The table given at the end of the Unit shows the area between O to Z, which is shown in the figure given below:



Example 6.1: Find the area under the normal curve between Z = 0.75 and Z = 1.85.

Solution:



Required Area

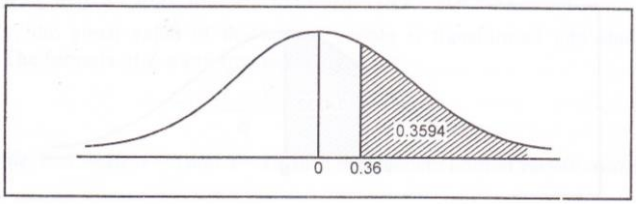
= (Area between Z = 0 and Z = 1.85) – (Area between Z = 0 and Z = 0.75)

= 0.4678 – 0.2734

= 0.1944

Example 6.2: Find the area to the right of Z = +0.36.

Solution:



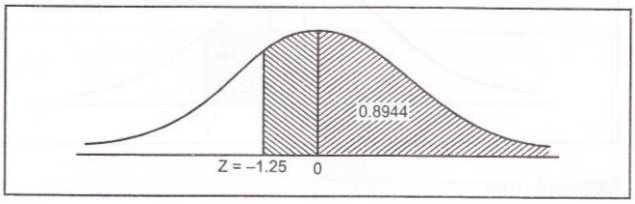
Required Area

= (Area between Z = 0) – (Area between Z = 0 and Z = 0.36)

= 0.5000 – 0.1406 = 0.3594

Example 6.3: Find the area to the right of Z = -1.25 or greater than Z = -1.25.

Solution:



Required Area

= (Area between Z = – 1 .25 and Z = 0) + (Area to the right of Z = 0)

= 0.3944 + 0.5000

= 0.8944

6.6 APPLICATIONS OF NORMAL DISTRIBUTION

Now, you will study the applications of normal distribution:

6.6.1 Finding areas when $\overline{X}$ and σ of normal variate are given

To find out the area under the normal curve, first you have to transform the given value of the normal variate into the Z-variate. For example, if $\overline{X}$ = 30, σ = 5 and X = 35, will be transformed into the standard normal variate as follows:

$$Z = \frac{35-30}{5} = 1 \qquad \text{where } Z = \frac{X - \overline{X}}{\sigma}$$

Thus, for X = 35, the standard normal variate (SNV) is 1.

After Z-transformation, the table of area under the normal curve is considered.

Example 6.4: An aptitude test for selecting officers was conducted on 1,000 candidates, and it was

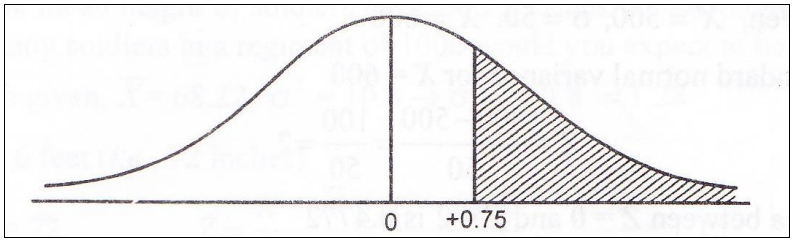
found that the average score is 42 and the standard deviation of scores is 24. Assume normal distribution for the scores, find (i) the number of candidates whose score exceeds 60 and (ii) the number of candidates whose score lies between 30 and 66.

Solution: Given, $\overline{X}$ = 42, σ = 24, N = 1000

(i) Exceeding 60

Z = SNV corresponding to 60

$$= \frac{X - \overline{X}}{\sigma} = \frac{60-42}{24} = \frac{18}{24} = \frac{3}{4} = +0.75$$



Required Proportion

= Area to the right of (Z = 0) – Area between (Z = 0 and Z = +0.75)

= 0.5000 – 0.2734 = 0.2266

Number of candidates whose score exceeds 60

= 1,000 x 0.2266 = 226.6 or 227

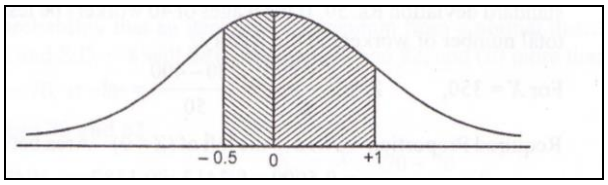
(ii) Between 30 and 66

Z_1= SNV corresponding to 30

$$= \frac{X_1 - \overline{X}}{\sigma} = \frac{30-42}{24} = \frac{-12}{24} = -0.5$$

Z_2= SNV corresponding to 30

$$= \frac{X_2 - \overline{X}}{\sigma} = \frac{66-42}{24} = \frac{24}{24} = +1$$



Required Proportion

= Area between (Z = -0.5 and Z=0) – Area between (Z=0 and Z = +1)

= 0.1915 + 0.3413= 0.5328

Number of candidates whose score lie between 30 and 66

= 1,000 x 0.5328 = 532.8 or 533

Example 6.5: Assume the mean height of soldiers to be 68.22 inches with a variance of 10.8 inches,

How many soldiers in a regiment of 1000 would you expect to be over six feet tall?

Solution: Given, $\overline{X}$ = 68.22, σ^2 = 10.8 or σ= $\sqrt{10.8}$ = 3.28

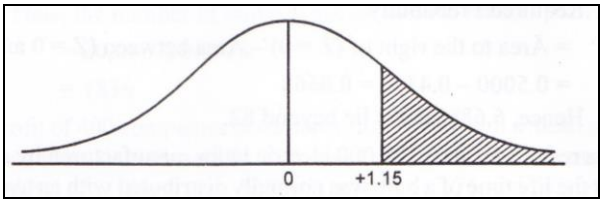
Above 6 feet (*i.e.,* 72 inches)

For $X = 72$, $$Z = \frac{X - \overline{X}}{\sigma} = \frac{72 - 68.22}{3.28} = \frac{3.78}{3.28} = 1.15$$

Required Proportion

= (Area to the right of Z=0) – (Area between Z=0 and Z=1.15)

= 0.5000 – 0.3749 = 0.1251



Thus, the expected number of soldiers having a height above 6 feet

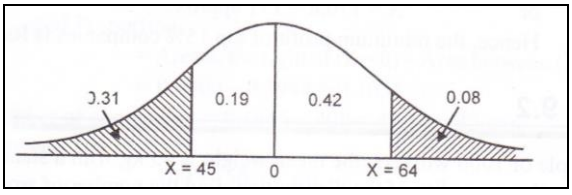
= 1000 x 0.1251= 125.1 or 125

6.6.2 Finding means and standard deviation when the area under the normal curve is given

When the area under the normal curve is given, then you can find the mean ($\overline{X}$) and standard deviation (σ) of the normal distribution.

Example 6.6: In a normal distribution, 31% of the items are under 45 and 8% are over 64. Find $\overline{X}$ and σ of the distribution.

Solution:



$$Z = \frac{X - \overline{X}}{\sigma}$$

Value of Z corresponding to 0.50 – 0.3 = 0.19 area = – 0.5 (From the table)

$$\therefore \quad -0.5 = \frac{45 - \overline{X}}{\sigma}$$

or $$-0.5\ \sigma = 45 - \overline{X}$$

or $$\overline{X} - 0.5\ \sigma = 45 \qquad ...(i)$$

Value of Z corresponding to 0.5 – 0.08 = 0.42 area = + 1.41 (From the table)

$$\therefore \quad 1.41 = \frac{64 - \overline{X}}{\sigma}$$

or $$1.41\sigma = 64 - \overline{X}$$

or $$\overline{X} + 1.41\sigma = 64 \qquad ...(ii)$$

Solving the two equations

$$\overline{X} - 0.5\ \sigma = 45$$

$$\overline{X} + 1.41\sigma = 64$$

– – –

$$-1.91\ \sigma = -19$$

$$\Rightarrow \sigma = 10$$

Substituting the values of σ in equations *(i)*

$$\overline{X} - 0.5\ (10) = 45$$

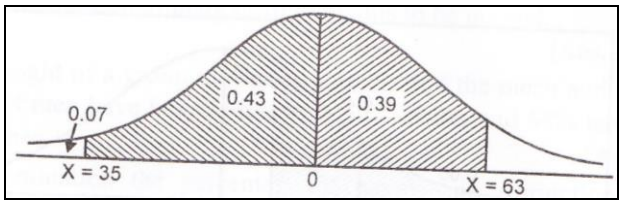
$$\overline{X} - 5 = 45$$

Or $$\overline{X} = 50$$

$$\therefore \quad \overline{X} = 50,\ \sigma = 10$$

Example 6.7: In a distribution exactly normal, 7 % of the items are under 35, and 89% are under 63. Find the mean and standard deviation of the distribution:

Solution: Value of Z corresponding to 0.43 area = - 1.48



$\therefore \quad -1.48 = \dfrac{35 - \overline{X}}{\sigma}$

or $\quad -1.48\ \sigma = 35 - \overline{X}$

or $\quad \overline{X} - 1.48\sigma = 35 \quad ...(i)$

Value of Z corresponding to 0.39 area = +1.23

$\therefore \quad 1.23 = \dfrac{63 - \overline{X}}{\sigma}$

or $\quad 1.23\sigma = 63 - \overline{X}$

or $\quad \overline{X} + 1.23\sigma = 63 \quad ...(ii)$

Solving the two equations

$\overline{X} - 1.48\sigma = 35$

$\overline{X} + 1.23\sigma = 63$

– – –

$-2.71\ \sigma = -28$

Or $\quad 2.71\ \sigma = 28$

or $\quad \sigma = \dfrac{28}{2.71}$

$= 10.33$

Substituting the values of σ in *(i)*

$\overline{X} - 1.48\ (10.33) = 35$

$$\overline{X} - 15.3 = 35$$

$$\Rightarrow \quad \overline{X} = 50.3$$

Thus, $\overline{X} = 50.3$, $\sigma = 10.33$

6.6.3 Finding the minimum and maximum score amongst the highest and lowest group

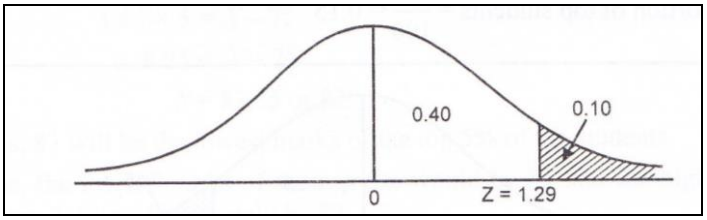
When the $\overline{X}$, σ and the proportion of the highest and lowest groups is given, then you can find the minimum and maximum score amongst the highest and lowest groups.

Example 6.8: The wages of 5,000 workers were found to be normally distributed with a mean Rs. 2,000 and a standard deviation of Rs. 120. What mark was the lowest wage amongst the richest 500 workers?

Solution: Given, N = 5,000, $\overline{X} = 2000$, $\sigma = 120$

Proportion of richest workers = $\frac{500}{5000} = \frac{1}{10} = 0.10$

The value of Z corresponding to 0.40 area = 1.29



We know that: $Z = \frac{X - \overline{X}}{\sigma}$

$$1.29 = \frac{X - 2000}{120}$$

$$\Rightarrow \quad X - 2000 = 154.8$$

$$\Rightarrow \quad X = \text{Rs. } 2154.8$$

Thus, the lowest wages among the richest 500 workers are Rs. 2154.8.

Example 6.9: A set of examination marks is approximately normally distributed with a mean of 75 and a standard deviation of 5. If the top 5% of the students get grade A and the bottom 25% get grade F, what is the lowest A, and what mark is the highest F?

Solution: Given, $\overline{X} = 75$, $\sigma = 5$

$$Z = \frac{X - \overline{X}}{\sigma}$$

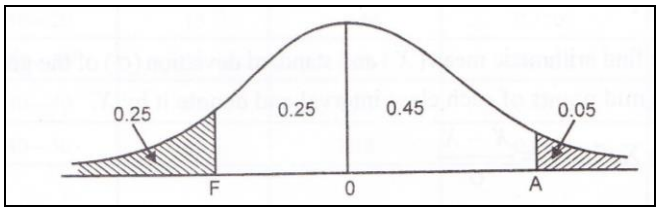
Value of Z corresponding to 0.5 – 0.25 = 0.25 area = – 0.68 (From the table)

$$\therefore \quad -0.68 = \frac{X - 75}{5}$$

or $\quad -0.68 \times 5 = X - 75$

or $\quad -3.4 = X - 75$

or $\quad X = 71.6$ or 72.



Thus, 72 will be the highest marks of the bottom 25% of the students.

Value of z corresponding to (0.50 – 0.05) = 0.45 area = 1.65 (From the table)

$$\therefore \quad 1.65 = \frac{X - 75}{5}$$

or $\quad 1.65 \times 5 = X - 75$

or $\quad 8.25 = X - 75$

or $\quad X = 83.25$ or 83.

Thus, 83 will be the lowest mark of the top 5% of the students.

Thus, the lowest marks of the top 5% would be 83, and the highest marks of the bottom 25% of students would be 72.

6.7 FITTING OF NORMAL CURVE

There are two methods for fitting the normal curve:

6.7.1 Ordinate Method

This method uses the Table of Ordinates of the Standard Normal Curve. This method involved the following steps:

(i) First, find the arithmetic mean ($\overline{X}$) and standard deviation (σ) of the given distribution.

(ii) Find the midpoints of each class interval and denote them by *X*.

(iii) For each X, find Z = $\frac{X - \overline{X}}{\sigma}$

(iv) Find ordinates at each of these values of Z from the table of ordinates.

(v) Multiply each of these values by $N \times \frac{i}{\sigma}$ and find the expected frequencies.

Here, N = number of items, *i* = size of class interval, σ = S.D.

Example 6.10: Fit a normal curve to the following data by the method of ordinance:

Variable:	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
Frequency:	3	5	8	3	1

Solution: For fitting the normal curve, compute $\overline{X}$ and σ

Computation of $\overline{X}$ and σ

Variable	***F***	**M.V. (*X*)**	***d***	***d′=d / i***	***fd′***	***fd′ 2***
0 – 10	3	5	-20	-2	-6	12
10 – 20	5	15	-10	-1	-5	5
20 – 30	8	25	0	0	0	0
30 – 40	3	35	+10	+1	+3	3
40 – 50	1	45	+20	+2	+2	4
	N=20				$\sum fd' = -6$	$\sum fd'^2 = 24$

$$\overline{X} = A + \frac{\sum fd'}{N} \times i = 25 + \frac{-6}{20} \times 10 = 22$$

$$\sigma = \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times i$$

$$= \sqrt{\frac{24}{20} - \left(\frac{-6}{20}\right)^2} \times 10 = 10.53$$

After finding the $\overline{X}$ and σ, adopt the following procedure.

Variable (1)	F (2)	M.V. (X) (3)	$Z = \frac{X - \overline{X}}{\sigma}$ (4)	Value of Ordinate from Ordinate Table (5)	$fe = \frac{Ordinate \times N \times i}{\sigma}$ (6)
0 – 10	3	5	-1.61	0.1092	2.07 = 2
10 – 20	5	15	-0.66	0.3209	6.09 = 6
20 – 30	8	25	0.28	0.3836	7.28 = 7

30 – 40	3	35	1.23	0.1872	3.55 = 4
40 – 50	1	45	2.18	0.0371	0.7046 = 1
					N=20

6.7.2 Area Method

This method uses the table of area under the Standard Normal Curve. It involves the following steps:

(i) First, find $\overline{X}$ and σ of the given distribution.

(ii) Write the lower limit for each class interval and denote it by '*X*'.

(iii) For each lower-class limit *X*, find Z = $\frac{X - \overline{X}}{\sigma}$

(iv) Find the area at each of these values of Z from the area table.

(v) Then, the successive differences between two area values are computed. These are a same sign and adding them when the Z's have opposite sign.

(vi) Multiply each of these values by N to find the expected frequencies.

Example 6.11: Fit a normal curve to the following data:

Variable:	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
Frequency:	3	5	8	3	1

Solution: From the above example, find that $\overline{X}$ = 22, σ = 10.53, N = 20,

After finding the $\overline{X}$ and σ We adopt the following procedure:

Variable	Lower Class Limit (*X*)	Z = $\frac{X - \overline{X}}{\sigma}$ (3)	Area from 0 to Z (4)	Area of each Class Interval (5)	*fe* = N x Area

(1)	(2)				(6)
0 – 10	0	-2.09	0.4817	0.1088	2.17 ≈ 2
10 – 20	10	-1.14	0.3729	0.2975	5.95 ≈ 6
20 – 30	20	-0.19	0.0753	0.3518	7.036 ≈ 7
30 – 40	30	0.76	0.2764	0.1800	3.6 ≈ 4
40 – 50	40	1.71	0.4564	0.0397	0.794 ≈ 1
50 – 60	50	2.66	0.4961		N = 20

6.8 SUMMARY

The exponential distribution describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate. It is the continuous analogue of the geometric distribution. The beta distribution is also a continuous probability distribution which is defined on the interval (0, 1) parameterized by two positive shape parameters, typically denoted by α and β. The beta distribution can be used for statistical modelling of proportions in applications where values of proportions equal to 0 or 1 do not occur. Normal distribution is one of the most important and widely used continuous probability distribution. It is mainly used to study the behaviour of continuous random variables like height, weight and intelligence of a group of students.

6.9 GLOSSARY

Exponential Distribution- A process in which events occur continuously and independently at a constant average rate.

6.10 CHECK YOUR PROGRESS

1. Binomial distribution is a discrete probability distribution, whereas normal distribution is a …… probability distribution.

2. Normal distribution was first discovered by an English Mathematician……. in 1733.

3. The normal curve extends to…….. on either side.

4. Area under the normal curve between $\overline{X}$ - 3σ and $\overline{X}$ + 3σ is ……...

6.11 ANSWERS TO CHECK YOUR PROGRESS

1. Continuous 2. Abraham Dimore 3. Infinity 4.0.9973,

6.12 TERMINAL QUESTIONS

1. Find the area under the normal curve in the following cases using the table:

   (i) Between Z = 0 and Z = 1.3

   (ii) Between Z = 0.75 and Z = 0.

   (iii)Between Z = -0.56 and Z = 2.45

   (iv)Between Z = 0.85 and Z = 1.96.

2. In a sample of 1000 workers, the mean weight is 45 kg with a standard deviation of 15 kg. Assuming the normality of the distribution, find the number of workers weighing between 40 and 60 kgs.

3. Find the probability that an item drawn at random from the normal distribution with mean 5 and S.D. 3 will be between 2.57 and 4.34.

4. A normal distribution has a mean (μ) = 12 and standard deviation (σ) =2. Find the area between X_1= 9.6 and X_2 = 13.8.

5. For a normal distribution, mean = 12 standard deviation = 2, find the area under the curve from X_1= 6 to X_2 = 18.

6. The marks obtained by the students in an examination are known to be normally distributed. If 10% of the students got less than 40 marks while 15% got over 80, what are the mean and standard deviation of marks?

7. In a certain examination, 15% of the candidates passed with distinction while 25% of them failed. It is known that a candidate fails if he obtains less than 40 marks (out of 100) while he must obtain at least 75 marks to pass with distinction. Calculate the mean and standard deviation of the distribution of marks, assuming this to be normal.

8. Assuming that the height of a group of men is normal, find the mean and standard deviation given that 84% of men have heights less than 65.2 inches and 68% have heights between 65.2 and 62.8 inches.

9. In a certain examination, the percentage of passes and distinctions was 46 and 9, respectively. Estimate the average marks obtained by the candidate and their standard deviation, the minimum pass and distinction marks being 40 and 75 respectively (Assume the distribution of marks to be normal).

10. The monthly incomes of 500 workers were found to be normally distributed with a mean of Rs. 2,000 and a standard deviation of Rs. 200. What was the lowest income among the richest 125 workers?

11. The incomes of a group of 5,000 persons were found to be normally distributed with mean = Rs. 900 and S.D. = Rs. 75. What was the highest income among the poorest 200?

    (Given: Area under the standard normal curve from Z=0 to Z=1.75 is 0.46).

12. The marks of students in a class are normally distributed with $\overline{X}$ = 6.7 and S.D.=1.2. Assuming the marks to be normally distributed, determine the maximum marks of the lowest 10% of the class.

13. In an intelligence test administered to 1,000 students, the average score was 42, and the standard deviation was 24. If the top 10% of the students get grade A, how many minimum marks, a student need to get to be able to get grade A?

14. (i) Name of the two methods available to fit the normal curve.

    (ii) Fit a normal curve to the following data by the Ordinate Method.

Class Interval:	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
***f*:**	5	8	12	8	7

15. Fit a normal curve to the following data by the area method:

Class Interval:	10.5 – 20.5	20.5 – 30.5	30.5 – 40.5	40.5 – 50.5	50.5 – 60.5	60.5 – 70.5	70.5 – 80.5
***f*:**	12	28	40	60	32	20	8

16. Fit a normal curve to the following data:

Mid-Interval:	61	64	67	70	73
***f*:**	5	18	42	27	8

17. Fit a normal curve to the following data:

Height (cm):	60 – 62	63 – 65	66 – 68	69 – 71	72 – 74
No. of Students	5	18	42	27	8

Given that $\overline{X}$ =67.45 cm and σ =2.92 cm, N=100

ANSWERS OF TERMINAL QUESTIONS

1. [(i) 0.4032, (ii) 0.2734, (iii) 0.7052, (iv) 0.1727]
2. 471
3. [0.2039]
4. [70.08%]
5. [99.74%]
6. [$\overline{X}$ = 62.15, σ = 17.16]
7. [$\overline{X}$ = 53.79, σ = 20.46]
8. [$\overline{X}$ = 64, σ = 1.2]
9. [$\overline{X}$ = 37.18, σ = 28.22]
10. [Rs. 2,134]
11. [Rs. 768.75]
12. [5.164 or 5]
13. [72.73 or 73]
14. [f = 3, 9, 14, 10, 4]
15. [f = 9, 26, 45, 54, 40, 20, 6]
16. [f = 4, 20, 41, 28, 7]
17. [f = 4,20,41,28,7]

6.13 SUGGESTED READINGS

1. Roy Ramendu, 'Principles of Statistics' Prayag Pustak Bhawan, Allahabad.
2. Gupta S. P. & Gupta M. P., 'Business Statistics' Sultan Chand & Sons, New Delhi.
3. Shukla S. M. & Sahai S. P., 'Advanced Statistics' Sahitya Bhawan Publications, Agra.
4. Goon, Gupta and Dasgupta, 'Basic Statistics' World Press Limited – Calcutta.
5. Fundamentals of Business Statistics – Sanchethi and Kappor. 6. Srivastava, Shenoy and Guptha, 'Quantitative Methods in Management'.

AREA UNDER THE NORMAL CURVE

The entries in the table are the probabilities that a random variable having the standard normal distribution assumes a value between o and z; they are given by the area under the curve shaded in the figure shown on the right hand side.

The Standard Normal Distribution.

Area under Normal Curve

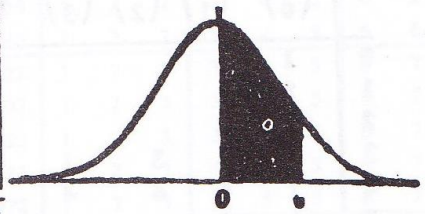
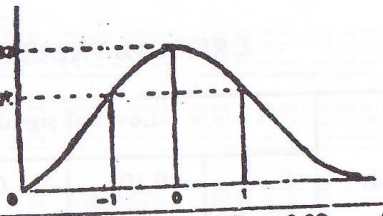


Table of Area

Z	0	1	2	3	4	5	6	7	8	9
0	·0000	·0040	·0080	·0120	·0160	·0199	·0239	·0279	·0319	·0359
·1	·0398	·0438	·0478	·0517	·0557	·0596	·0636	·0675	·0714	·0753
·2	0793	·0832	·0871	·0910	·0948	·0987	·1026	·1064	·1103	·1141
3	1179	·1217	·1255	·1293	·1331	·1368	·1406	·1443	·1480	·1517
4	1554	·1591	·1628	·1664	·1700	·1736	·1772	·1808	·1844	·1879
·5	·1915	·1950	·1985	·2019	·2054	·2083	·2123	·2157	·2190	·2224
·6	·2257	·2291	·2324	·2357	·2389	·2422	·2454	·2486	·2517	·2549
·7	·2580	·2611	·2642	·2673	·2703	·2734	·2764	·2794	2823	·2852
·8	·2881	·2910	·2939	·2967	·2995	·3023	·3051	·3078	·3106	·3133
·9	·3159	·3186	·3212	·3238	·3264	·3289	·3315	·3340	·3365	·3389
1·0	·3413	·3438	·3461	·3485	·3508	·3531	·3554	·3577	·3599	·3621
1·1	·3643	·3665	·3686	·3708	·3729	·3749	·3770	·3790	·3810	·3830
1·2	·3849	·3869	·3888	·3907	·3925	·3944	·3962	·3930	·3997	·4015
1·3	·4032	·4049	·4066	·4082	·4099	·4115	·4131	·4147	·4162	·4177
1·4	·4192	·4207	·4222	·4236	·4251	·4265	·4279	·4292	·4306	·4319
1·5	·4332	·4345	·4357	·4370	·4382	·4394	·4406	·4418	·4429	·4441
1·6	·4452	·4463	·4474	·4484	·4495	·4505	·4515	·4525	·4535	·4545
1·7	·4554	·4564	·4573	·4582	·4591	·4599	·4608	·4616	·4625	·4633
1·8	·4641	·4649	·4656	·4664	·4671	·4678	·4686	·4693	·4699	·4706
1·9	·4713	·4719	·4726	·4732	·4738	·4744	·4750	·4756	·4761	·4767
2·0	·4772	·4778	·4783	·4788	·4793	·4798	·4803	·4808	·4812	·4817
2·1	·4921	·4826	·4830	·4834	·4838	·4842	·4846	·4850	·4854	·4857
2·2	·4961	·4864	·4868	·4871	·4875	·4878	·4881	·4884	·4887	·4890
2·3	·4993	·4896	·4898	·4901	·4904	·4906	·4909	·4911	·4913	·4916
2·4	·4918	·4920	·4922	·4925	·4927	·4929	·4931	·4932	·4934	·4936
2·5	·4938	·4940	·4941	·4943	·4945	·4946	·4948	·4959	·4951	·4952
2·6	·4953	·4955	·4956	·4957	·4959	·4980	·4961	·4962	·4963	·4964
2·7	·4965	·4966	·4967	·4968	·4969	·4970	·4071	·4972	·4973	·4974
2·8	·4974	·4975	·4976	·4977	·4977	·4978	·4979	·4979	·4980	·4981
2·9	·4981	·4982	·4982	·4983	·4984	·4984	·4985	·4985	·4986	·4986
3·0	·4987	·4987	·4987	·4988	·4988	·4989	·4989	·4989	·4990	·4990
3·1	·4990	·4991	·4991	·4991	·4992	·4992	·4992	·4992	·4990	·4993
3·2	·4993	·4993	·4994	·4994	·4994	·4994	·4994	·4995	·4995	·4995
3·3	·4995	·4995	·4995	·4996	·4996	·4996	·4996	·4996	·4967	·4967
3·4	·4997	·4997	·4997	·4997	·4997	·4997	·4997	·4997	·4997	4978
3·5	·4998	·4998	·4998	·4998	·4999	·4998	·4998	·4998	·4998	·4998
3·6	·4998	·4998	·4999	·4999	·4999	·4999	·4999	·4999	·4999	·4999
3·7	·4999	·4999	·4999	·4999	·4999	·4999	·4999	·4999	·4999	.4999
3·8	·4999	·4999	·4999	·4999	·4999	·4999	·4999	·4999	·4999	.4999
3·9	·5000	·5000	·5000	·5000	·5000	·5000	·5000	·5000	·5000	.5000

ORDINATES (Y) of the STANDARD NORMAL CURVE



0	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0·0	·3989	·3989	·3989	·3988	·3986	·3984	·3982	·3980	·3977	·3973
0·1	·3970	·3965	·3961	·3956	·3951	·3945	·3939	·3932	·3925	·3918
0·2	·3910	·3902	·3894	·3885	·3876	·3867	·3857	·3847	·3836	·3825
0·3	·3814	·3802	·3790	·3778	·3765	·3752	·3739	·3725	·3712	·3697
0·4	·3683	·3668	·3653	·3637	·3621	·3605	·3589	·3572	·3555	·3538
0·5	·3521	·3503	·3485	·3467	·3448	·3429	·3410	·3391	·3372	·3352
0·6	·3332	·3312	·3292	·3271	·3251	·3230	·3209	·3187	·3166	·3144
0·7	·3123	·3101	·3079	·3056	·3034	·3011	·2989	·2966	·2943	·2920
0·8	·2897	·2874	·2850	·2827	·2803	·2780	·2756	·2732	·2709	·2685
0·9	·2661	·2637	·2613	·2589	·2565	·2541	·2516	·2492	·2468	·2444
1·0	·2420	·2396	·2371	·2347	·2323	·2299	·2275	·2251	·2227	·2203
1·1	·2179	·2155	·2131	·2107	·2083	·2059	·2036	·2012	1989	·1965
1·2	·1942	·1919	·1895	·1872	·1849	·1826	·1804	·1781	1758	·1736
1·3	·1714	·1691	·1669	·1647	·1626	·1604	·1582	·1561	1539	·1518
1·4	·1497	·1476	·1456	·1435	·1415	·1394	·1374	·1354	1334	·1315
1·5	·1295	·1276	·1257	·1238	·1219	·1200	·1182	·1163	1145	·1127
1·6	·1109	·1092	·1074	·1057	·1040	·1023	·1006	·0989	·0973	·0957
1·7	·0940	·0925	·0909	·0893	·0878	0863	·0848	·0833	·0818	·0804
1·8	0790	·0775	·0761	·0748	·0734	·0721	·0707	·0694	·0681	·0669
1·9	·0656	·0644	·0632	·0620	·0608	·0596	·0584	·0573	·0562	·0551
2·0	·0540	·0529	·0519	·0508	·0498	·0488	·0478	·0468	·0459	·0449
2·1	·0440	·0431	·0422	·0413	·0404	0396	·0387	·0379	·0371	·0363
2·2	·0355	·0347	·0339	·0332	·0325	·0317	·0310	·0303	·0297	·0290
2·3	·0283	·0277	·0270	·0264	·0258	·0252	·0246	·0241	·0235	·0229
2·4	·0224	·0219	·0213	·0208	·0203	·0108	·0194	·0189	·0184	0180
2·5	·0175	·0171	·0167	·0163	·0158	·0154	·0151	·0147	·0143	·0139
2·6	·0136	·0132	·0129	·0126	·0122	·0119	·0116	·0113	0110	·0107
2·7	0104	·0101	·0099	·0096	·0093	·0091	·0088	·0086	·0084	·0081
2·8	·0079	·0077	·0075	·0073	·0071	0099	·0067	·0065	·0063	·0061
2·9	·0060	·0058	·0056	·0055	·0053	·0051	·0050	·0048	·0047	0046
3·0	·0044	·0043	·0042	·0040	·0039	·0038	·0037	·0036	·0035	0034
3·1	·0033	·0032	·0031	·0030	0029	·0028	·0027	·0026	·0025	·0250
3·2	·0024	·0023	·0022	·0022	·0021	·0020	·0020	·0019	·0018	·0018
3·3	·0017	·0017	·0016	·0016	·0015	·0015	·0014	·0014	·0013	·0013
3·4	·0012	·0012	·0012	·0011	·0011	·0010	·0010	·0010	·0009	·0009
3·5	·0009	·0008	·0008	·0008	·0008	·0007	·0007	·0007	·0007	·0006
3.6	·0006	·0006	·0006	·0005	·0007	·0005	·0005	·0005	·0005	·0004
3.7	·0004	·0004	·0004	·0004	·0004	·0004	·0003	·0003	·0003	·0003
3.8	·0003	·0003	·0003	·0003	·0003	·0002	·0002	·0002	·0002	·0002
.9	·0002	·0002	·0002	·0002	0002	·0002	·0002	·0002	·0001	·0001

UNIT 7 : PROCEDURE OF TESTING A HYPOTHESIS

Structure

OBJECTIVES

At the end of this unit, you will be able to:

- Explain the different methods of testing hypothesis.
- Describe the application of testing of hypothesis.

7.1 INTRODUCTION

In a statistical investigation, the totality of units (objects) under consideration is called the population. A population that has a finite number of units is called a finite population. A population having an infinite number of units is called an infinite population. The units belonging to the population have characteristics such as height, weight, etc. Thus, in a statistical investigation, we may refer of the heights of students who study in a college. In another case, we may refer to the weights of mangoes that are grown on a tree.

When the population is vast, while conducting a statistical investigation, we may not be able to contact each and every unit in the population. And so, the investigation may be based on a sample (a representative portion of the population). In this case, the investigation is called a sample survey.

Suppose n units are selected from the population; these selected units form a sample of size n. While drawing a sample from the population, if the units are selected according to certain pre-assigned probabilities, such a sample is called a random sample. In case the probabilities are the same for all the units, the selection is called simple random sampling.

For a variable in the population, suppose we find constants such as mean, standard deviation, etc., these constants are called parameters of the population. On the other hand, if we find the mean, standard deviation, etc., of the sample, they are called statistics.

The parameter is a statistical constant of the population. The statistic is a function of the sample values:

Thus, the mean height of students in a college is a parameter. Whereas, the mean height of 50 randomly selected students of the college is a statistic. Statistical distribution of a population may be identified by one parameter (Poisson distribution) or it may be identified by more than one parameter (Normal distribution has two parameters).

The set of all the admissible values of the population parameter is called the Parameter Space. If one parameter alone is enough to describe the statistical distribution of the population, the parameter space is one-dimensional. On the other hand, if two parameters describe the population, the parameter space is two two-dimensional, and so on.

7.2 SAMPLING DISTRIBUTION AND STANDARD ERROR

Suppose a sample of size n is drawn from a population and the sample mean $\overline{X}$ is calculated. From the population, many such samples of the same size can be drawn. The set of all samples of size n that can be drawn from a population is called the Sample space. For each of the samples, it can be calculated. And so, there can be many values of $\overline{X}$ Suppose in the sample space, these different values of $\overline{X}$ are tabulated in the form of a frequency distribution, the resulting distribution is called the Sampling distribution of $\overline{X}$. The standard deviation of this sampling distribution is called the Standard error (S.E)

The distribution of values of a statistic for different samples of the same size is called the sampling distribution of the statistic.

Standard error (S.E.) of a statistic is the standard deviation of the sampling distribution of the statistic.

Sampling distributions of other statistics, such as sample variable, sample median, etc, can also be written down. In each of these cases, the corresponding standard deviation would be the standard error (S.E).

Consider a population whose mean is μ and the standard deviation is σ. In order that the discussions would be comfortable, here we restrict ourselves only to a large population. Then, the sampling distribution of $\overline{X}$ has mean μ and standard error $\sigma/\sqrt{n}$. That is $E(\overline{X}) = \mu$ and $S.E.(\overline{X}) = \sigma/\sqrt{n}$

Let a random sample of size n_1 be drawn from a population whose mean is μ_1 and the standard deviation is σ_1. Also, let a random sample of size n_2 be drawn from another population whose mean is μ_2, and standard deviations are σ_2 Let $\overline{X_1}$ be the mean of the first sample and $\overline{X_2}$ be the mean of the second sample. Then

$$E(\overline{x_1} - \overline{x_2}) = (\mu_1 - \mu_2) \text{ and } S.E.(\overline{x_1} - \overline{x_2}) = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$$

Standard Error or Proportions

In a population, suppose it is possible to make a dichotomous classification (classification into two classes) of units as those that possess an attribute and those that do not possess the attribute.

For example:

1. People of the village are classified as ‘literates’ and ‘illiterates’
2. Students of college are classified as ‘poor’ and ‘not poor’
3. Fruits are classified as ‘ripe’ and ‘unripe’

In a population, let P be the proportion of units that possess the attribute. From such a population, suppose a random sample of size n is drawn. Let x of these n units belong to the class that possesses the attribute. Then, $P = \frac{x}{n}$ is the sample proportion of the attribute.

Here, $P = \frac{x}{n}$ mean E(p)= P and standard error $S.E.(p) = \sqrt{\frac{PQ}{n}}$ where Q =1-P

Let a random sample of size n_1 be drawn from a population with a proportion P_1 of an attribute. *Let* x_1 Units in the sample possess the attribute. Then, the sample proportion is $P_1 = \frac{x_1}{n_1}$. Let a random sample of size n_2 be drawn from a population with a proportion P_2 of the attribute. Let x_2 units in this sample possess the attribute. Then, the sample proportion is $P_2 = \frac{x_2}{n_2}$. Here is the difference in the sample proportions $p_1 - p_2$ has mean $E(p_1 - p_2) = (p_1 - p_2)$ and standard error. $S.E.(p_1 - p_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$ Where $Q_1 = 1 - P_1$ and $Q_2 = 1 - P_2$.

Here, if $P_1 = P_2 = P$ then the standard error is $\sqrt{PQ\left[\frac{1}{n_1} + \frac{1}{n2}\right]}$

Thus, the following are the means and standard errors of some statistics.

Statistic	Mean	Standard Error

$\bar{x}$ (*Sample mean*)	μ	$\frac{\sigma}{\sqrt{n}}$
$\bar{x}_1 - \bar{x}_2$ (*difference of means*)	$\mu_1 - \mu_2$	$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
P (*Sample mean*)	P	$\sqrt{\frac{PQ}{n}}$
$P_1 - P_2$ (*difference of proportions*)	$P_1 = P_2$	$\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$
$P_1 - P_2$ (*when* $P_1 = P_2 = P$)	0	$\sqrt{pQ\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$

Utility of Standard Error

Standard Error (S.E) is a measure of variability of the statistic. It is useful in the estimation and testing of hypotheses.

1. In the theory of estimation, standard error is used to decide the efficiency and consistency of the statistic as an estimator.
2. In interval estimation, the standard error is used to write down the confidence intervals.
3. In testing of hypotheses, the standard error of the test statistic is used to standardize the distribution of the test statistic.

Statistical Inference

Statistical inference is that branch of statistics that deals with the theory and techniques of making decisions regarding the statistical nature of the population using samples drawn from the population.

Statistical inference has two branches. They are (i) Theory of estimation and (ii) Testing of hypotheses.

7.3 THEORY OF ESTIMATION

In statistics, we often come across a situation where we would talk about the likely value of a parameter of a population. Suppose a horticulture research centre has evolved a hybrid variety of banana. The centre wants us to know the mean weight (μ) of this variety of bananas. For this purpose, some bananas are randomly chosen and their mean weight is calculated. This mean weight (sample mean) is proposed as a likely mean of the population. Thus, the sample mean $\left(\overline{X}\right)$ is an estimator of the population mean (μ) For a specified sample, suppose the sample mean weight is 84 gms., This specific value of $\left(\overline{X}\right)$ It is called the estimate of the population mean.

An estimator of an unknown parameter is a statistic that specifies the likely value of that parameter. The estimate is a specific value of the estimator for a specified sample. An estimator is a statistic, whereas an estimate is a numerical value.

Estimation deals with the methods and techniques adopted for finding the likely value of a population parameter using statistics from a sample drawn from the population. There are two types of estimation. They are –

i) Point Estimation and

ii) Interval estimation

Point Estimation:

While estimating an unknown parameter, if a single value is proposed as the estimate, such estimation is called Point Estimation. Thus, based on the sample mean $\overline{X} = 84\ gms$. If we conclude that the population mean is 84 gms., it is a point estimation. Here $\overline{X}$ is a point estimator of the population mean μ. The specific value 84 gms is the point estimate of μ. Usually, a point estimator of n_1 is denoted by $\hat{\mu}$. And so, $\hat{\mu} = \overline{X}$.

Interval Estimation

In point estimation, we propose a single value as the estimate of the unknown parameter. In most situations, the value proposed is unlikely to be the actual value of the parameter. Instead, if we

propose a small interval around the point estimate as the likely interval to contain the parameter, our proposition would be stronger. This interval, which is likely to contain the parameter, is called an interval estimate.

In interval estimation, an interval $(T_1=T_2)$ which is likely to contain the parameter, is proposed as an estimator of the parameter. The interval $(T_1=T_2)$ It is called a confidence interval. The probability that the confidence interval contains the parameter is called the confidence coefficient. The limits T_1 and T_2 of the confidence interval is called confidence limits. These limits are based on the sampling distribution of the associated point estimator (statistic).

The probability that a confidence interval contains the parameter is called the confidence coefficient. It is denoted by $(1-\alpha)$ The confidence coefficient, when expressed as a percentage, is $(1-\alpha)\%$ In interval estimation, we may write down confidence intervals of different confidence coefficients, say, 95%, 99%, etc.

Example:

Let us consider the example cited earlier regarding the weight of bananas. Here, the unknown mean weight of bananas (population mean μ) is estimated using a sample of size n=100, say, for which the sample mean is $\overline{X}=84\ gms$ 84gms. Let the population standard deviation be $\overline{\sigma}=5$ gms. Let the weight of bananas be normally distributed. Then, $\overline{X}$ is normally distributed with a mean and a standard deviation $\frac{\sigma}{\sqrt{n}}=\frac{5}{100}$

1. $\overline{X}$ = 84 gms, is a point estimate of μ. And so, we say that bananas on average weigh 84gms.
2. 95% confidence interval for μ is

   $(\bar{x}-1.96\frac{\sigma}{\sqrt{n}},\bar{x}+1.96\frac{\sigma}{\sqrt{n}})$

   It is $(84-2.58\times\frac{5}{\sqrt{100}},84+2.58\times\frac{5}{\sqrt{100}})$.

And so, with 95% confidence, we say that bananas on average weigh between 83.02 and 84.98 gms.

3. 99% confidence interval for μ is $(\bar{x} - 2.58\frac{\sigma}{\sqrt{n}}, \bar{x} + 2.58\frac{\sigma}{\sqrt{n}})$

   It is $(84 - 2.58 \times \frac{5}{\sqrt{100}}, 84 + 2.58 \times \frac{5}{\sqrt{100}})$.

   It is (81,62,85,29). And so, with 99% confidence, we say that bananas on an average weight between 81.61 and 85.29 gms

Note: 1: We know that $\bar{x}$ is an estimator of the mean np of a binomial population. And so, an estimate of p is

$$\hat{p} = \frac{\bar{x}}{n} = \frac{3.38}{7} = 0.48$$

Note 2: The estimate of the mean λ of Poisson population is $\hat{\lambda} = \bar{x} = 0.2$

Note 3: The estimate of the mean n_1 of Poisson population is $\hat{\lambda} = \bar{x} = 1.2$

7.4 TESTING OF HYPOTHESES

Testing of hypotheses deals with the verification of the validity of presumptions regarding the parameter of the population using samples drawn from the population.

Suppose we presume that the mean income of Bangalore is Rs. 18 per day. In order to test this presumption, some Bangalorians are randomly selected and their mean daily income is found. If this sample mean is in the close neighborhood of Rs. 18, we conclude that the mean income of a Bangalorian differs from Rs. 18. Thus, if the sample mean is Rs. 18.10 since this value is close to Rs. 18, we conclude that mean income of a Bangalorian is Rs. 18 per day. On the other hand, if the sample mean is Rs. 19.60, since this value is far away from Rs. 18, we conclude that 'the mean income of a Bangalorian differs from Rs. 18'.

Statistical Hypothesis

A statistical hypothesis is an assertion regarding the statistical distribution of the population. It is a statement regarding the parameters of the population.

A statistical hypothesis is denoted by H.

Examples

1. H: The population has a mean $\mu = 25$.
2. H: The population is normally distributed with a mean $\mu = 25$ and standard deviation $\sigma = 2$

A hypothesis that completely specifies the statistical distribution of the population is called the Simple Hypothesis. It is a hypothesis that specifies all the parameters of the population. For example,

H: The population is normally distributed with a mean $\mu = 25$ It is a simple hypothesis

In a test procedure, to start with, a hypothesis is made. The validity of this hypothesis is tested. If the hypothesis is found to be true, it is accepted. On the other hand, if it is found to be untrue, it is rejected.

The hypothesis that is being tested for possible rejection is called the null hypothesis. The null hypothesis is denoted by H_0

If the null hypothesis is found to be untrue, another hypothesis that contradicts the null hypothesis is accepted. The hypothesis, which is accepted when the null hypothesis is rejected, is called the alternative hypothesis. The alternative hypothesis is denoted by H_1

Suppose a null hypothesis is $H_0: \mu = Rs.100$ (Mean wage is Rs. 100) Depending on the situation, the alternative hypothesis may be any one of the following.

$H_1: \mu \neq Rs.100$ (*Mean wage differs from Rs.*100)

$H_2: \mu > Rs.100$ (*Mean wage is more than Rs.*100)

$H_3: \mu < Rs.100$ (*Mean wage is less than Rs.*100)

$H_4: \mu \neq Rs.120$ (*Mean wage is* $Rs.120$)

Test Procedure:

The steps in the application of a statistical test procedure for testing a null hypothesis are as follows:

1. Setting up the null hypothesis.
2. Setting up the alternative hypothesis.
3. Identifying the test statistic and its null distribution
4. Identifying the critical region.
5. Drawing a random sample and conducting the test.
6. Making the decision (Giving the inference)

1. Setting up the null hypothesis

The hypothesis that is being tested for possible rejection is the null hypothesis. The null hypothesis may be that

a. The parameter is equal to a given value $(\mu = \mu_0)$
b. The parameters for the two population are equal. $(\mu_1 = \mu_2)$
c. The difference is insignificant (not significant)
d. The distribution is a good fit.
e. The attributes are independent, and so on.

2. Setting up the alternative hypothesis

The hypothesis that is to be accepted when the null hypothesis is rejected is the alternative hypothesis. It may be that

a. The parameter is not equal to the given value $(\mu \neq \mu_0)$
b. The parameter is greater than the given value $(\mu > \mu_0)$
c. The parameter for the two populations are not equal $(\mu_1 \neq \mu_2)$
d. The parameter for the first population is lesser than the parameter for the second population $(\mu_1 = \mu_2)$
e. The difference is significant

f. The distribution is not a good fit.

g. The attributes are dependent (not independent) and so on.

3. Identifying the test statistic and its null distribution

The test statistic is the statistic based on whose distribution the test is conducted. The statistical distribution of the test statistic under H_0 is called null distribution.

For testing $H_0 : \mu = \mu_0$ (*mean is euqual to* μ_0), the test statistic is $\bar{x}$ the sample mean. Under H_0 The distribution of $\bar{x}$ is $N(\mu_{0,} \sigma^2 / n)$ This is the null distribution of $\bar{x}$

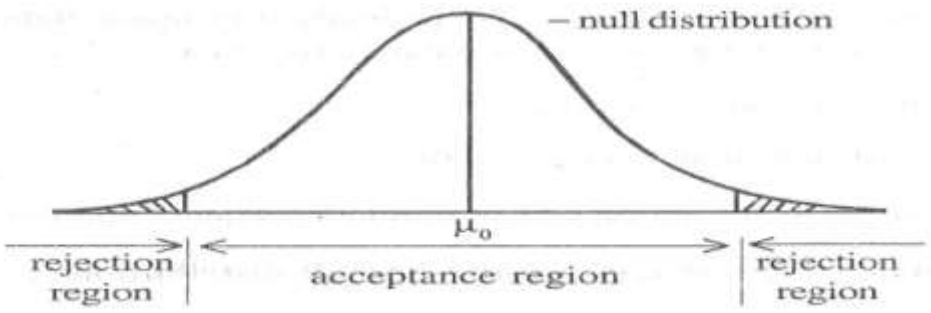
However, for convenience, the standardized form $Z = \frac{\overline{x - \mu_0}}{\sigma n}$ may be considered in which case, the null distribution is N (0,1)

4. Identifying the critical region.

The set of those values of the test statistic that lead to the rejection of the null hypothesis is called the critical Region (Rejection Region). The set of those values of the test statistic that leads to the acceptance of the null hypothesis is called the acceptance region.

The demarcation limits of the critical region are called Critical values.

The critical region is so formed that when H_0 true, the probability of its rejection is a small is pre-decided value. This pre-decided value is called the level of significance.



Level of Significance is a pre-decided upper limit for the probability of rejection of the null hypothesis when it is actually true. The level of significance is denoted by α. Usually, the pre-decided value α would be 0.05 or 0.01. In other words, it would be 5% or 1%.

Depending on the nature of the test and also on the nature of the alternative hypothesis, the critical region may have one part or it may have two parts. If the critical region has one part, the test is called a one-sided test (one-tailed test). If the critical region has two parts, the test is called a two-sided test (two-tailed test).

5. Drawing a random sample and conducting the test

A random sample of size n is drawn from the population. The value of the test statistic is calculated from the sample value. If the calculated value (observed value) belongs to the critical region, H_0 is rejected in favor of H_1. On the other hand, if the calculated value belongs to the acceptance region, H_0 is accepted.

6. Making the decision (Giving the inference)

The final decision (inference, conclusion) is announced. The decision may be

a. The new medicine is more effective than the old one

b. Average neonatal growth among male infants is the same as that among female infants.

Errors of the First and Second Kind

(Type I and Type II errors)

While testing a null hypothesis against an alternative hypothesis, one of the following situations arises.

	Actual fact	**Decision based on the sample**		**Error**
1	H_0 Is true	Accept H_0	Correct decision	---
2	H_0 is true	Reject H_0	Wrong decision	Type I
3	H_0 is not true	Accept H_0	Wrong decision	Type II

4	H_0 is not true	Reject H_0	Correct decision	---

Here, in situations (2) and (3), wrong decisions are made. These wrong decisions are termed as Error of the first kind (Type I error) and Error of the second kind (Type II error), respectively. Thus,

(i) Error of the first kind (Type I error) is making a wrong decision to reject the null hypothesis when it is true.

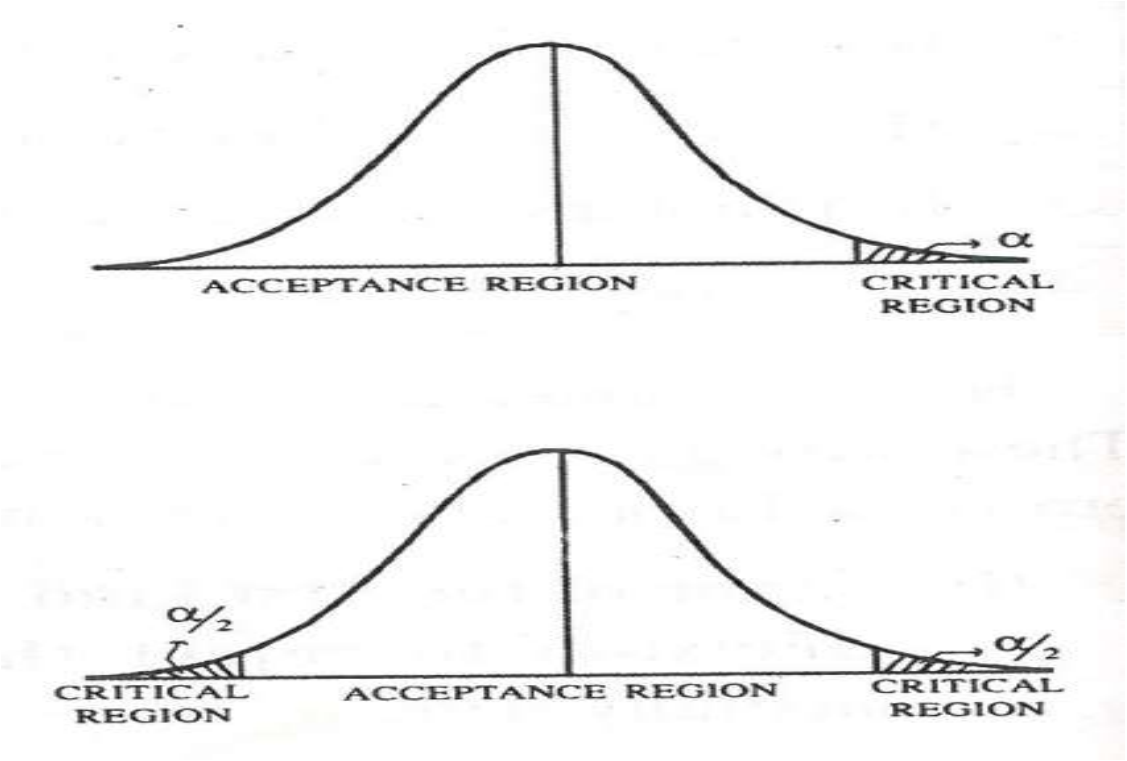
(ii) Error of the second kind (Type II error) is making a wrong decision to accept the null hypothesis when it is not true.

The probability of the occurrence of the first kind of error is α. It is the size of the test.

The probability of occurrence of the second kind of error is denoted by β.

The value $(1-\beta)$ It is called the power of the test. The power of a test is the probability of rejecting H_0 when it is not true. While testing, the level of significance α It is decided in advance. Then, the critical region is determined in such a way that the power $(1-\beta)$ is maximum. Thus, the critical values are based on the level of significance.

Two-Tailed and One-Tailed Tests (Two-sided and one-sided tests)



While testing a null hypothesis H_0 against an alternative hypothesis, if the critical region is considered at one tail of the null distribution of the test statistic, the test is one-tailed test (one-sided test).

On the other hand, if the critical region is considered at both the tails of the null distribution of the test statistic, the test is two tailed test (two-sided test)

The following are some of the one-tailed tests.

1. Testing $H_0: \mu = \mu_0$ *against* $H_1: \mu > \mu_0$
2. Testing $H_0: \mu_1 = \mu_2$ *against* $H_1: \mu_1 < \mu_2$
3. Testing *for goodness of fit*
4. Testing *for independence of attributes in acontigency table.*

The following are some of the two-tailed tests.

1. Testing $H_0: \mu = \mu_0 \; against \; H_1: \mu \neq \mu_0$
2. Testing $H_0: \mu_1 = \mu_2 \; against \; H_1: \mu_1 \neq \mu_2$
3. Testing $H_0: \sigma = \sigma_0 \; against \; H_1: \sigma \neq \sigma_0$

Exercise 1:

Past experience says that cashew kernels have a standard deviation of weight of 1.2 gms. Find the standard errors (S.E) of the mean weight of 40 randomly chosen kernels.

Solution:

Here, $\sigma = 1.2 gms. \; and \; n = 40$

The standard error of the sample mean is

$$S.E(\overline{X}) = \frac{\sigma}{n} = \frac{1.2}{\sqrt{40}} = 0.1897 \; gms.$$

Exercise 2:

The mean and standard deviation of the weight of the boys of a college are 47 kgs and 3.1 kgs, respectively. The mean and standard deviation of the weight of girls in the college are 45 kgs and 2.8 kgs. From the college, 16 boys and 9 girls are randomly selected.

i) Find the mean and standard deviation (S.E.) of the mean weight of the 9 selected boys.
ii) Find the mean and standard deviation of the mean weight of the 9 selected girls.
iii) Find the mean and standard deviation of the difference between the mean weight of the selected boys and the mean weight of the selected girls.

Solution:

Here,

μ_1 47 *kgs.*, σ_1 3.1 *kgs.*, $n_1 = 16$,

μ_2 45 *kgs.*, $\sigma_2 = 2.8$ *kgs.*, $n_2 = 9$

Let $\overline{x_1}$ be the mean weight of the selected boys

Let $\overline{x_2}$ the mean weight of the selected girls.

(i) $Mean = E(\overline{x_1}) = \mu_1 = 47$ *kgs.*

$$S..E.(\overline{x}) = \frac{\sigma_1}{\sqrt{n_1}} = \frac{3.1}{\sqrt{16}} = 0.775 \text{ kgs}$$

(ii) $Mean = E(\overline{x_2}) = \mu_2 = 45$ *kgs.*

$$S..E.(\overline{x_2}) = \frac{\sigma_2}{\sqrt{n_2}} = \frac{2.8}{\sqrt{9}} = 0.993 \text{ kgs}$$

(iii) $Mean = E(\overline{x_1} - \overline{x}_2) = \mu_1 = \mu_2.$

$= 47 - 45 = 2$*kgs.*

$$S..E.(\overline{x_1} - \overline{x}_2) = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}} + = \sqrt{\frac{(3.1)^2}{16}} + \frac{(2.8)^2}{9} = 1.213 \text{ kgs}$$

A sample investigation produces results, and with these results, decisions are made on the population. But such decisions involve an element of uncertainty, causing wrong decisions. A Hypothesis is an assumption that may or may not be true about a population parameter. For example, by tossing a coin 300 times, one may get 190 heads and 110 tails. In this instance, we are interested in testing whether the coin is unbiased or not. Therefore, we may conduct a test to

judge the significance of whether the difference is due to sampling. The procedure for carrying out a significance test is as follows:

Laying Down of Hypothesis

To verify our assumption, which is based on a sample study, we collect data and find out the difference between the sample value and the population value. If there is no difference or if the difference is very small, then our hypothesized value is correct, generally, two hypotheses must be constructed, and if one hypothesis is correct, the other one is rejected.

a) **Null Hypothesis:**

It is a very useful tool to test the significance of the difference. Any hypothesis concerning a population is called a statistical hypothesis. In the process of a statistical test, the hypothesis is rejected or accepted based on a sample drawn from the population. The statistician tests the hypothesis through observation and gives a probability statement. The simple hypothesis reveals that the value of the sample and the value of the population under study do not show any difference.

The hypothesis we have assumed is said to be a null hypothesis; it means that the true difference between the mean of the sample and the mean of the population is nil; the least difference found is unimportant. The rejection of the null hypothesis means that the true difference between the mean of the sample and the mean of the population is nil. The rejection of the null hypothesis reveals that the decision is correct.

For example:

i) The average height of the students of a university is 155 cms.
ii) The average daily sales of a firm are Rs. 1500.
iii) The average income of mean of a particular village is Rs. 100.

All these statements will have to be verified on the basis of sample tests. Generally, a hypothesis states that there is no difference between the mean of the sample and the population. A statistical hypothesis is a null hypothesis if it is accepted. A null hypothesis is denoted by H_0.

b) **Alternative Hypothesis:**

Rejection of H_0 leads to the acceptance of the alternative hypothesis, which is denoted by

H_2.

For e.g.

$H_0 = \mu = 155$ (Null hypothesis)

$H_1 = \mu \neq 155$ i.e., $\mu > 155$ or $\mu < 155$ (Alternative hypothesis)

When there are two hypotheses set up, the acceptance or rejection of a null hypothesis is based on a sample study. Thus, it leads to two wrong conclusions, i.e., (i) Rejecting H_0, when H_0 is true and (ii) Accepting H_0, when H_0 is false. This can be expressed in the following table:

	Decision from the sample	
	Accept H_0	Reject H_0
H_0 true	Correct	Wrong (Type I error)
H_0 false (H_1 true)	Wrong (Type II error)	Correct

By rewriting

Reject H_0 when it is true (Type I error) = a

Accept H_0 when it is false (Type II error) = β

Accept H_0 when it is true (correct decision)

Reject H_0 when it is false (correct decision)

Level of significance:

The maximum probability of committing a Type I error, which we specified in a test, is known as the level of significance. Generally, a 5% level of significance is fixed in statistical tests. This implies that we can have 95% confidence in accepting a hypothesis, or we could be wrong 5%.

Critical region:

The range of variation has two regions – the acceptance region and the critical region or rejection

region. If the sample statistics fall in a critical region, we have to reject the hypothesis, as it leads to a false decision. We go for H_1; if the computed value of the sample statistic falls in the rejecting region.

One-tailed and two-tailed tests:

The critical region under a normal curve, as stated earlier, can be divided in two ways,

(a) Two sides under a curve

(b) One side under a curve, and both are either at the right tail or at the left tail.

Making a decision or conclusion:

Finally, we come to a conclusion either to accept or reject the null hypothesis. The decision is based on the basis of computed value, whether it lies in the acceptance region or rejected region.

Standard Error

The standard deviation of the sampling distribution is called the standard error. E.g., $\overline{X}_1$, $\overline{X}_2$, $\overline{X}_3$,..... etc., are the means of all the samples drawn from the population. The standard deviation of all these means is the standard error of the mean. The formula for this is $\frac{\sigma}{\sqrt{n}}$.

Utility:

1. It is a useful instrument for testing the hypothesis. We may test the hypothesis at a 5% level of significance, which means that if the difference between observed and expected means is more than 1.96 S.E., the hypothesis is not accepted and one has to go for the alternative hypothesis. The level of significance can be 1%. Generally, the hypothesis is accepted if the difference is less than 3 S.E.; a 5% level is popular.
2. The reliability of a sample can be known.
3. The value of the parameters can be determined along with limits.

Exercise 3:

In city A, 32% of voters voted for Congress. In city B, 29% of voters voted for Congress.

i) Among 70 randomly selected votes from city A, if p_1 is the proportion of voters who voted for Congress. Find the mean and the standard error of p_1

ii) Among 60 randomly selected voters from city B, if p_2 is the proportion of voters who voted for Congress. Find the mean and the standard error of p_2

iii) Find the mean and standard error of $(p_1 - p_2)$

Solution

Here, $p_1 = \frac{32}{100} = 0.32$ *and* $p_2 = \frac{29}{100} = 0.29$

$n_1 = 70$ *and* $n_2 = 60$

(i) $Mean = E(p_1) = p_1. = 0.32$

$$S..E.(p_1) = \sqrt{\frac{P_1 Q_1}{n_1}} = \sqrt{\frac{0.32 \times 0.68}{70}} = 0.05575$$

(ii) $Mean = E(p_2) = p_2. = 0.29$

$$S..E.(p_2) = \sqrt{\frac{P_2 Q_2}{n_2}} = \sqrt{\frac{0.29 \times 0.71}{60}} = 0.05858$$

(iii) $Mean = E(p_1 - p_2) = p_1 - p_2. = 0.32 - 0.29 = 0.03$

$$S..E.(p_1 - p_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$$

$$= \sqrt{\frac{0.323 \times 0.68}{70}} + \frac{0.29 \times 0.71}{60}$$

$$= 0.08087$$

Exercise 4:

The proportion of women in a society is 0.48. Among 64 randomly selected people of the society, let p_1 be the proportion of women. In another selection of 86 people, let p_2 be the proportion of women.

Find –

i) Standard error of P_1

ii) Standard error of p_2

iii) Standard error of the difference $(P_1 - P_2)$

Solution:

Here $P_1 = P_2 = 0.48 = P$ (say)

$n_1 = 64$ *and* $= n_2 = 86$

$$(i)\ S..E.(p_1) = \sqrt{\frac{PQ}{n_1}} = \sqrt{\frac{0.42 \times 0.52}{64}} = 0.06245$$

$$(i)\ S..E.(p_1) = \sqrt{\frac{PQ}{n_1}} = \sqrt{\frac{0.48 \times 0.52}{86}} = 0.05387$$

$$S..E.(p_1 - p_2) = \sqrt{PQ\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$$

$$= \sqrt{0.48 \times 0.52\left[\frac{1}{64} + \frac{1}{86}\right]}$$

$$= \sqrt{0.48 \times 0.52\left[\frac{86 + 64}{64 \times 86}\right]}$$

$$= 0.08248$$

Exercise 5:

It is required to estimate the mean germination time needed for a variety of seeds. Ten randomly selected seeds are sown, and their germination time is noted as under.

Time (days): 28, 32, 27, 38, 30, 31, 30, 30, 27, and 33

By taking sample means as the estimator of a population mean, estimate the mean germination time.

Solution:

The sample mean is

$$\bar{x} = \frac{\sum x}{n} = \frac{28 + 32 + ... + 33}{10}$$

$$= \frac{306}{10} = 31\ day\ (approx)$$

Thus, the estimate of mean germination time is 31 days.

Hence, we conclude that, on average, the variety of seeds germinates in 31 days.

Exercise 6:

To know how much husbands, on average, are taller than wives, a random sample consisting of 8 couples is considered. The heights of husbands and wives in each case are measured, and their differences are noted as below:

Couple	1	2	3	4	5	6	7	8
Difference (in cms) :	5	7	12	23	6	-4	10	9

Using the sample mean $\bar{x}$ as an estimator of a population mean μ estimate μ

Solution:

The estimate is

$$\mu = \overline{x} = \frac{\sum x}{n} = \frac{5\_7 + ... + 93}{8}$$

$$\frac{68}{8} = 8.5\ cms$$

Thus, we conclude that husbands, on average, are 8.5 cm taller than their wives.

Exercise 7:

The owner of a bakery wants to estimate the average daily demand for cakes in the town. On 156 days, the survey revealed the following demand.

Demand (# cakes):	120-130	130-140	140-150	150-160	160-170
Number of days:	13	43	70	27	3

Estimate the average demand

Solution:

Demand	**Days (f)**	**Mid-value (x)**	**f.x.**
120-130	13	125	1625
130-140	43	135	5805
140-150	70	145	10150
150-160	27	155	4185
160-170	3	165	495
Total	156	-	22260

The estimate is $\bar{x} = \frac{\sum fx}{N} = \frac{22260}{156} = 143(approx).$

Thus, the estimated average demand is 143 cakes/day.

Decide on which distribution to use in Hypothesis testing

After deciding what level of significance to use, our next task in hypothesis testing is to determine the appropriate probability distribution. We have a choice between the normal distribution and the t distribution. The rules for choosing the appropriate distribution are similar to those we encountered in the unit on estimation. The Table below summarizes when to use the normal and t distributions in making tests of means. Later in this unit, we shall examine the distributions appropriate for testing hypotheses about proportions.

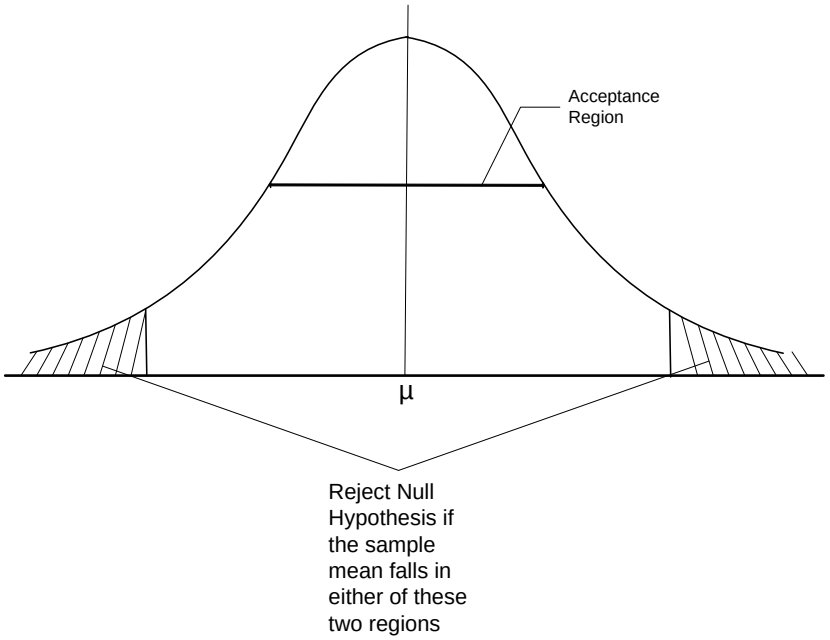
Remember one more rule when testing the hypothesized value of a mean. In estimation, use the finite population multiplier whenever the population is finite in size, sampling is done without replacement, and the sample is more than 5 percent of the population.

Conditions for using the Normal and t distributions in Testing the Hypothesis about means		
	When the Population Standard Deviation is Known	**When the Population Standard Deviation Is Not Known**
Sample size n is larger than 30	Normal distribution, z – table	Normal distribution, z – table
Sample size n is 30 or less, and we assume the population is normal or approximately so	Normal distribution, z – table	t Distribution, t – t-table

Two-tailed tests and One-Tailed tests

Two-Tailed Tests:

A ***two-tailed test*** of a hypothesis will reject the null hypothesis if the sample mean is significantly higher than or lower than the hypothesized population mean. Thus, in a two-tailed test, there are ***two rejection regions***. This is shown in the figure below:

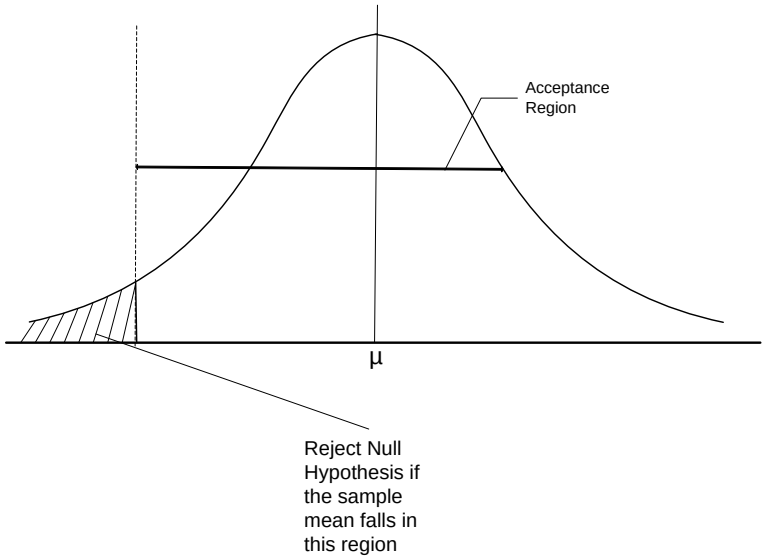


A two-tailed test is appropriate when the null hypothesis is $\mu = \mu_{Ho}$(where μ_{Ho} is some specified value) and the alternative hypothesis is $\mu \neq \mu_{Ho}$.

Assume that a manufacturer of light bulbs wants to produce bulbs with a mean life of $\mu = \mu_{Ho}$ = 1,000 hours. If the lifetime is shorter, he will lose customers to his competition; if the lifetime is longer, he will have a very high production cost because the filaments will be excessively thick. In order to see whether his production process is working properly, he takes a sample of the output to test the hypothesis H_o: μ = 1,000. Because he does not want to deviate significantly from 1,000 hours in either direction, the appropriate alternative hypothesis is H_1: $\mu \neq$ 1,000, and he uses a two-tailed test. That is, he rejects the null hypothesis if the mean life of bulbs in the sample is either ***too far above*** 1,000 hours or ***too far below*** 1,000 hours.

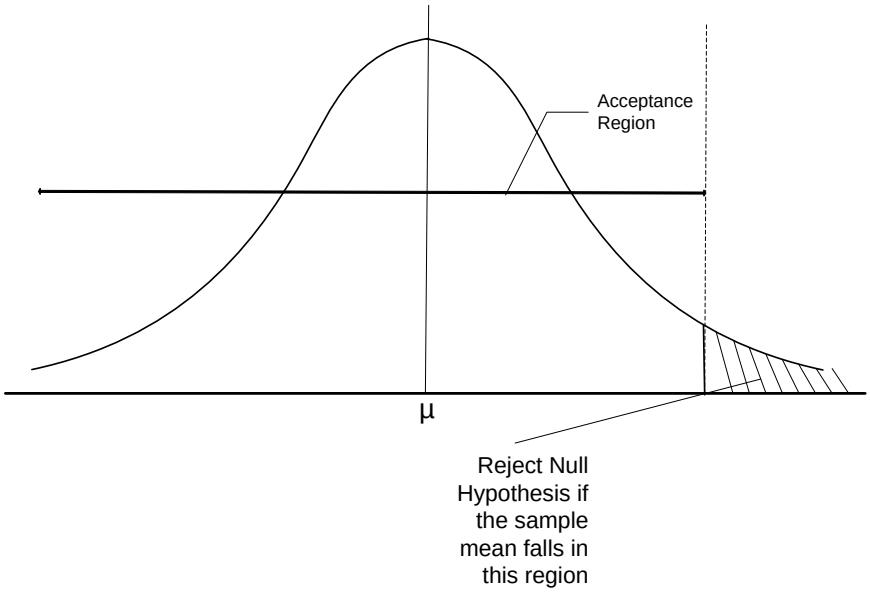
However, there are situations in which a two-tailed test is not appropriate, and we must use a one-tailed test. Consider the case of a wholesaler that buys light bulbs from the manufacturer discussed earlier. The wholesaler buys bulbs in large lots and does not want to accept a lot of bulbs unless their mean life is at least 1,000 hours or a minimum of 1000 hours. As each shipment arrives, the wholesaler tests a sample to decide whether it should accept the shipment. The company will reject the shipment only if it feels that the mean life is below 1,000 hours. If it feels that the bulbs are better than expected (with a mean life above 1,000 hours), it certainly will not reject the shipment because the longer life comes at no extra cost. So, the wholesaler's hypotheses are Ho: μ = 1,000

hours and H_1: $\mu < 1{,}000$ hours. It rejects H_o only if the mean life of the sampled bulbs is significantly below 1,000 hours. This situation is illustrated in the figure below. From this figure, we can see why this test is called a left-tailed test (or a lower-tailed test).



In general, a left-tailed (lower-tailed) test is used if the hypotheses are Ho: $\mu = \mu_{Ho}$ and H_1: $\mu < \mu_{Ho}$. In such a situation, it is sample evidence with the sample mean significantly below the hypothesized population mean that leads us to reject the null hypothesis in Favor of the alternative hypothesis. Stated differently, the rejection region is in the lower tail (left tail) of the distribution of the sample mean, and that is why we call this a lower-tailed test.

A left-tailed test is one of the two kinds of one-tailed tests. As you have probably guessed by now, the other kind of one-tailed test is a right-tailed test (or an upper-tailed test). An upper-tailed test is used when the hypotheses are Ho: $\mu = \mu_{Ho}$ and H_1: $\mu > \mu_{Ho}$**.** Only values of the sample mean that are significantly above the hypothesized population mean will cause us to reject the null hypothesis in favour of the alternative hypothesis. This is called an upper-tailed test because the rejection region is in the upper tail of the distribution of the sample mean.



This is to remind you again that, in each example of hypothesis testing, when we accept a null hypothesis based on ***sample information***, we are really saying that there is no ***statistical evidence*** to reject it. We are ***not*** saying that the null hypothesis is true. The ***only way*** to prove a null hypothesis is to know the population parameter, and that is not possible with sampling. Thus, we accept the null hypothesis and behave as if it is true simply because we can find no evidence to reject it.

Assumptions in Hypothesis Testing

- Don't use sample results to decide whether to use a two-tailed, upper-tailed, or lower-tailed test.
- Before any data are collected, the form of the test is determined by what the decision maker believes or wants to detect.

7.5 SUMMARY

Hypothesis testing begins with an assumption, called a *hypothesis, that* we make about a population parameter. Then we collect sample data, produce sample statistics, and use this information to decide how likely it is that our hypothesized population parameter is correct. Say that we assume a certain value for a population mean. To test the validity of our assumption, we gather sample data and determine the difference between the hypothesized value and the actual value of the sample

mean. Then we judge whether the difference is significant. The smaller the difference, the greater the likelihood that our hypothesized value for the mean is correct. The larger the difference will lead to the smaller the likelihood.

Unfortunately, the difference between the hypothesized population parameter and the actual statistic is more often neither so large that we automatically reject our hypothesis nor so small that we just as quickly accept it. So, in hypothesis testing, as in most significant real-life decisions, clear-cut solutions are the exception, not the rule. In estimation, use the finite population multiplier whenever the population is finite in size, sampling is done without replacement, and the sample is more than 5 percent of the population.

7.6 GLOSSARY

- **Hypothesis:** A condition from which something follows. It is an illusion
- **Simple hypothesis:** A hypothesis that specifies the exact distribution

7.7 CHECK YOUR PROGRESS

1. For the following cases, specify which probability distribution to use in hypothesis testing:

   a. H_o: $\mu = 27$, H_1: $\mu \neq 27$, $\bar{x} = 33$, sample $\sigma = 4$, $n = 25$

   b. H_o: $\mu = 98.6$, H_1: $\mu > 98.6$, $\bar{x} = 99.1$, $\sigma = 1.5$, $n = 50$

   c. H_o: $\mu = 3.5$, H_1: $\mu < 3.5$, $\bar{x} = 2.8$, sample $\sigma = 0.6$, $n = 18$

   d. H_o: $\mu = 382$, H_1: $\mu \neq 382$, $\bar{x} = 363$, $\sigma = 68$, $n = 12$

   e. H_o: $\mu = 57$, H_1: $\mu > 57$, $\bar{x} = 65$, sample $\sigma = 12$, $n = 42$

2. Theatre owners in India know that a hit movie ran for an average of 84 days with a standard deviation of 10 days in each city where the movie was screened. A particular movie distributor was interested in comparing the popularity of movies in his region with that of the population. He randomly chose 75 theatres at random in the region and found a popular movie that ran for 81.5 days.

I State appropriate hypotheses for testing whether there was a significant difference between theatres in the distributor's region and the population.

II. At a 1% significance level, test these hypotheses.

7.8 ANSWERS TO CHECK YOUR PROGRESS

This means the rejection area under both tails is .01, and the area of the acceptable region is 0.99. There for are of, half of the acceptable region is $\frac{0.99}{2}$ = .4950 means a z value of 2.58.

Therefore, the limits of the acceptable region are:

z = ± 2.58 or $\bar{x} = \mu_{Ho} \pm \frac{z\sigma}{\sqrt{n}} = 84 \pm 2.58x\frac{10}{\sqrt{75}}$

= 81.02 lower limit and 86.98 as the upper limit.

Because the observer value is in the acceptance region, we do not reject the null hypothesis H_0. The length of the run of the movie is the same as the other theatres.

Or in another way:

The observed z value is $\bar{x} - \frac{\mu_{H\circ}}{SE}$ where SE = $\frac{\sigma}{\sqrt{n}} = \frac{81.4-84}{(1.155)}$

= – 2.17. The acceptable z region is ± z = ± 2.58

7.9 TERMINAL QUESTION

1. a) t with 24 df (degrees of freedom)
   b) z or normal distribution: right-tailed test
   c) t with 17 df
   d) z or normal distribution: two-tailed test
   e) In this case, it's actually t with 41 df. Since 41 df is not there, we use a normal distribution test table.

2. Given the following data:

   σ = 10 days, n = 75 theatres $\bar{x}$ = 81.5

   Ho: μ = 84 days H_1: μ ≠ 84 days α = 0.01

3. Microsoft estimated the previous year that 35% of the potential software buyers were planning to wait to purchase the new OS Windows Vista until an upgrade had been released. After an

advertising campaign to reassure the public, Microsoft surveyed 3000 buyers and found 950 who were still skeptical. At a 5% level of significance, can the company conclude that the proportion of sceptical people has decreased? (Null hypothesis is rejected. Use Z distribution.

7.10 SUGGESTED READINGS

1. Basic Statistics – Goon, Guptha and Dasgupta – World Press Limited – Calcutta.
2. Fundamentals of Business Statistics – Sanchethi and Kappor.
3. Quantitative Methods in Management – Srivastava, Shenoy and Guptha.
4. Business Statistics – Guptha and Guptha.

UNIT 8 : SIGNIFICANCE TEST IN ATTRIBUTES

Structure

OBJECTIVES

At the end of this unit, you will be able to:

- Explain the different methods of testing the hypothesis.
- Describe the test of significance for attributes.
- Explain the Moment generating function.
-

8.1 INTRODUCTION

Industrial applications of statistics are often concerned with making decisions about populations and population parameters. For example, decisions about which of two processes is better or whether to discontinue production on a particular machine because it is producing an economically unacceptable number of defective components are often based on determining the mean or

standard deviation of a population, calculated using sample data drawn from the population. In reaching these decisions, certain assumptions are made, which may or may not be true. The assumptions made are called statistical hypothesis or just hypothesis and are usually concerned with statements about the probability distribution of populations.

8.2 TESTING OF HYPOTHESIS

A sample investigation produces results, and with these results, decisions are made on the population. But such decisions involve an element of uncertainty, causing wrong decisions. A Hypothesis is an assumption that may or may not be true about a population parameter. For example, tossing a coin 300 times, one may get 190 heads and 110 tails. At this instance, we are interested in testing whether the coin is unbiased or not. Therefore, we may conduct a test to judge whether the difference is due to sampling. The procedure for carrying out a significance test is as follows:

Laying Down of Hypothesis

To verify our assumption, which is based on a sample study, we collect data and find out the difference between the sample value and the population value. If there is no difference or if the difference is very small, then our hypothesized value is correct, generally, two hypotheses must be constructed, and if one hypothesis is correct, the other one is rejected.

a) **Null Hypothesis:**

It is a very useful tool to test the significance of the difference. Any hypothesis concerning a population is called a statistical hypothesis. In the process of a statistical test, the hypothesis is rejected or accepted based on a sample drawn from the population. The statistician tests the hypothesis through observation and gives a probability statement. The simple hypothesis reveals that the value of the sample and the value of the population under study do not show any difference.

The hypothesis we have assumed is said to be the null hypothesis; it means that the true difference between the mean of the sample and the mean of the population is nil; the least difference found is unimportant. The rejection of the null hypothesis means that the true difference between the mean of the sample and the mean of the population is nil. The rejection of the null hypothesis reveals that the decision is correct.

For example:

i) The average height of the students of a university is 155 cms.

ii) The average daily sales of a firm are Rs. 1500.

iii) The average income of mean of a particular village is Rs. 100.

All these statements will have to be verified on the basis of sample tests. Generally, a hypothesis states that there is no difference between the mean of the sample and the population. A statistical hypothesis is a null hypothesis if it is accepted. A null hypothesis is denoted by H_0.

b) **Alternative Hypothesis:**

Rejection of H_0 leads to the acceptance of the alternative hypothesis, which is denoted by H_2. For example,

$H_0 = \mu = 155$ (Null hypothesis)

$H_1 = \mu \neq 155$ i.e., $\mu > 155$ or $\mu < 155$ (Alternative hypothesis)

When there are two hypotheses set up, the acceptance or rejection of a null hypothesis is based on a sample study. Thus, it leads to two wrong conclusions, i.e. (i) Rejecting H_0, when H_0 is true (ii) Accepting H_0, when H_0 is false. This can be expressed in the following table:

	Decision from the sample	
	Accept H_0	Reject H_0
H_0 true	Correct	Wrong (Type I error)
H_0 false (H_1 true)	Wrong (Type II error)	Correct

By rewriting

Reject H_0 when it is true (Type I error) = a

Accept H_0 when it is false (Type II error) = β

Accept H_0 when it is true (correct decision)

Reject H_0 when it is false (correct decision)

Level of significance:

The maximum probability of committing Type I error, which we specified in a test, is known as the level of significance. Generally, a 5% level of significance is fixed in statistical tests. This implies that we can have 95% confidence in accepting a hypothesis, or we could be wrong 5%.

Critical region:

The range of variation has two regions – acceptance region and critical region or rejection region. If the sample statistics falls in the critical region, we have to reject the hypothesis, as it leads to a false decision. We go for H_1; if the computed value of the sample statistic falls in rejecting region.

One-tailed and two-tailed tests:

The critical region under a normal curve, as stated earlier, can be divided in two ways,

(a) Two sides under a curve

(b) One side is under a curve, and both are either at the right tail or at the left tail.

Making a decision or conclusion:

Finally, we come to a conclusion either to accept or reject the null hypothesis. The decision is based on the basis of computed value, whether it lies in the acceptance region or rejected region.

8.3 STANDARD ERROR

The standard deviation of the sampling distribution is called the standard error. E.g., $\overline{X}_1$, $\overline{X}_2$, $\overline{X}_3$,..... etc., are the means of all the samples drawn from the population. The standard deviation of all these means is the standard error of the mean. The formula for this is $\frac{\sigma}{\sqrt{n}}$.

Utility:

1. It is a useful instrument in testing the hypothesis. We may test the hypothesis at a 5% level of significance, which means, if the difference between observed and expected means is more than 1.96 S.E., the hypothesis is not accepted, and one has to go for the alternative hypothesis. The level of significance can be 1%. Generally, the hypothesis is accepted if the difference is less than 3 S.E.; a 5% level is popular.
2. The reliability of a sample can be known.
3. The value of the parameters can be determined along with limits.

Now we discuss the various tests of significance to be applied on various situations. They are:

1. Tests of significance for attributes

2. Tests of significance for variables

8.4 TESTS OF SIGNIFICANCE FOR ATTRIBUTES

The sampling of attributes may be regarded as the drawing of samples from a population whose members consist of the presence or absence of a particular characteristic. For example, in the study of attribute blindly, a sample may be drawn, and its members are classified as blind and not blind. The presence of the attribute may be represented by p, and the absence of the attribute may be represented by q. Thus, of 1000 people, 25 are blind and the remaining are not blind. In other words, $p = \frac{25}{1000}$ or 0.025 and q = 0.975. The various types of significance test may be studied under the following heads:

A) **Test of a number of success:**

This follows a binomial distribution. Formula:

S.E. of No. of success = $\sqrt{npq}$

n = size of sample

p = probability of success in each trial

q = (1 – p), i.e., probability of failure

Example 1:

Out of a consignment of 1,00,000 tennis balls, 400 were selected at random and examined. It was found that 20 of these were defective. How many defective balls can you reasonably expect to have in the whole consignment at a 95% confidence level?

Solution:

Here $p = \frac{20}{400} = 0.05$

q = 0.95

$\overline{X} = np = 1,00,000 \times (0.05) = 5,000$

$S.E. = \sqrt{npq} = \sqrt{1,00,000 \times .05 \times .95}$

$= \sqrt{4750} = 68.9$

95% confidence limits are

$5,000 \pm 1.96 \times 68.9 = 5,000 \pm 135.044$ (or)

5135 and 4.865

Example 2:

In a sample of 500 people from a village in Rajasthan, 280 are found to be rice eaters and the rest wheat eaters. can we assume that both the food articles are equally popular?

Solution:

We take the hypothesis that the food articles are equally popular.

Then, the expected frequency of wheat eaters and rice eaters is:

250:250.

$S.E. = \sqrt{npq} = \sqrt{500 \times \frac{1}{2} \times \frac{1}{2}} = 11.18$

Difference between actual and observed = 280 – 250 = 30

$\frac{\text{Difference}}{\text{S.E.}} = \frac{30}{11.18} = 2.68$

The difference is more than 2.58 S.E. at the 1% level. It is not because of sampling fluctuations. Therefore, we may assume that the food articles are not equally popular.

Example 3:

A coin is tossed 400 times and it turns up head 216 times. Discuss whether the coin may be an unbiased one, and explain briefly the theoretical principles you would use for this purpose.

Solution:

An unbiased coin turns up head $= \frac{1}{2}$

The expected number of heads in tosses of 400 = 200

But the observed number of heads = 216

$$S.E. = \sqrt{npq}$$

$$= \sqrt{400 \times \frac{1}{2} \times \frac{1}{2}} = \sqrt{100} = 10$$

Deviation from actual = 216 – 200 = 16

$$Z = \frac{Difference}{S.E.} = \frac{16}{10} = 1.6$$

Since the observed deviation is 1.6 times the S.E., which is less than 1.96 S.E., (5% level), it can be concluded that the hypothesis is accepted. Therefore, the coin is an unbiased one.

B) **Tests of proportion of success:**

Instead of taking the number of success in each sample, a proportion of success i.e., $\frac{1}{n}$ is recorded.

Formula:

$$S.E. = \sqrt{\frac{pq}{n}}$$

Example 4: Random samples of 500 pineapples were taken from a large consignment, and 65 were found to be bad. Show that the standard error of the population of bad ones in this sample is of size 0.015, and deduce that the percentage of bad pineapples in the consignment almost certainly lies between 8.5 and 17.5.

Solution:

Here $p = \frac{65}{500} = 0.13$,

$$q = 1 - 0.13 = 0.87$$

$$S.E. = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.13 \times 0.87}{500}}$$

$$= \sqrt{\frac{0.1131}{500}}$$

$$= \sqrt{.000226} = 0.015$$

The limits of the percentage of bad pineapples in the consignment are:

$$(0.13 \pm 3S.E.) \times 100 = (0.13 \pm 3 \times 0.015)100$$

$$= (.13 \pm .045)\ 100$$

$$= (13 \pm 4.5) = 17.5 \text{ and } 8.5$$

Note: 3 S.E. limits are "almost certain".

Example 5: A wholesaler of apples claims that only 4% of the apples supplied by him are defective. A random sample of 600 apples contained 36 defective apples. Test the claim of the wholesaler.

Solution:

$$\text{S.E.} = \sqrt{\frac{pq}{n}}$$

$$= \sqrt{\frac{.96 \times .04}{600}} = 0.008$$

95% confidence limit = $p \pm 1.96$ S.E.

$$= p \pm 1.96 \times 0.008$$

$$= .96 \pm 0.01568$$

$$= 0.94432 \text{ to } = .97568$$

Out of 600 apples, good apples may be between 94432+600 = 566.59 to .97568 × 600 = 585.4 or 567 to 585. Therefore, the number of defective is expected to be between 15 to 33 apples. He claims that 4% of apples are defective. But the actual number is 36 defectives. Hence, his claim cannot be accepted.

Example 6: A sample size of 600 persons selected at random from a large city shows

that the percentage of males in the sample is 53. It is believed that the ratio of males to the total population in the city is $\frac{1}{2}$. Test whether this belief is confirmed by the observation.

Solution:

Let the null hypothesis be that the number of males to total population is $\frac{1}{2}$ or .5

The observed value = .53

$$S.E. = \sqrt{\frac{pq}{n}} = \sqrt{\frac{\frac{1}{2}\times\frac{1}{2}}{600}} = \sqrt{\frac{\frac{1}{4}}{600}} = \sqrt{\frac{1}{2400}} = 0.02$$

$$S.E. = \frac{0.53-0.05}{S.E.} = \frac{(0.53-0.5)}{0.02} = 1.5$$

Since Z is less than 1.96, the difference is not significant at the 5% level of confidence and could have arisen because of sampling fluctuations. Therefore, the null hypothesis cannot be rejected. The belief is confirmed.

C) **Tests of Difference in Proportions:**

We draw two samples from different populations and verify whether the proportion of success is significant or not.

Formula:

$$S.E.(p_1 - p_2) = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

If S.E. $(p_1 - p_2) < 1.96$

The difference is regarded as due to random sampling fluctuations.

Example 7: One thousand articles from a factory are examined and found to be 3% defective. Fifteen hundred similar articles from a second factory are found to be only 2% defective. Can it reasonably be concluded that the product of the first factory is inferior to the second?

Solution:

Let us set up the null hypothesis

$$H_0 : p_1 = p_2$$

$$p_1 = \frac{30}{1000} = 0.03$$

$$p_2 = \frac{30}{1500} = 0.02$$

$$S.E.(p_1 - p_2) = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$p = \frac{(1000 \times 0.03) + (1500 \times 0.02)}{1000 + 1500} = \frac{(30 + 30)}{2500} = 0.024$$

$$S.E. = \sqrt{0.024 \times 0.976\left(\frac{1}{1000} + \frac{1}{1500}\right)}$$

= 0.006

$$Z = \frac{0.03 - 0.02}{0.006} = 1.67$$

At a 95% level of confidence, $Z = 1.96$, the difference is not significant. The null hypothesis that $p_1 = p_2$is accepted.

Example 8: A machine puts out 16 imperfect articles in a sample of 500. After the machine is overhauled, it puts out 3 imperfect articles in a batch of 100. Has the machine improved?

Solution:

$p_1 = \frac{16}{500} = 0.032$ (in the first sample)

$p_2 = \frac{3}{100} = 0.03$ (in the second sample)

Let us assume that the machine has not improved after the overhauling.

$$S.E.(p_1 - p_2) = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$p = \frac{500 \times 0.032 + 100 \times 0.3}{500 + 100}$$

$$= \frac{16 + 3}{600} = 0.03$$

q = 1 – 0.03 = 0.97

$$S.E.(p_1 - p_2) = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$= \sqrt{(0.03)(0.97)\left(\frac{1}{500}\right) + \left(\frac{1}{100}\right)}$$

$$= \sqrt{(0.03)(0.97)[0.002 + 0.01]}$$

= 0.0187

$$Z = \frac{0.032 - 0.03}{0.0187} = \frac{0.002}{0.0187} = 0.106$$

Since the difference is less than 2.58 S.E. (1% level), the result of the experiment supports the hypothesis. Therefore, we conclude that the machine has not improved after the overhauling.

Example 9: In a random sample of 1000 persons from town A, 400 are found to be consumers of wheat. In a sample of 800 from town B, 400 are found to be consumers of wheat. Do these data reveal a significant difference between town A and town B as far as the proportion of wheat consumers is concerned?

Solution:

Let us assume the hypothesis that the two towns do not differ so far as the proportion of wheat consumption.

H_0: $p_1 = p_2$

$$p_1 = \frac{400}{1000} = 0.4,\ p_2 = \frac{400}{800} = 0.5$$

$$p = \frac{(1000 \times 0.4) + (800 \times 0.5)}{1000 + 800}$$

$$= \frac{4}{9} \qquad \therefore q = \frac{5}{9}$$

$$S.E.(p_1 - p_2) = \sqrt{\frac{4}{9} \times \frac{5}{9}\left(\frac{1}{1000} + \frac{1}{800}\right)}$$

$$= \sqrt{\frac{20}{81} \times \frac{9}{4000}} = 0.024$$

$$p_1 - p_2 = 0.4 - 0.5 = 0.1$$

$$\frac{\text{Difference}}{\text{S.E.}} = \frac{0.1}{0.024} = 4.17$$

Since the difference is more than 2.58 S.E. (1% level), it could not have arisen due to fluctuations of sampling. Hence, the data reveal a significant difference between town A and town B as far as the proportion of wheat consumers is concerned.

10.5 MOMENT GENERATING FUNCTION

Suppose that X is a random variable: that is, X is a function from the sample space to the real numbers, in computing various characteristics of the random variable X, such as E(X) or V(X), we work directly with the probability distribution of X. [the probability distribution is given by a function: either the p.d.f (Probability distribution function) in the continuous case, or the point probabilities $p(x_i) = P(X = x_i)$ in the discrete case.]

Let X be a random variable for an experiment taking values in a subset S of R. The moment generating function of X is the function Mx defined by

$M_x(t) = E[\exp(tx)]$ for t in R`

Note that since exp(tx) is a nonnegative random variable, $M_x(t)$ exists as a real number or positive infinity for any t.

(1) Show that if X has a discrete distribution with density function f, then

$$M_x(t) = \sum_{x \in S} e^{tx} f(x)$$

(2) Show that if X is continuous with a density function f, then

$$M_x(t) = \int_S e^{tx} f(x) dx$$

Since the exponential function is positive, the moment generating function of X always exists,

either as a real number or as positive infinity.

Example-10

Suppose that X is uniformly distributed over the interval [a, b] therefore, the m.g.f is given by

$$M_x(t) = \int_a^b \frac{e^{tx}}{b-a} dx$$

$$= \frac{1}{(b-a)t}\left[e^{bt} - e^{at}\right], t \neq 0$$

Note: The m.g.f. of the sum of a number of independent variables is the product of their m.g.f.

$$E\left\{e^{t(x_1+x_2+x_3+\ldots)}\right\} = E\left(e^{tx_1}\right).E\left(e^{tx_2}\right).E\left(e^{tx_3}\right)\ldots$$

m.g.f. of Binomial Distribution

We know that the relative frequency of x successes is ${}^nC_x p^x q^{n-x}$ in the case of the Binomial distribution.

Therefore, the m.g.f. about origin will be given by

$$M_0(t) = \sum_{x=0}^{n} e^{tx} {}^nC_x p^{n-x}$$

$$= \sum_{x=0}^{n} {}^nC_x \left(pe^t\right)^x q^{nx}$$

$= (q+pe^t)^n$

$$= \left[q+p\left(1+t+\frac{t^2}{2!}+\ldots\right)\right]^n = \left[1+pt+\frac{pt^2}{2!}+\ldots\right]^n$$

$$M_0(t) = \left[1+pt+\frac{pt^2}{2!}+...\right]^n$$

Also, we know that the mean in this distribution is given by m=np and

$M_a(t) = e^{-at} M_0(t)$

The m.g.f. The about mean will be given by

$M_m(t) = e^{-mt} M_0(t)$ where m = np

$= e^{-npt}(q+pe^t)^n$

$= [qe^{-pt}+pe^{qt}]^n$

m.g.f. of Poisson distribution

We know that the probability of x successes in the case of the Poisson distribution is given by

$$e^{-m}\frac{m^x}{x!}$$

m.g.f. about origin will be given by

$$M_0(t) = \sum_{x=0}^{\infty} e^{tx}\left(\frac{e^{-m}m^x}{x!}\right)$$

$$M_0(t) = e^{-m}\sum_{x=0}^{\infty}\frac{\left(me^t\right)^x}{x!} = e^{-m}e^{me^t}$$

$$e^t = \mu'_2 = \sum_{x=0}^{n}\left\{x(x-1)+x\right\}^n C_x p^x q^{n-x}$$

or $M_0(t) = e^{m(e^t-1)}$

Also we know that the mean in this distribution is m, and where

$M_a(t) = e^{-at} M_0(t)$

m.g.f. The about mean will be given by

$M_m(t) = e^{-mt} M_0(t)$

$$_m(t) = e^{m\left(e^{t} - 1 - t\right)}$$

Example-11

Show that if X_1 and X_2 be two independent random variables with Poisson distribution with parameters m_1 and m_2, respectively, then the sum (X_1+X_2) is a random variable with Poisson distribution with parameter (m_1+m_2).

Solution

If $M_1(t)$ and $M_2(t)$ be the m.g.f. of X_1 and X_2, then

$M_1(t) = e^{m_1(e^t - 1)}$ and $M_2(t) = e^{m_2(e^t - 1)}$

also, we know that the m.g.f of the sum of several independent variables is the product of the m.g.f.

m.g.f. of $(X_1 + X_2)$, where X_1 , X_2 are independent variables

= product of m.g.f of X_1 and X_2

$= M_1(t) \times M_2(t)$

$= e^{m_1(e^t - 1)} \times e^{m_2(e^t - 1)}$

$= e^{(m_1+m_2)(e^t-1)}$

= m.g.f. of Poisson distribution with parameter (m_1+m_2).

Hence proved

m.g.f. of Normal distribution

In the case of normal distribution, we know the probability function with the mean at the origin is given by

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\left(\frac{x^2}{2\sigma^2}\right)}$$

$$M_0(t) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{tx} e^{-\left(\frac{x^2}{2\sigma^2}\right)} dx$$

$$M_0(t) = e^{mt+(1/6)t^2\sigma^2}$$

m.g.f. with respect to mean m

$M_m(t) = e^{-mt} M_0(t)$

$$= e^{-mt}\{ e^{mt+(1/2)t^2\sigma^2} \}$$

$$M_m(t) = e^{\frac{t^2\sigma^2}{2}}$$

Example -12

Prove that if X_1 and X_2 are independent normal variable with means m_1 and m_2 and variance $\sigma_1^2 \text{and} \sigma_2^2$ respectively, then the variable (X_1+X_2) is also a normal variable with mean (m_1+m_2)

and variance $\sigma_1^2 + \sigma_2^2$.

Solution

If $M_1(t)$ and $M_2(t)$ be the m.g.f. of X_1 and X_2 about origin, then we have

$$M_1(t) = e^{m_1 t + \frac{1}{2}t^2\sigma_1^2} \text{ and } M_2(t) = e^{m_2 t + \frac{1}{2}t^2\sigma_2^2}$$

also we know that the m.g.f of the sum of a number of independent variates is the product of the m.g.f.

m.g.f. of (X_1+X_2), where X_1 and X_2 are independent variates

= product of m.g.f. of X_1 and X_2

$= M_1(t) \times M_2(t)$

$$= e^{m_1 t + \frac{1}{2}t^2\sigma_1^2} \times e^{m_2 t + \frac{1}{2}t^2\sigma_2^2}$$

$$= e^{(m_1+m_2)(t) + \frac{1}{2}t^2\left(\sigma_1^2+\sigma_2^2\right)}$$

= m.g.f. of normal variate with mean (m_1+m_2) and variance $\sigma_1^2 + \sigma_2^2$.

Function of random variables

Function of random variable situation often arise in systems analysis where knowledge of some characteristic of the system, together with knowledge of the input, will allow some estimate of the behavior at the output. For example, the input random variable X and its density f(x) are known, and the input output behavior is characterized by:

$$Y = \phi(x)$$

We are interested in computing the density of the random variable Y. Note that for a given random

variable X and a function ϕ, Y may not satisfy the definition of a random variable. But if we assume that ϕ is continuous, then Y = $\phi(x)$ will be a random variable.

Example -13

Let Y = $\phi(x) = X^2$. As an example, X could denote the measurement error in a certain physical experiment, and Y would then be the square of the error.

Note that $F_Y(y) = 0$ for $y \leq 0$, for $y > 0$

$$F_Y(y) = P(Y \leq y)$$

$$= P(X^2 \leq y)$$

$$= P(-\sqrt{y} \leq X \leq \sqrt{y})$$

$$= F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

And by differentiation, the density of Y is:

$$f_Y(y) = \frac{1}{2\sqrt{y}}\left[f_X\left(\sqrt{y}\right) + f_X\left(-\sqrt{y}\right)\right], \quad y>0$$

$$= 0$$

Example -14

Let X be uniformly distributed on (0, 1). We show that $Y = -\lambda^{-1}\text{In}(1 - X)$ has an exponential distribution with parameter $\lambda > 0$.

Observe that Y is a nonnegative random variable implying $F_y(y) = 0$ for $y \leq 0$ for $y > 0$, we have:

$$F_y(y) = P(Y \leq y) = P[-\lambda^{-1}\text{In}(1 - X) \leq y]$$

$$= P[\text{In}(1 - X) \geq -\lambda y]$$

$$= P[(1 - X) \geq e^{-\lambda y}]$$

Since e^x is an increasing function of x,

$$= P(X \leq 1 - e^{-\lambda y})$$

$$= F_x(1 - e^{-\lambda y})$$

But since X is uniform over (0, 1), $F_x(x) = x$, $0 \leq x \leq 1$, thus:

$F_y(y) = 1 - e^{-\lambda y}$

Therefore, Y is exponentially distributed with parameter λ.

Sampling theory

If data is collected only from a part of the population i.e., only from some units of the population, it is called sampling. We have under consideration a specified population of objects (people, manufactured items, etc) about which we want to make some inference without looking at every single object. Thus, we sample, i.e. we try to consider some typical objects from which we hope to extract some information that is some sense is characteristic of the entire population.

Suppose that we label each member of the finite population with a number, consecutively, say, so that without loss of generality, a population consisting of N objects may be represented as 1, 2,…, N. Now choose n items in a way to will be described below. Define the following random variables.

X_i = population value obtained when the i^{th} item is chosen i = 1,2…,n

The probability distribution of the random variables X_1, X_2, …,X_n depends on how we go about sampling. If we sample with replacement, each time choosing an object at random, the random variables X_1,X_2,…, X_n are independent and identically distributed. That is for each X_i, i=1,2, ..., n we have

$P(X_i=j) = 1/N$, j = 1,2…N

Instead, their joint probability distribution is given by

$$P[X_i=j_1,\ldots,X_n=j_n]= \frac{1}{N(N-1)(N-2)..(N-n+1)}$$

Where $j_1,\ldots j_n$ are any n values from (1,2…N)

Method of drawing a sample

The following are some of the methods to draw a sample:

1. Simple random sampling
2. Stratified random sampling

Point estimation

If X is a random variable with probability distribution f(x), characterized by the unknown parameter θ, and if $X_1,X_2,\ldots X_n$ is a random sample of size n from X, the statistic $\hat{\theta}=h(X_1,X_2,\ldots X_n)$ is called a point estimator of θ. Note that $\hat{\theta}$ is a random variable because it is a function of the random variable. After the sample has been selected, $\hat{\theta}$ takes on a particular numerical value $\hat{\theta}$ called the point estimate of θ.

In general, A point estimate of some population parameter θ is a numerical value $\hat{\theta}$ of a statistic $\hat{\theta}$. The statistic $\hat{\theta}$ is called the point estimator.

Unbiased Estimators

The point estimator $\hat{\theta}$ is an unbiased estimator for the parameter θ if

$E(\hat{\theta}) = \theta$

If the estimator is not unbiased, then the difference $E(\hat{\theta}) - \theta$

is called the bias of the estimator $\hat{\theta}$.

When an estimator is unbiased, the bias is zero; i.e. $E(\hat{\theta}) - \theta = 0$

Example -15

Suppose that X is a random variable with mean μ and variance σ^2. Let

$X_1, X_2, \ldots X_n$ be a random sample of size n from the population represented by X. show that the sample mean $\overline{X}$ and the sample variance S^2 are unbiased estimators of μ and σ^2, respectively.

Solution

First, consider the sample mean. We know $E(\overline{X}) = \mu$

Therefore, the sample mean $\overline{X}$ is an unbiased estimator of the population mean μ.

Now consider the sample variance. We have

$$E(S^2) = E\left[\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}\right] = \frac{1}{n-1}E\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

$$= \frac{1}{n-1}E\sum_{i=1}^{n}\left(X_i^2 + \overline{X^2} - 2\overline{X}X_i\right)$$

$$= \frac{1}{n-1}E\left(\sum_{i=1}^{n}X_i^2 - n\overline{X^2}\right)$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} E\left(X_i^2\right) - nE\left(\overline{X^2}\right)\right]$$

Since $E(X_i^2) = \mu^2 + \sigma^2$ and $E(\overline{X}^2) = \mu^2 + \sigma^2/n$, we have

$$E(S^2) = \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(\mu^2 + \sigma^2\right) - n\left(\mu^2 + \sigma^2/n\right)\right]$$

$$= \frac{1}{n-1}\left(n\mu^2 + n\sigma^2 - n\mu^2 - \sigma^2\right)$$

$$= \sigma^2$$

Therefore, the sample variance S^2 is an unbiased estimator of the population variance σ^2.

Variance of a Point Estimator

If we consider all unbiased estimator of θ, the one with the smallest variance is called the minimum variance unbiased estimator (MVUE)

If $X_1, X_2, \ldots, Xn$ is a random sample of size n from a normal distribution with mean μ and variance σ^2, the sample mean $\overline{X}$ is the MVUE for μ.

8.6 SUMMARY

In this unit, we studied that a statistical hypothesis test is a method of making decisions using data, whether from a controlled experiment or an observational study (not controlled). In statistics, a result is called statistically significant if it is unlikely to have occurred by chance alone, according to a pre-determined threshold probability, the significance level. The phrase "test of significance" was coined by Ronald Fisher: "Critical tests of this kind may be called tests of significance, and

when such tests are available, we may discover whether a second sample is or is not significantly different from the first."

Hypothesis testing is sometimes called confirmatory data analysis, in contrast to exploratory data analysis. In frequency probability, these decisions are almost always made using null-hypothesis tests. These are tests that answer the question, assuming that the null hypothesis is true, what is the probability of observing a value for the test statistic that is at least as extreme as the value that was observed? More formally, they represent answers to the question, posed before undertaking an experiment, of what outcomes of the experiment would lead to rejection of the null hypothesis for a pre-specified probability of an incorrect rejection. One use of hypothesis testing is deciding whether experimental results contain enough information to cast doubt on conventional wisdom. The Bayesian approach to hypothesis testing is to base rejection of the hypothesis on the posterior probability. Other approaches to reaching a decision based on data are available via decision theory and optimal decisions. The *critical region* of a hypothesis test is the set of all outcomes that cause the null hypothesis to be rejected in favor of the alternative hypothesis. The critical region is usually denoted by the letter *C*.

8.7 GLOSSARY

- **Estimator:** the values in the given sample of attributes which is used for finding the required mean
- **Upper limit:** the upper value in the given population

8.8 CHECK YOUR PROGRESS

1. Thompson Press hypotheses that the average life of its latest web–offset press is 14,500 hours. They know the SD of the press life is 2,100 hours. From a sample of 25 presses, the company finds a sample mean of 13,000 hours. At a 0.01 significance level, should the company conclude that the average life of the presses is less than the hypothesized 14,500 hours?

2. Theatre owners in India know that a hit movie ran for an average of 84 days with a standard deviation of 10 days in each city where the movie was screened. A particular movie distributor

was interested in comparing the popularity of movie in his region with that of the population. He randomly chose 75 theatres at random in the region and found a popular movie ran for 81.5 days.

a. State appropriate hypotheses for testing whether there was a significant difference between theatres in the distributor's region and the population.

b. At a 1% significance level, test these hypotheses.

8.9 ANSWERS TO CHECK YOUR PROGRESS

1. Here, the hypotheses can be written as follows:

H_o: $\mu = 14500$ and H_1: $\mu < 14500$ and significance level $\alpha = 0.01$

For $\alpha = 0.01$ the lower region of acceptance, region $z = -2.33$ or

(For a probability of 0.4901, the value of z is 2.33 from the tables. Acceptance under left region of the tail).

$$\bar{x} = \frac{\mu H \circ -2\sigma}{\sqrt{n}} = \frac{145000 - 2.33(2100)}{5} = 13521.4 \text{ hours}$$

Since the sample mean $\bar{x}$ is far less than the hypothesised value, the null hypothesis is rejected.

2. Given the following data:

$\sigma = 10$ days, $n = 75$ theatres $\bar{x} = 81.5$

Ho: $\mu = 84$ days H_1: $\mu \neq 84$ days $\alpha = 0.01$

This means the rejection area under both the tails is .01, and the area of the acceptable region is 0.99. There for are of, half of the acceptable region is $\frac{0.99}{2}$ = .4950 means a z value of 2.58. therefore, the limits of the acceptable region are :

z = ± 2.58 or $\bar{x} = \mu_{Ho} \pm \frac{z\sigma}{\sqrt{n}} = 84 \pm 2.58 x \frac{10}{\sqrt{75}}$

= 81.02 lower limit and 86.98 as upper limit.

Because the observer value is in the acceptance region, we do not reject the null hypothesis H_0. The length of the run of the movie is the same as the other theatres.

Or in another way:

The observed z value is $\bar{x} - \frac{\mu_{H_\circ}}{SE}$ where SE = $\frac{\sigma}{\sqrt{n}} = \frac{81.4 - 84}{(1.155)}$

= – 2.17. The acceptable z region is ± z = ± 2.58

8.10 TERMINAL QUESTIONS

1. What do you mean by the test of differences in proportion?

2. Write a note on the Moment generating function

3. What do you understand by standard error?

8.11 SUGGESTED READINGS

1. Basic Statistics – Goon, Guptha and Dasgupta – World Press Limited – Calcutta.

2. Fundamentals of Business Statistics – Sanchethi and Kappor.

3. Quantitative Methods in Management – Srivastava, Shenoy and Guptha.

4. Business Statistics – Guptha and Guptha.

UNIT 9: SIGNIFICANCE TEST IN VARIABLES (LARGE SAMPLES)

Structure

OBJECTIVES

At the end of this unit, you will be able to

- Apply a significance test for large samples
- Analyze a given data by significance test

9.1 INTRODUCTION

A test of hypothesis is based on the sampling distribution of the test statistic. And so, to define the critical region, the sampling distribution must be known. Here, the sampling distribution may be one of those that are discussed in earlier units or it may be one different from these. For small samples $n > 30$ the test would be based on the respective sampling distribution. However, for large samples $n \geq 30$ most of the sampling distributions tend to normality, and so, the test may

be based on normal distribution. Let us consider some large sample tests which are based on normal distribution.

9.2 TESTS FOR MEAN

Suppose the mean μ of a population is not known. We want to test whether the men is a given value μ_0

The null hypothesis is –

$H_0 : \mu = \mu_0$

For a large random sample of size

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \text{ is } N(0,1) \text{ and so, the test staristic is } Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

The alternative hypothesis may be any one of the following:

1. $H_1 : \mu = \mu_0$ *Here, the test is two – tailed*
2. $H_1 : \mu > \mu_0$ *Here, the test is one – tailed with critical region at the upper tail.*
3. $H_1 : \mu < \mu_0$ *Here, the test is one – tailed with critical region at the lower tail.*

4. Testing *for independence of attributes in acontigency table.*

The alternative hypothesis may be any one of the following:

1. $H_1 : \mu_1 \neq \mu_2$ *Here, the test is two – tailed*
2. $H_1 : \mu_1 > \mu_2$. *Here, the test is one – tailed with critical region at the upper tail.*
3. $H_1 : \mu < \mu_2$ *Here, the test is one – tailed with critical region at the lower tail.*

In the case of a two-tailed test, if α is the level of significance, the critical values are $-k_{\alpha/2}$ *and* $k_{\alpha/2}$. In the case of upper upper-tailed test, the critical value is k_α. In the case of lower lower-tailed test, it is $-k_\alpha$

Note: Here, if α_1 and α_2 are not known, the test statistic is

$$Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^{\,2}}{n_1} + \frac{S_{21}^{\,2}}{n_2}}} \quad where$$

S_1 *and* S_2 *aresample s*tan*dard deviations.*

Test for Proportion

Suppose the proportion of an attribute in a population is not known. We want to test whether the proportion is a given value P_0

The null hypothesis is –

$H_0 : P = P_0$ (*Populaton is* P_0)

The alternative hypotheses is may be any one of the following.

1. H_1: $P \neq P_0$ *Here, the test is two – tailed* .
2. H_1: $P > P_0$. *Here, the test is one – tailed with critical region at the upper tail*.
3. H_1: $P < P_0$ *Here, the test is one – tailed with critical region at the lower tail*.

In a large random sample of size of size n from the population, let x units possess the attribute. Then, the sample proportion is $p = \frac{x}{n}$

And so, under $H_0, Z = \frac{P - P_0}{\sqrt{\frac{P_0 Q_0}{n}}}$ *is* N (0,1)

and it is the test statistic.

In the case of a two-tailed test, if α is the level of significance, the critical values are -$-k_{\alpha/2}$ *and* $k_{\alpha/2}$. In the case of lower lower-tailed test, it is $-k_\alpha$

Test for Equality of Proportions

Suppose there are two populations with unknown proportions P_1 and P_2 of a certain attribute. We wish to test whether the proportions are equal.

The null hypothesis is –

$H_0 : P_1 = P_2$ (*Population proportions are equal*).

The alternative hypothesis may be any one of the following.

1. $H_1 : P_1 \neq P_2$ *Here, the test is two – tailed*
2. $H_1 : P_1 > P_2$. *Here, the test is one – tailed with critical region at the upper tail.*
3. $H_1 : P_1 > P_2$ *Here, the test is one – tailed with critical region at the lower tail.*

Under H_0 let P is the common proportion. Let a large random sample of size n_1 be drawn from the first population. Among these n_1 units, let x_1 units possess the attribute, so that the sample proportion is $P_1 = \frac{\overline{x_1}}{n_1}$ Also, let a large random sample of size n_2 be drawn from the second population. Among these n_2 units, let x_2 units possess the attribute, so that the sample proportion is $P_2 = \frac{\overline{x_2}}{n_2}$

Then, under $H_0, Z = \frac{P_1 - P_2}{PQ\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$ is N(0,1)

and it is the test statistic,

Generally, the common proportion P will not be known. And so, it is estimated from the samples.

The estimate is $\hat{P} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

Thus, the test statistic is $Z = \dfrac{P_1 - P_2}{\sqrt{\hat{P}\hat{Q}\left[\dfrac{1}{n_1} + \dfrac{1}{n_2}\right]}}$

In the case of a two-tailed test, the critical values are $-K_{\alpha/2}$ and $K_{\alpha/2}$. In the case of upper upper-tailed test, the critical value is $K_{\alpha.}$. In the case of lower lower-tailed test, it is $-K_{\alpha.}$

Example-1

'Mandakini Milks' markets milk in sachets of 500ml. each. Milk is filled in sachets by a machine for which the standard deviation of fillings is 5ml. Three different possible situations are –

a) It is required to verify whether the machine is filling 500ml. milk on average. That is, the machine is set properly.
b) There is a complaint from the customers that the sachets have less than 500ml. milk.
c) The management requires that, on average, the fillings should not exceed 500ml.
d) Suppose one of the above situations arises, and we are required to reach a conclusion. Among the filled sachets. 72 are randomly picked and their contents are measured. The mean of these measurements is found to be 501. 1 ml. What is your conclusion? (Test at 5% level of significance)

Solution:

Here, $\mu_{0.} = 500ml., \sigma = 5\,ml., n = 72, \bar{x} = 501.1ml.\ and\ \alpha = 0.05ml.$

a) It is required to verify whether the population mean μ is 500ml. Or not. And so, the null hypothesis is –

$H_{0.}: \mu \neq 500ml.$ (*The machine is set properly* .)

The Alternative hypothesis is

$H_{1.}: \mu \neq 500ml.$ (*The machine is not properly set*.)

Under H_0, *the testastic* $Z = \dfrac{\bar{x}\mu_{0.}}{\sigma/\sqrt{n}}$ *is* $N(0,1)$

Here, the test is two-tailed because

$H_{1.}: \mu \neq 500ml$

At 5% level of significance, the critical values are $-K_{\alpha/2} = -1.96$ *and* $K_{\alpha/2} = 1.96$ For the given sample, the desired value of Z *is*

$$Z_{obs} = \frac{\bar{x}\mu_{0.}}{\sigma/\sqrt{n}} = \frac{501.1 - 500}{5/\sqrt{72}} = 1.87$$

Since $Z_{obs} = 1.87$ *is a valuein the* int *erval* $(-1.96, 1.96)$ H_0 *is accepted*.

Conclusion: The mean is 500 ml. That is, the machine is set properly.

b) It is required to verify whether the population mean is less than 500ml. And so, the null hypothesis is –

$H_{0.}: \mu = 500ml.$ (*The mean is* 500 *ml.*)

The alternative hypothesis is –

$H_{1.}: \mu < 500ml.$(*The mean is less than* 500 *ml*)

Under $H_{0.}$, *the test statistic* $Z = \frac{\bar{x}\mu_{0.}}{\sigma/\sqrt{n}}$ *is* $N(0,1)$

Here, the test is lower-tailed because $H_{1.}: \mu < 500ml.$

At a 5% level of significance, the critical value is $-K_{\alpha/2} = -1.645$ and the critical region is $Z = -1.645$

$$Z_{obs} = \frac{\bar{x}\mu_{0.}}{\sigma/\sqrt{n}} = \frac{501.1 - 500}{5/\sqrt{72}} = 1.87$$

$Z_{obs} = 1.87 > -1.645$, H_0 *is accepted*.

Conclusion: The mean is 500 ml. It is not less than 500 ml.

c) It is required to verify whether the population mean is more than 500 ml. And so, the null hypothesis is –

$H_{0.}: \mu < 500ml.$(*The mean is* 500 *ml*)

The alternative hypothesis is –

$H_{1.}: \mu < 500ml.(The\ mean\ is\ greater\ than\ 500\ ml)$

$Under\ H_{0.}, the\ test\ statistic\ Z = \frac{\bar{x}\mu_{0.}}{\sigma/\sqrt{n}}\ is\ N(0,1)$

Here, the test is upper-tailed because $H_{1.}: \mu > 500ml.$

At a 5% level of significance, the critical value is $K_{\alpha} = 1.645$ and the critical region is

$Z > 1.645$

$$Z_{obs} = \frac{\bar{x}\mu_{0.}}{\sigma/\sqrt{n}} = \frac{501.1 - 500}{5/\sqrt{72}} = 1.87$$

$Since\ Z_{obs} = 1.87 > 1.645, H_0\ is\ rejected$

Conclusion: The mean is greater than 500 ml. That is, the average filling is more than 500 ml.

Example-2

A firm manufactures resistors. The standard deviation of their resistance is known to be 0.02 ohms. It is required to test whether their mean resistance is 1.4 ohms. A random sample consisting of 64 resistors has a mean of 1.39 ohms. Based on this sample, can we conclude that the mean resistance of the whole lot is 1.4 ohms?

Solution:

Here, $\mu_0 = 1.4\ ohms, \sigma = 0.02\ ohms, n = 64\ and\ \bar{x} = 1.39\ ohms$

The level of significance is not mentioned. In such cases, we assume it as $\alpha = 5\% - 0.05$. The null hypothesis is –

$H_{0.}: \mu = 1.4\ ohms\ (The\ mean\ resis\tan ce\ is\ 1.4\ ohms)$

The alternative hypothesis is –

$H_{1.}: \mu \neq 1.4 ohms\ (The\ mean\ resis\tan ce\ is\ not\ equal\ to\ 1.4\ ohms)$

$Under\ H_{0.}, the\ test\ statistic\ Z = \frac{\bar{x}\mu_{0.}}{\sigma/\sqrt{n}}\ is\ N(0,1)$

Here, the test is two-tailed.

At a 5% level of significance, the critical values are

$K_{\alpha/2} = -1.96$

$$Z_{obs} = \frac{\bar{x}\mu_{0.}}{\sigma/\sqrt{n}} = \frac{1.39 - 1.4}{0.02/64} = -4$$

Since $Z_{obs} = -4$ *is a value outside the* int *erval* $(-1.96, 1.96)$, H_0 *is rejected*

Conclusion: The Mean resistance of the resistors is not equal to 1.4 ohms.

Exercise-3

It is required to test the hypothesis that, on average, Punjabis are taller than 180 cms. For this, 50 Punjabis are randomly selected and their heights are measured. If the mean height is 181.1 cms. And the standard deviation is 3.3 cms., What is your conclusion? (Use a 1% level of significance.)

Solution:

Here, $\mu_{0.} = 180 cms., n = 50, \bar{x} = 181.1 cms., s = 3.3 cms. and\ \alpha = 0.01$

The null hypothesis is –

$H_{0.} \mu = 180 cms., (Mean\ height\ is\ Punjab\ is\ 180\ cms.)$

The alternative hypothesis is –

$H_{1.} \mu > 180 cms., (Mean\ height\ is\ greaterthan\ 180\ cms.)$

$Under\ H_{0.}, the\ test\ statistic\ Z = \frac{\bar{x}\mu_{0.}}{\sigma/\sqrt{n}}\ is\ N(0,1)$

Here, the test is upper-tailed.

At a 1% level of significance, the critical value is $K_{\alpha} = 2.33$ and the critical region is

$Z > 2.23$

$$Z_{obs} = \frac{\bar{x}\mu_{0.}}{s/\sqrt{n}} = \frac{181.1 - 180}{3.3/\sqrt{50}} = 2.36$$

Since $Z_{obs} = 2.36$ *is greater than* 2.33, H_o *is rejected*

Conclusion: On average, Punjabis are taller than 180 cms.

Exercise-4

From a population with a mean of 836, a random sample containing 225 observations is drawn. The mean and standard deviation for the sample is 840.5 and 45, respectively. At a 1% level of significance, test whether the sample mean differs significantly from the population mean.

Solution:

Here, $\mu_0 = 836, n = 225, \bar{x} = 840.5, s = 45$ *and* $\alpha = 0.01$

The null hypothesis is –

$H_{0.}: \mu = 836$ (*The sample mean does not differ significantly from thepopulation mean*)

The alternative hypothesis is –

$H_{1.}: \mu \neq 836$ (*The sample mean differs significantly from the population mean*)

Under H_0, *the test statistic* $Z = \frac{\overline{x} - \mu_{0.}}{s/\sqrt{n}}$ *is* $N(0,1)$

Here, the test is two-tailed.

At 1% level of significance, the critical values are $-K_{\alpha/2} = -2.58$ *and* $K_{\alpha/2} = 2.58$

$$Z_{obs} = \frac{\bar{x} - \mu_{0.}}{s/\sqrt{n}} = \frac{1.39 - 1.4}{45/\sqrt{50}} = 1.5$$

Since $Z_{obs} = 1.5$ *is a value outside the* int *erval* $(-2.58, 2.58)$, H_0 *is accepted*

Conclusion: The sample mean does not differ significantly from the population mean.

Exercise-5

It is known that IQ of boys has standard deviation 10, and that IQ of girls has standard deviation 12. Mean IQ of 200 randomly selected boys is 99 and the mean IQ of 300 randomly selected girls is 97.

(i) Can we conclude that on average, boys and girls have the same *IQ* ?

(ii) Can we conclude that on average, boys have *IQ* greater than girls?

Solution:

Here, $\sigma_1 = 10, \sigma_2 = 12, n_1 = 200, n_2 = 200, n_2 = 300, \bar{x}_1 = 99$ *and* $\bar{x}_2 = 97$

Since α is not specified, we consider $\alpha = 5\% = 0.05$

(i) The null hypothesis is –

$H_0 : \mu_1 = \mu_2$ (*Boys and girls have the same IQ*)

The alternative hypothesis is –

$H_1 = U_1 \neq U_2$ (*Boys and girls have different IQ*)

$$\text{Under } H_0 \text{ the test statistic } Z = \frac{\bar{x}_1 - x_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \text{ is } N(0,1)$$

Here, the test is two-tailed.

At 5% level of significance, the critical values are $-K_{\alpha/2} = -1.96$

And $K_{\alpha/2} = 1.96$

$$Z_{obs} = \frac{\bar{x}_1 - x_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{599 - 97}{\sqrt{\frac{10^2}{200} + \frac{12^2}{3200}}} = 1.87$$

Since $Z_{obs} = 2.02$ *is a value outside the* int *erval* $(-1.96, 1.96)$, H_0 *is rejected*

Conclusion: Boys and girls have different *IQ*

(ii) Here, the alternative hypothesis is –

$H_1 : \mu_1 > \mu_2$ (*Boys have IQ greater than that of girls*.)

Here, the test is upper-tailed.

At a 5% level of significance, the critical value is $K_\alpha = 1.645$

Since $Z_{obs} = 2.02$ *is greater that* 1.645, H_0 *is rejected.*

Conclusion: Boys have IQ greater than that of girls.

Exercise-6

A random sample of 1000 apples from an orchard has a mean weight 187gms. and standard deviation 8 gms. A random sample of 800 apples from another orchard has a mean weight 188.4gms. and standard deviation 10 gms. Test the hypothesis that the mean weights of apples of the two orchards are the same.

Solution:

Here,

$$n_1 = 1000 \qquad \overline{x_1} = 187 gms, \qquad s_1 = 8 gms.$$

$$n_2 = 800 \qquad \overline{x_2} = 188.4 gms, \qquad s_2 = 10 gms.$$

The null hypothesis is –

$H_0 = \mu_1 = \mu_2$ (means are equal)

The alternative hypothesis is –

$H_1 = \mu_1 \neq \mu_2$ (means are not equal)

Under H_0 *thetestatistic* $Z = \dfrac{\overline{x_1} - x_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ *is* $N(0.1)$

The test is two-tailed.

At a 5% level of significance, the critical values are $-K_{\alpha/2} = 1.96$ and $K_{\alpha/2} = 1.96$

$$Z_{obs} = \frac{\overline{x_1} - x_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \frac{187 - 188.4 = 1.87}{\sqrt{\dfrac{8^2}{1000} + \dfrac{10^2}{800}} = -3.22}$$

Since $Z_{obs} = -3.22$ *is a value outside the*int *erval* $(-1.96, 1.96)$, H_0 *is rejected.*

Conclusion: The Mean weights of apples of the two orchards are not the same.

Exercise-7

A study of the systolic blood pressure of a randomly selected group of 36 patients suffering from a disease and another group of 36 persons who do not suffer from the disease gave the following results.

	Suffering	Not suffering
Sample size	36	36
Mean systolic pressure	178	141
Standard deviation	24	12

Test whether the average systolic pressure of the patients suffering from the disease is greater than that of those who do not suffer. Conduct the test at 1% level of significance.

Solution:

Here, $n_1 = 36$ $\overline{x_1} = 178$ $s_1 = 24$

$n_2 = 36$ $\overline{x_2} = 141$ $s_1 = 12 \; and \; \alpha = 0.01$

The null hypothesis is –

$H_0 : \mu_1 = \mu_2$ (Average systolic pressure of the two groups is the same).

The alternative hypothesis is –

$H_1 : \mu_1 > \mu_2$ (Average systolic pressure of those who suffer from the disease is greater than the average systolic pressure of those who do not suffer.

Under H_0, $the test statistic Z = Z = \dfrac{\overline{x_1} - x_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \; is \; N(0,1)$

Here, the test is upper-tailed.

At a 1% level of significance, the critical value is $K_\alpha = 2.33$

$$Z_{obs} = \frac{\overline{x_1} - x_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{178 - 141}{\sqrt{\frac{24^2}{36} + \frac{12^2}{36}}} = -8.27$$

Since $Z_{obs} = 8.27$ *isgreatertr than* 2.33, H_0 *is rejected*.

Conclusion: The Average systolic pressure of those who suffer from the disease is greater than the average systolic pressure of those who do not suffer.

Exercise-8

An intelligence test on two groups of boys and girls gave the following results.

	Mean marks,	Standard deviation,	Sample size
Boys	70	20	250
Girls	75	15	150

Can we conclude at a 1% level of significance that the mean marks of girls are more than those of boys?

Solution:

Here,

$n_1 = 250,$ $\overline{x_1} = 70,$ $s_1 = 20.$

$n_2 = 150$ $\overline{x_2} = 75$ $s_2 = 15$ *and* $\alpha = 1\% = 0.01$

The null hypothesis is –

$H_0 = \mu_1 = \mu_2$ (means are equal)

The alternative hypothesis is –

$H_1 = \mu_1 \neq \mu_2$ (The mean marks of boys are less than the mean marks of girls)

Under H_0, *thetest statistic* $Z = \dfrac{\overline{x_1} - x_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ *is* $N(0.1)$

The test is two-tailed.

At 1% level of significance, the critical values is $-K_{\alpha/2} = -2.33$ and $K_{\alpha/2} = -2.33$

$$Z_{obs} = \frac{\overline{x_1} - x_2}{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}} = \frac{70 - 75}{\sqrt{\dfrac{20^2}{250} + \dfrac{15^2}{150}}} = -2.84$$

Since $Z_{obs} = -2.84$ *is lessthan* -2.33, H_0 *is rejected.*

Conclusion: The Mean marks of boys are less than those of girls. (The mean marks of girls are more than those of boys.)

Exercise-9

Certain dose of an analgesic (pain reliever) when administered to each of 32 women patients, the average duration of pain relief was 3.5 hours. The same dose, when administered to 36 men patients, an average duration of pain relief was 4 hours. By past experience, it is known that the standard deviation of the duration of pain relief is 0.5 hours. Test whether, on average, the duration of pain relief is the same among men and women.

Solution:

Here, $\sigma_1 = \sigma_2 = 0.5 hrs., n_1 = 32,\ \overline{x}_1 = 3.5\ hrs$

$n_2 = 36\ and\ \overline{x_2} = 4 hrs$

Since α is not specified, we consider $\alpha = 5\% = 0.05$

(i) The null hypothesis is –

$H_0 : \mu_1 = \mu_2$ (*Boys and girls have the same IQ*)

The null hypothesis is -

$H_0 = \mu_1 = \mu_2$ (*Average duration of pain relief is the same among men and women*)

The alternative hypothesis is –

$H_1 : \mu_1 \neq \mu_2 n$(*Average duration of pain releif is the same among men and women*).

Under H_0 *the test statistic* $Z = \dfrac{\overline{x_1} - x_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$ *is* $N(0,1)$

Here, the test is two-tailed.

At 5% level of significance, the critical values are $-K_{\alpha/2} = -1.96$

And $K_{\alpha/2} = 1.96$

$$Z_{obs} = \frac{\overline{x_1} - x_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \frac{3.5 - 4}{\sqrt{\dfrac{(0.5)^2}{32} + \dfrac{(0.5)^2}{36}}} = -33.88$$

Since $Z_{obs} = 33.88$ *is a value outside the* int *erval* $(-1.96, 1.96)$, H_0 *is rejected*

Conclusion: Average duration of pain relief is not the same among men and women.

Exercise-10

The manufactures of Brand R pens contend that the proportion of college students of Bangalore who use R pens is more than 0.3. In order to test this contention, 40 college students were randomly picked and questioned in this regard. Among these 40 students, 10 were found to use Brand R pens. At a 0.05 level of significance, test whether the manufacturer's contention is acceptable.

Solution:

Here, $P_0 = 0.3$, $n = 40$, $x = 10$ *and* $\alpha = 0.05$

And so, the sample proportion is

$$p = \frac{x}{n} = \frac{10}{40} = 0.25$$

The null hypothesis is –

$H_0 = P = 0.3$ (The proportion of users of Brand R pens is 0.3.)

The alternative hypothesis is –

$H_0 = P > 0.3$ (*The proportion of users of Brand R pens is o*.3.)

$$\textit{Under } H_0 \textit{ the test statistic } Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n_1}}} \textit{ is } N(0,1)$$

Here, the test is upper-tailed because

$H_1 = P > 0.3$

At 5% level of significance, the critical value is $K_\alpha = 1.645$

$$Z_{obs} = \frac{p - P_0}{\frac{P_0\ Q_0}{n}} = \frac{0.25 - 0.3}{\sqrt{\frac{0.3x0.7}{40}}} = -0.69$$

$Z_{obs} = -0.69$ *is less than* 1.645, H_0 *is accepted*

Conclusion: The proportion of users of Brand R pens is 0.3, and it is not greater than 0.3.

Exercise-11

The author of this book opines that more than 90% Statistics students of PUC of Karnataka refer to the book 'Statistics by Rajmohan'. A survey of 225 randomly picked Statistics students from all over Karnataka revealed that 93.1% of them refer to the book. Can we conclude at a 1% level of significance that the author's opinion is valid?

Solution:

Here, $p_0 = \frac{90}{100} = 0.9,, n = 225, p = \frac{93.1}{100} = 0.931 \; and \; \alpha = 0.01$

The null hypothesis is –

$H_0 : P = 0.9$ (The book is referred to by 90% students).

The alternative hypothesis is –

$H_1 : P > 0.9$ (The book is referred to by more than 90% students).

$Under \; H_0 \; the \; test \; statistic \; Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}} \; is \; N(0,1)$

Here, the test is upper tailed.

At 1% level of significance, the critical value is $K_\alpha = 2.33$

$$Z_{obs} = \frac{p - P_0}{\frac{P_0 \, Q_0}{n}} = \frac{0.931 - 0.9}{\sqrt{\frac{0.9x0.1}{225}}} = -1.55$$

$Z_{obs} = 1.55 is \; is \; less \; than \; 2.33, \; H_0 \; is \; accepted$

Conclusion: The book is referred to by 90% of the students and not by more than 90% students. Thus, the author's opinion is not valid.

EXERCISE -12

It is required to verify whether a coin is biased. The coin is tossed 32 times, and the results are noted. 19 of the 32 tosses resulted in the occurrence of a head. Can we conclude that the coin is biased?

Solution:

We know that the probability of a head for an unbiased coin is 0.5. Therefore, we test whether $P = 0.5$

Thus, $P_0 = 0.5,\ n = 32,\ and\ x = 19$

Therefore, $p = \frac{x}{n} = \frac{19}{32} = 0.5938$

The null hypothesis is –

$H_0 : P = 0.5 (The\ coin\ is\ un\ biased).$

The alternative hypothesis is –

$H_1 : P \neq 0.5 (The\ coin\ is\ un\ biased).$

$Under\ H_0\ the\ test\ statistic\ Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}}\ is\ N(0,1)$

Here, the test is two-tailed.

At 5% level of significance, the critical values are $-K_{\alpha/2} = -1.96$

And $K_{\alpha/2} = 1.96$

$$Z_{obs} = Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}} = \frac{0.5938 - 0.5}{\frac{0.5 \times 0.5}{32}} = 1.06$$

$Since\ Z_{obs} = 1.06\ \ is\ a\ value\ in\ the\ interval\ (-1.96, 1.96),\ we\ \ is\ accepted\ H_0$

Conclusion: The coin is unbiased.

Exercise-13

According to Probability theory, the probability that in a family that has two children, both children are sons is 0.25. In a locality, among 136 families that have 2 children each, 46 families have 2 sons. Does this information support the theory? (Test at 1% level.)

Solution:

Here, $P_0 = 0.25,\ n = 136,\ \ x = 46\ and\ \alpha = 0.01$

Therefore, $p = \frac{x}{n} = \frac{46}{136} = 0.3382.$

The null hypothesis is –

$$H_0 : P = 0.25 (\textit{The proportion of families with } 2 \textit{ sons is } 0.25).$$

The alternative hypothesis is –

$$H_1 : P \neq 0.25 (\textit{The proportion differes from } 0.25).$$

$$\textit{Under } H_0 \textit{ the test statistic } Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}} \textit{ is } N(0,1)$$

Here, the test is two-tailed.

At 1% level of significance, the critical values are $-K_{\alpha/2} = -2.58$

And $K_{\alpha/2} = 2.58$

$$Z_{obs} = Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}} = \frac{0.3382 - 0.25}{\sqrt{\frac{0.25 \times 0.75}{136}}} = 2.375$$

$\textit{Since } Z_{obs} = 2.375 \textit{ is a value in the interval } (-2.58, 2.58), \textit{ we is accept } H_0.$

Conclusion: The proportion of families with 2 sons is 0.25

Exercise -14

It is required to test whether the proportion of smokers among students is less than that among the lecturers. Among 60 randomly picked students, 2 were smokers. Among 17 randomly picked lecturers, 5 were smokers. What would be your conclusion?

Solution:

Here $n_1 = 60,$ $x_1 = 2,$ $p_1 = \frac{x_1}{n_1} = \frac{2}{60} = 0.0333$

$n_2 = 17,$ $x_2 = 5,$ $p_2 = \frac{x_2}{n_2} = \frac{5}{17} = 0.2941$

The null hypothesis is –

$H_0 = p_1 = p_2$ (The proportion of smokers among students is the same as the proportion among lecturers.)

The alternative hypothesis is –

$H_1 = p_1 < p_2$ (The proportion of smokers among students is less than the proportion among lecturers.)

$$Under H_{0,}\ the\ test\ statistic\ Z = \frac{P_1 - P_2}{\sqrt{\hat{P}\hat{Q}\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}\ is\ N(0.1)$$

Here, the test is lower-tailed.

At a 5% level of significance, the critical value is $-K_\alpha = -2.33$

Since the common proportion P is unknown, it is estimated from the given data. The estimate is –

$$\hat{P} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{2 + 5}{60 + 17} = 0.0909$$

$$Z_{obs} = \frac{P_1 - P_2}{\sqrt{\hat{P}\hat{Q}\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} = \frac{0.0333 - 0.2941}{\sqrt{0.0909 \times 0.9091\left[\frac{1}{60} + \frac{1}{17}\right]}}$$

$$\frac{-0.2608}{\sqrt{\frac{0.0909 \times 0.9091}{60 \times 17}}} - 3.302$$

$$Since Z_{obs} = -3.302 is less\ than - 2.33,\ H_0 is\ rejected$$

Conclusion: The proportion of smokers among students is less than the proportion among lecturers.

Exercise-15

Among 326 scooters that crossed the Mysore Bank junction at Bangalore during a period of one hour, 143 were Brand B scooters. Among 213 scooters that crossed the Shivaji Statue junction at Pune during a period of one hour, 137 were Brand B scooters. Test whether the proportion of Brand B scooters on the roads of Bangalore differs from the proportion in Pune.

Solution:

Here $n_1 = 3326,$ $x_1 = 143,$ $p_1 = \frac{x_1}{n_1} = \frac{143}{326} = 0.4387$

$n_2 = 213,$ $x_2 = 137,$ $p_2 = \frac{x_2}{n_2} = \frac{137}{213} = 0.6432$

The null hypothesis is –

$H_0 = p_1 = p_2$ (The proportion of smokers among students is the same as the proportion among lecturers.)

The alternative hypothesis is –

$H_1 = p_1 < p_2$ (The proportion of smokers among students is less than the proportion among lecturers.)

$$Under H_{0,}\ the\ test\ statistic\ Z = \frac{P_1 - P_2}{\sqrt{\hat{P}\hat{Q}\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}\ is\ N(0.1)$$

Here, the test is lower-tailed.

At 5% level of significance, the critical value is $-K_{\alpha} = -1.96 and K_{\alpha/2} = -1.96$

Since the common proportion P is unknown, it is estimated from the given data. The estimate is –

$$Z_{obs} = \frac{P_1 - P_2}{\sqrt{\hat{P}\hat{Q}\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} = \frac{0.4387 - 0.6332}{\sqrt{0.5195 \times 0.4805\left[\frac{1}{326} + \frac{1}{213}\right]}}$$

$Since Z_{obs} = -4.646 is a value outside the interval (-1.96, 1.96). H_0 is rejected$

Conclusion: The proportion of Brand B scooters on the roads of Bangalore differs from the proportion in Pune.

Exercise 16- From the following data, test whether the difference between the proportions in the two samples is significant.

	Size	Proportion
Sample I	1000	0.02
Sample II	1200	0.01

Solution:

Here $n_1 = 1000, n_2 = 1200, P_1 = 0.02 \text{ and } P_2 = 0.01$

$$Therefore, \hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 p_1} = \frac{100 \times 0.02 + 1200 \times 0.01}{1000 + 1200} = 0.0146$$

The null hypothesis is –

$H_0 = p_1 = p_2$ (the proportions are equal)

The alternative hypothesis is –

$H_1 = p_1 \neq p_2$ (The proportion differ)

$$Under H_0, the test statistic Z = \frac{P_1 - P_2}{\sqrt{\hat{P}\hat{Q}\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} is N(0.1)$$

Here, the test is two-tailed.

At 5% level of significance, the critical values are $-K_{\alpha/2} = -1.96 \text{ and } K_{\alpha/2} = 1.96$

$$Z_{obs}=\frac{P_1-P_2}{\sqrt{\hat{P}\hat{Q}\left[\frac{1}{n_1}+\frac{1}{n_2}\right]}}=\frac{0.02-0.01}{\sqrt{0.0146\times 0.9854\left[\frac{1}{1000}+\frac{1}{1200}\right]}}=1.9471$$

$Since Z_{obs}=1.9471 \, lies \; with \, \text{int} \, he \, \text{int} \, erval(-1.96, 1.96). \, H_0 is \; accepted$

Conclusion: The proportions are the same.

Exercise -17 It is required to test whether a coin is biased.

a) Suppose the coin is tossed 40 times and the proportion of tosses resulting in head is 0.4. What is the conclusion?
b) Suppose the coin is tossed 100 times and the proportion of tosses resulting in a head is 0.4. What is the conclusion?

Solution:

In both cases, we have,

$H_0 : P = 0.5$ (The coin is unbiased.)

$H_1 : P \neq 0.5$ (The coin is biased.)

$Under \; H_0 \, the \, test \; statistic \; Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}} \; is \; N(0,1)$

Here, the test is two-tailed.

At a 5% level of significance, the critical values are $-K_{\alpha/2}=-1.96$

and $K_{\alpha/2}=1.96$

a) Here, n=40 and p-0.4

$$Z_{obs} = Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}} = \frac{0..4 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{40}}} = -1.2649$$

$Since\ Z_{obs} = -1.2649\ \ is\ wihtin\ the\ \text{int}\ erval\ (-1.96, 1.96),\quad H_0 is\ accepted$

b) Here, n=100 and p=0.4

Conclusion: The coin is unbiased.

$$Z_{obs} = Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}} = \frac{0.0.4 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{100}}} = -2$$

$Since\ Z_{obs} = -2\ \ is\ outside\ \text{int}\ erval\ (-1.96, 1.96),\ H_0\ \ is\ rejected$

Conclusion: The coin is biased.

Note: In case (a) and case (b), the proportions P are the same (0.4). In spite of this, the decisions are not the same. It is because, for larger n The SE is smaller.

9.3 LARGE SAMPLES

It is very difficult to distinguish between large and small samples. If the sample size is greater than 30, i.e., if n > 30, then those samples may be regarded as large samples. There is a difference between large and small samples in using the test of significance, because the assumptions we make for the two samples are also not the same. The assumptions made for large samples are:

1. The random sample distribution of statistics is approximately normal.
2. Sampling values are sufficiently close to the population value and can be used for the calculation of the standard error of estimate.

In the case of large samples, when we are testing the significance of a statistic, the concept of

standard error is used. The following are the formula for finding out the standard error for different statistics.

The standard error of the mean:

It measures only sampling errors. Sampling errors are involved in estimating a population parameter from a sample, instead of including all the essential information in the population.

i) When the standard deviation of the population is known, the formula is

$$S.E.\,\bar{X} = \frac{\sigma_p}{\sqrt{n}}$$

S.E. $\bar{X}$ = The standard error of the mean

σ_p = Standard deviation of the population

n = Number of observation in the sample

ii) When the standard deviation of the population is not known, we have to use the standard deviation of the sample in calculating the standard error of the mean; the formula is

$$S.E.\,\bar{X} = \frac{\sigma(sample)}{\sqrt{n}}$$

σ = standard deviation of the sample

If the standard deviation of the sample and population are available, then for the calculation of the standard error of the mean, we must use the standard deviation of the population.

Example -1

Calculate the standard error of the mean from the following data, showing the amount paid by 100 firms in Calcutta on the occasion of Durga Pooja.

Mid Value (Rs.):	39	49	59	69	79	89	99
No. of firms:	2	3	11	20	32	25	7

Solution:

$$S.E.\,\bar{X} = \frac{\sigma}{\sqrt{n}}$$

Computation of Standard Deviation

Mid value m	f	$\frac{m-69}{10} = d'$	fd'	fd'^2
39	2	– 3	– 6	18
49	3	– 2	– 6	12
59	11	– 1	– 11	11
69	20	0	0	0
79	32	+ 1	+ 32	32
89	25	+ 2	+ 50	100
99	7	+ 3	+ 21	63
	N = 100		$\Sigma fd' = 80$	$\Sigma fd'^2 = 236$

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times C$$

$$= \sqrt{\frac{236}{100} - \left(\frac{80}{100}\right)^2} \times 10$$

$$= \sqrt{2.36 - 0.64} \times 10 = \sqrt{1.72} \times 10$$

$= 1.311 \times 10 = 13.11$

$$S.E.\,\bar{X} = \frac{13.11}{\sqrt{100}} = \frac{13.11}{10} = 1.311$$

Hypothesis Testing of Proportions: Large Samples

Assumptions:

When doing the hypothesis testing involving proportions, we use the binomial distribution as the sampling distribution. Unless np and nq are both at least 5, we can use the normal distribution as an approximation of the binomial distribution.

Example 2-Mr. X owns a hardware store and sells a particular brand of garden scissors. He wants to compare them with those sold all over the country. He knows from experience that 15% of the scissors sold all over the country require repairs in the first year itself. He sampled 120 customers and found only 22 of them required repairs in the first year of buying them. At a 2% level of significance, is there enough evidence that his scissors sold differ in reliability from those sold all over the country?

Answer:

n = 120, H_o: P = 0.15 H_1: P ≠ 0.15, mean p = $\frac{22}{120}$ = 0.1833

At α = 2% or 0.02, the area of acceptance of the region under both halves is .98, and therefore the area of the acceptance region under one half of the normal curve is 0.4950.

This gives a value of z as 2.33 (critical value)

Therefore, the limits of the acceptance region are ± z or ± 2.33

Observed values of $\pm z = \pm \frac{p - P_{Ho}}{\sqrt{\frac{p_{Ho} q_{Ho}}{n}}} = \frac{0.1833 - 0.15}{\sqrt{\frac{0.15(0.85)}{120}}} = \pm 1.02$

As the observed value 1.02 is less than the critical value 2.33, the null hypothesis is accepted. This means that scissors sold at Mr. X's store is not significantly reliable than the one's sold all over the country.

9.4 SUMMARY

In statistics, a result is called **statistically significant** if it is unlikely to have occurred by chance. The phrase *test of significance* was coined by Ronald Fisher. As used in statistics, *significant* does not mean *important* or *meaningful*, as it does in everyday speech. Research analysts who focus solely on significant results may miss important response patterns that individually may fall under the threshold set for tests of significance. Many researchers urge that tests of significance should always be accompanied by effect-size statistics, which approximate the size and thus the practical importance of the difference.

The amount of evidence required to accept that an event is unlikely to have arisen by chance is known as the **significance level** or critical p-value. In traditional statistical hypothesis testing, the p-value is the probability of observing data at least as extreme as that observed, given that the null hypothesis is true. If the obtained p-value is small, then it can be said that either the null hypothesis is false or an unusual event has occurred. P-values do not have any repeated sampling interpretation.

An alternative (but nevertheless related) statistical hypothesis testing framework is the Neyman–Pearson frequentist school which requires both a null and an alternative hypothesis to be defined and investigates the repeat sampling properties of the procedure, i.e. the probability that a decision to reject the null hypothesis will be made when it is true and should not have been rejected (this is called a "false positive" or Type I error) and the probability that a decision will be made to accept the null hypothesis when it is false (Type II error). Fisherian p-values are philosophically different from Neyman–Pearson Type I errors.

9.5 GLOSSARY

- **Significance level:** The amount of evidence required to accept that an event is unlikely to have arisen by chance
- **Sampling errors:** Sampling errors are involved in estimating a population parameter from a sample

9.6 CHECK YOUR PROGRESS

1. A ketchup manufacturer is in the process of deciding whether to produce a new, extra spicy brand of ketchup. The company's market research team found in a survey of 6000 households that 355 households would buy the extra spicy brand. In an earlier, more extensive study carried out 2 years ago showed that 5% of the households would buy the brand then. At 2% level of significance, should the company conclude that there is an increased interest in the extra spicy flavour?

2. A sample of 1000 students from Bombay University was taken, and their average weight was found to be 112 Ibs with a standard deviation of 20 Ibs. Could the mean weight of students in the population be 120 pounds?

3. A company manufacturing electric light bulbs claims that the average life of its bulbs is 1600 hours. The average life and standard deviation of a random sample of 100 such bulbs were 1570 hours and 120 hours, respectively. Should we accept the claim of the company?

9.7 ANSWERS TO CHECK YOUR PROGRESS

n = 6000, p = 335/6000 = 0.05583

H_o: P = 0.05 H_1: P > 0.05 □ = 0.02

For □ = 0.4800, the upper limit of the acceptance region is 0.4800, which gives a z value of 2.05 from the z table.

$$p = p_{Ho} + z\sqrt{\frac{p_{Ho}\,q_{Ho}}{n}} = 0.05 + 2.05\sqrt{\frac{0.05 \times 0.95}{6000}} = 0.05577$$

Because the observed value P (0.05577) is > than P (0.05), we should just barely reject H_o.

or in another way

As the observed z value is $\dfrac{p - P_{H_\circ}}{\sqrt{\dfrac{p_{Ho}\,q_{Ho}}{n}}} = \dfrac{0.05583 - 0.05}{\sqrt{\dfrac{0.05\,(0.95)}{6000}}} = 2.07$ is greater than the critical value

of z that is 2.05 we barely reject H_o

9.8 TERMINAL QUESTIONS

1. Explain the test for the mean.

2. What do you understand by a large sample?

9.9 SUGGESTED READINGS

1. Basic Statistics – Goon, Guptha and Dasguptha – World Press Limited – Calcutta.
2. Fundamentals of Business Statistics – Sanchethi and Kappor.
3. Quantitative Methods in Management – Srivastava, Shenoy and Guptha.
4. Business Statistics – Guptha and Guptha.

UNIT 10: SIGNIFICANCE TEST IN VARIABLES (SMALL SAMPLES)

Structure

OBJECTIVES

At the end of this unit, you will be able to:

- Apply significance test for small samples
- Analyze a given data among small samples by significance test
- Explain the t test and Chi Square Test

10.1 INTRODUCTION

If the sample size is less than 30 i.e., $n < 30$, then those samples may be regarded as small samples. As a rule, the methods and the theory of small samples apply to large samples; but the methods and the theory of large samples do not apply to small samples. The small samples are used in testing a given hypothesis, to find out the observed values, which could have arisen by sampling fluctuations from some values given in advance, For example, if a sample of 12 gives

a correlation coefficient of + 0.5, we can test whether the value is significant of correlation in the parent population.

In a small sample, the investigator's estimate will vary widely from sample to sample. An inference drawn from a smaller sample result is less precise than an inference drawn from a large sample result.

10.2 STUDENT'S t – DISTRIBUTION

The greatest contribution to the theory of small samples was made by Sir William Gossett and R. A. Fisher. Gossett published his discovery in 1905 under the pen name 'students' and it is popularly known as t-test or student's t-distribution or students' distribution.

When the sample size is 30 or less and the population standard deviation is unknown, we can use the t-distribution.

The formula is

$$t = \frac{\left(\bar{X} - \mu\right)}{\delta} \times \sqrt{n}$$

where $\delta = \sqrt{\frac{\Sigma\left(X - \bar{X}\right)^2}{n-1}}$

under the assumption of a normally distributed population, the t-distribution has been derived mathematically i.e.,

$$f\left(t\right) = C\left(1 + \frac{t^2}{v}\right)^{\frac{-(v+1)}{2}}$$

where $t = \frac{\bar{X} - \mu}{\delta} \times \sqrt{n}$

C = a constant required to make the area under the curve equal to unity.

v = n – 1, the number of degrees of freedom.

Example

The following results are obtained from a sample of 10 boxes of biscuits:

Mean weight of contents = 490 gms.

The standard deviation of the weight = 9 gms.

Could the sample come from a population having a mean of 500 gms.

Solution:

Let us take the hypothesis that μ = 500 gms.

$$t = \frac{\bar{X} - \mu}{\sigma}\sqrt{n}$$

$$\bar{X} = 490 : \mu = 500 ; \sigma = 9 ; n = 10$$

$$t = \frac{490 - 500}{9}\sqrt{10}$$

df =10 – 1 = 9

$$= \frac{10}{9}\sqrt{10}$$

$$= \frac{10}{9} \times 3.16 = \frac{31.6}{9} = 3.51$$

df = 9, $t_{0.01}$ = 3.25

3.51 > 3.25 our hypothesis is rejected.

Example:

An operator claims that he produces 40 articles in an hour. A sample of 10 random hours shows the turns out as 43, 45, 38, 37, 41, 42, 44, 39, 43 and 38. Is the claim of the operator reasonable at a 5% significance level? Assume the distribution of hourly turnout of the operator to be a normal and critical region at 5% level employing a one-tailed test for 9 df to be 1.833.

Solution: Let the null hypothesis be that the average turnout of the operator is 40. This hypothesis is against the alternative hypothesis that turnout is less than 40.

$$\overline{X} = \frac{1}{10}\left(43+45+38+37+41+42+44+39+43+38\right) = 41$$

$$\delta^2 = \frac{1}{9}\left(43-41\right)^2 + \left(45-41\right)^2 + \left(38-41\right)^2 + \left(37-41\right)^2 + \left(41-41\right)^2 +$$
$$\left(42-41\right)^2 + \left(44-41\right)^2 + \left(39-41\right)^2 + \left(43-41\right)^2 + \left(38-41\right)^2$$

$$= \frac{1}{9} \times 72 = 8$$

$$\delta = \sqrt{8} = 2.8284$$

$$t = \frac{\sqrt{X} - \mu}{\delta}\sqrt{n} = \frac{41-40}{2.8284} \times \sqrt{10}$$

$$= 1.118 < 1.833$$

Therefore the difference is not significant. Hence it could have arisen because of chance, the hypothesis cannot be rejected. The claim of the operator is tenable.

10.3 t-Tests

A test of hypothesis is based on the sampling distribution of the statistic. And so, in order to define a critical region, it is necessary that the sampling distribution is known.

For a random sample of size n from normal population $N(\mu,\sigma^2)$, the sample mean $\bar{x}$ normally distributed with mean μ standard deviation $\sigma/\sqrt{n}$. Thereore $Z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}=is\ N\ (0,1)$. Also, $t=\frac{\bar{x}-\mu}{s/\sqrt{n-1}}$ is a Student t-vitiate with $n-1$ d.f. For two random samples of size n_1 and n_2 drawn respectively from $N(\mu_1,\sigma^2)$ and

$N(\mu_2,\sigma^2)$ populations, $\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{n_1s_1^2+n_2s_2^2}{n_1+n_2-2}\left[\frac{n_1+n_2}{n_1n_2}\right]}}$ is a student's

t-variate with $(n_1+n_2-2)\ d.f.$ When population standard deviations are not known, the above t variate is used for testing of hypothesis. However, for large samples, the t is approximately normal.

Small Sample Test for Mean

Suppose in a $N(\mu,\sigma^2)$, population, both μ and σ are not known.

We want to test whether the mean is a given value μ_0

The null hypothesis is –

$H_0=\mu=\mu_0$ $(poulation\ mean\ is\ \mu_0)$

For a random sample of size n, Under H_0, the test statistic

$t=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}-1}$ is a student's t-variate with $(n-1)d.f.$

The alternative hypothesis may be any one of the following.

1. $H_1: \mu\neq\mu_0.$ *Here, the test is two−tailed*

2. $H_1: \mu > \mu_0$. *Here, the test is one – tailed with critical region at the upper tail.*

3. $H_1: \mu < \mu_0$. *Here, the test is one – tailed with critical region at the lower tail.*

In the case of a two-tailed test, if α is the level of significance, the critical values are $-t_{\alpha/2}$ and $t_{\alpha/2}$. In the case of the upper-tailed test, the critical value is t_α. In the case of a lower-tailed test, to critical value $-t_\alpha$

The critical values for different α and for different degrees of freedom are obtained from the table of t distribution provided at the end of the book.

Note: This test is based on the assumption that the population is normal.

Small Sample Test for Equality of Means

Suppose there are two populations $N(\mu_1, \sigma^2)$, and $N(\mu_2, \sigma^2)$, with unknown μ_1 and μ_2 and σ^2

We want to test whether the means μ_1 and μ_2 are equal. The null hypothesis is –

$H_0 = \mu_1 = \mu_2$ (Populatio n means are equal)

For a random sample of size n_1 and n_2 from these populations,

under, the test statistic $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left[\dfrac{n_1 + n_2}{n_1 n_2}\right]}}$ is student' s t - variate with

$(n_1 + n_2 - 2)\ d.f.$

The alternative hypothesis may be any one of the following.

1. $H_1: \mu_1 \neq \mu_2$. *Here, the test is two – tailed*

2. $H_1: \mu_1 > \mu_2$. *Here, the test is one – tailed with critical region at the upper tail.*

3. $H_1: \mu_1 < \mu_2$. *Here, the test is one – tailed with critical region at the lower tail.*

In the case of a two-tailed test, if α is the level of significance, the critical values are $-t_{\alpha/2}$ and $t_{\alpha/2}$.

In the case of the upper-tailed test, the critical value is t_α.

In the case of a lower-tailed test, to the critical value $-t_\alpha$

The critical values for different α and for different degrees of freedom are obtained from the table of t - distribution

Note 1: This test is based on the assumptions –

a) The populations are normal
b) The population standard deviations are equal (unknown)

Note 2: In the case of paired observations with n random pairs, the test statistic $t = \dfrac{\bar{d}}{s_d / \sqrt{n-1}}$

is a student's t-variate with $(n-1)$

$d.f$ Here, d is the deviation between the observations in the pair.

TEST FOR EQUALITY OF MEANS WHEN OBSERVATIONS ARE PAIRED
(Paired t-test, dependent samples):

In a bivariate normal population, let the units in the population have two variable characteristics x and y which are $N(\mu_2, \sigma^2{}_2)$, and $N(\mu_1, \sigma^2{}_1)$ respectively. For example,

1. Couple have husband's height (x) and wife's height (y).
2. Patients undergoing 'Yoga treatment' for high B.P have two measurement of B.P. – one before treatment (x) and the other after treatment (y).
3. Students attending coaching classes score marks (x) in a test before coaching and (y) in another test after coaching.

In such situations, suppose we want to test whether the means μ_1 *and* μ_2 are equal. The null hypothesis is –

$H_0 = \mu_1 = \mu_2$ (means are equal)

For n random pairs of observations $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$, *let* $d_1 = x_i - y_i$

be the deviations. Let $\bar{d}$ be the sample mean and s_d be the sample standard deviation of these deviations. Then, under H_0

The test statistic $t = \dfrac{\bar{d}}{s_d / \sqrt{n-1}}$ is a student's t-variate with $(n-1)$ d.f.

The alternative hypothesis may be any one of the following.

1. $H_1 : \mu_1 \neq \mu_2$. *Here, the test is two – tailed*

2. $H_1 : \mu_1 > \mu_2$. *Here, the test is one – tailed with critical region at the upper tail.*

3. $H_1 : \mu_1 < \mu_2$. *Here, the test is one – tailed with critical region at the lower tail.*

In the case of a two-tailed test, if α is the level of significance, the critical values are $-t_{\alpha/2}$ and $t_{\alpha/2}$. In the case of the upper-tailed test, the critical value is t_α.

In the case of a lower-tailed test, to critical value $-t_\alpha$

The critical values for different α and for different degrees of freedom are obtained from the table of t - distribution

Exercise:

On eight random days, the time taken by a city bus to reach the college is noted below. Test the hypothesis that the mean time for the bus to reach the college is 30 minutes.

Day	:	1	2	3	4	5	6	7	8
Time (minutes)	:	27	34	30	35	31	30	29	32

Solution:

$Here \mu_0 = 30\ Minutes\ and\ n = 8$

The sample is small and α is not known. Assuming normality, we use student's t-test.

The null hypothesis is –

$H_0 : \mu = 30\ (\ mean\ time\ \ is\ 30\ \text{min}\ utes)$

Under H_0, the test statistic $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n-1}}$ is a student's

t-variate with $(n-1) = (8-1) = 7\ d.f.$.

Here, the test is two-tailed.

At 5% level of significance, for 7 d.f., the critical values are

$-t_{\alpha/2} = -2.37$ and $t_{\alpha/2} = 2.37$

$Time(x)$	x^2
27	729
34	1156
30	900
35	1225
31	961
30	900
29	841
32	1024
248	7736

$$\bar{x} = \frac{\sum x}{n} = \frac{248}{8} = 31$$

$$s = \sqrt{\frac{\sum x^2}{n} - \left[\frac{\sum x}{n}\right]^2} = \sqrt{\frac{7736}{8} - \left[\frac{248}{8}\right]^2} = 2.4495$$

$$t_{0bs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n-1}} = \frac{31-30}{2.4495/\sqrt{8-1}} = 1.08$$

Since $t_{0bs} = 1.08$ is within the interval (-2.37, 2.37), is accepted.

Conclusion: The mean time for the bus to reach the college is 30 minutes.

Exercise:

The management of a factory contends that the mean sound intensity in the factory is less than 120 decibels. 23 random measurements have 117 decibels and a standard deviation of 8 decibels. Test at 1% level of significance whether the contention of the management is acceptable.

Solution:

$Here\ \mu_0 = 120\ decibels\ and\ n = 23, \bar{x} = 117\ decibels, s = 8\ decibels\ and\ \alpha = 1\% = 0.01$

Since the sample is small and σ is not known, assuming a normal distribution of sound intensity, we use the student's t-test.

The null hypothesis is –

$H_0 : \mu = 120$ (*mean sound* int *ensity is* 120 *decibles*)

Under H_0, the test statistic $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n-1}}$ is a student's t-variate with $(n-1) = (23-1) = 22\ d.f.$

The alternative hypothesis is –

$H_1 : \mu < 120$ (Mean sound intensity is 120 decibles.)

The test is lower-tailed.

Here, the test is two-tailed.

At a 1% level of significance, for 22 d.f., the critical values are

$$-t_\alpha = -2.51$$

$$t_{0bs} = \frac{\bar{x} - \mu_0}{s / \sqrt{n-1}} = \frac{117-120}{8/\sqrt{23-1}} = -1.76$$

Since t_{0bs} = -1.76 is not less than - 2.51, H_0 is accepted.

Conclusion: Mean sound intensity is 120 decibels and not less.

Exercise:

The weights in grams of hearts of 5 female cats and 8 male cats are given below.

Female cats	:	7.5	7.3	7.1	9.0	7.6			
Male cats	:	12.7	15.6	9.1	12.8	8.3	11.2	9.4	8.2

Test at 1% level of significance that the mean weight of hearts of male cats is more than that of female cats

Solution:

$Here n_1 = 5, n_2 = 8$ *and* $\alpha = 0.01$

The null hypothesis is –

$H_0 : \mu_1 = \mu_2$ (*Mean weights are equal*)

$under H_0$, *the test statistic* $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left[\dfrac{n_1 + n_2}{n_1 n_2}\right]}}$ is student' s t - variate with

student's t-variate with $n_1 + n_2 - 2 = (5 + 8 - 2) = 11\, d.f.$

The alternative hypothesis is –

$H_1 : \mu_1 < \mu_2$ (mean weigh t of hearts of male cats is more than that of female cats).

The test is lower–tailed.

At 1% level of significance, for 11d.f., the critical value is $-t_{\alpha/2} = -2.72 =$ -2.72.

Female cats		Male cats	
x	x^2	x	x^2
7.5	56.25	12.7	161.29
7.3	53.29	15.6	243.36
7.1	50.41	9.1	82.81
9.0	81.00	12.8	163.84
7.6	57.76	8.3	68.89
		11.2	125.44
		9.4	88.36
		8.2	67.24
38.5	298.71	87.3	1001.23

Female cats:

$$\bar{x}_1 = \frac{\sum_1 x}{n_1} = \frac{38.5}{5} = 7.7$$

$$s_1^2 = \frac{\sum_1 x^2}{n_1} - \left[\frac{\sum_1 x}{n_1}\right]^2 = \frac{298.71}{5} - \left[\frac{38.5}{5}\right]^2 = 0.452$$

Male cats:

$$\bar{x}_2 = \frac{\Sigma_2 x}{n_2} = \frac{87.3}{8} = 10.91$$

$$s_1^2 = \frac{\Sigma_2 x^2}{n_2} - \left[\frac{\Sigma_2 x}{n_2}\right]^2 = \frac{1001.23}{8} - \left[\frac{87.3}{8}\right]^2 = 6.12$$

$$t_{0bs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left[\frac{n_1 + n_2}{n_1 n_2}\right]}} = \frac{7.7 - 10.1}{\sqrt{\frac{5 \times 0.452 + 8 \times 6.12}{5 + 8 - 2}\left[\frac{5+8}{5 \times 8}\right]}} = -2.61$$

Since $t_{0bs} = -2.61$ is not less than -2.72, H_0 is accepted.

Conclusion: Mean weight of the hearts of male cats is the same as that of female cats. The evidence is sufficient to conclude that the mean weight of hearts of male cats is more than that of female cats.

Exercise:

The following data regarding heights of randomly selected boys and girls of SSLC class test whether SSLC boys, on average, are taller than SSLC girls.

	Boys	Girls
Sample size	9	12
Mean height (Cms.)	171	169
Standard deviation (Cms	3	2

Solution: *Here* $,n_1 = 9, n_2 = 12, \ \bar{x}_1 = 171, \bar{x}_2 = 169, s_1 = 3$ *and* $s_2 = 2$

The null hypothesis is –

$H_0 : \mu_1 = \mu_2$ (*Mean heights of boys and grils are equal*)

$$under H_0, \text{ the test statistic } t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left[\frac{n_1 + n_2}{n_1 n_2}\right]}} \text{ is student's t - variate with}$$

$(n_1 + n_2 - 2) = (9 + 12 - 2) = 19\ d.f.$

The alternative hypothesis is –

$H_1 : \mu_1 > \mu_2$ (Boys on an average are taller than girls).

The test is upper–tailed.

At 5% level of significance, for 19d.f., the critical value is $t_\alpha = 1.73$

$$t_{0bs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left[\frac{n_1 + n_2}{n_1 n_2}\right]}} = \frac{171 - 169}{\sqrt{\frac{9 \times 3^2 + 12 \times 2^2}{9 + 12 - 2}\left[\frac{9 + 12}{9 \times 12}\right]}} = 1.74$$

Since $t_{0bs} = 1.74$ is greater than 1.73, H_0 is rejected..

Conclusion: Boys, on average, are taller than girls.

Exercise:

The manufacturer of a cattle feed claims that cows fed on his product yield more milk. To substantiate his claim, the following cow experiments were conducted.

a) 6 cows were fed on unusual feed and 8 cows were fed on the manufacturers fed. The 6 cows had a mean milk yield of 9.7 liters and a standard deviation of 1.3 liters. The 8 cows had a mean milk yield of 10.5 liters and a standard deviation of 2.7 liters.

b) 8 cows were fed on the manufacturer's feed. Their milk yield in liters when they were under usual feed and also when they were under the manufacturer's feed are as below.

Usual feed	:	6.3	7.4	9.7	12.4	11.1	10.4	9.6	7.1
Manufacturer's	:	7.4	7.2	14.6	13.6	10.5	11.6	10.4	8.7

In each of the above cases, verify whether there is support for the manufacturer's claim.

Solution:

The case (a), there are two independent samples of cows that are fed on the usual feed and on the manufacturer's feed. Therefore, the t-test for independent samples is applied. In case (b), the same set of cows is observed under the usual feed and under the manufacturer's feed. Here, the data indicate a change in yield in individual cases. Therefore, the paired t-test is applied.

$a)$ *Here* $,n_1=6, n_2=8, \ \overline{x_1}=9.7, \overline{x_2}=10.5, s_1=1.3$ *and* $s_2=2.7$

The null hypothesis is –

$H_0:\mu_1=\mu_2$ (*Mean are equal*)

$$under H_0, \text{ the test statistic } t=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{n_1s_1^2+n_2s_2^2}{n_1+n_2-2}\left[\frac{n_1+n_2}{n_1n_2}\right]}} \text{ is student' s t - variate with}$$

$(n_1+n_2-2)=(6+8-2)=12 d.f.$

The alternative hypothesis is –

$H_1:\mu_1<\mu_2$ (Manufactur er's f*eed* i*s* better)

The test is lower–tailed.

At 5% level of significance, for 12d.f., the critical value is $-t_\alpha=1.78$

$$t_{0bs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left[\frac{n_1 + n_2}{n_1 n_2}\right]}} = \frac{9.7 - 10.5}{\sqrt{\frac{6\times(13.3)^2 + 8\times(2.7)^2}{6+8-2}\left[\frac{6+8}{6\times 8}\right]}} = 0.62$$

Since $t_{0bs} = -0.62$ is not less than -1.78, H_0 is accected..

Conclusion: Means are equal (There is no evidence that the manufacturer's feed is better yielding.)

b) Here, n =8. Let d= x-y where x is the yield on the usual feed and y is the yield on the manufacturer's feed.

Under H_0, $t = \frac{\bar{d}}{s_d / \sqrt{n} - 1}$ is a student's t-variate with $(n-1) = (8-1) = 7\ d.f.$.

The null and alternative hypotheses are the same as in case (a), the test is lower–tailed.

At 5% level of significance, for 7 d.f., the critical value is = $-t_\alpha = -1.90$

x	y	$d = x - y$	d^2
6.3	7.4	-1.1	1.21
7.4	7.2	0.2	0.04
9.7	14.6	-4.9	24.01
12.4	13.6	-1.2	1.44
11.1	10.5	0.6	0.36
10.4	11.6	-1.2	1.44
9.6	10.4	-0.8	0.64
7.1	8.7	-1.6	2.56

-	-	-10.0	31.70

$$\bar{d} = \frac{\sum d}{8} = \frac{-100}{8} = -1.25$$

$$s_d = \sqrt{\frac{\sum d^2}{n} - \left[\frac{\sum d}{n}\right]^2} = \sqrt{\frac{31.70}{8} - \left[\frac{-10}{8}\right]^2} = 1.55$$

$$t_{0bs} = \frac{\bar{d}}{s_d / \sqrt{n-1}} = \frac{-1.25}{1.55/\sqrt{8-1}} = 2.13$$

Since $t_{0bs} = -2.13$ is less than -1.90, H_0 is rejected.

Conclusion: There is support for the manufacturer's claim that cows fed on his feed yield more milk.

Exercise:

There is a coaching class for CET. 10 randomly selected students were given a test before coaching and they also were given a test after coaching. The test scores are as follows.

Before coaching	:	35	39	47	53	27	19	36	46	08	17
After coaching	:	41	37	45	56	31	21	47	41	05	12

Can we conclude that the coaching is effective?

Solution:

Here, the marks before coaching and the marks after coaching can be paired and so a paired t-test is applied.

Let x denote marks before coaching and y denote marks after coaching. Let d=x-y be the deviations.

The null hypothesis is –

$H_0 : \mu_1 = \mu_2$ (*mean are equal*)

Under H_0, the test statistic $t = \dfrac{\bar{d}}{s_d / \sqrt{n-1}}$ is a student's

t-variate with $(n-1) = (10-1) = 9\ d.f.$

The alternative hypothesis is –

$H_1 : \mu_1 < \mu_2$ (Mean has increased, coaching is effefctive .)

The test is lower–tailed.

At 5% level of significance, for 9 d.f., the critical value is $-t_\alpha = 1.83$

x	y	$d = x - y$	d^2
35	41	-6	36
39	37	2	4
47	45	2	4
53	56	-3	9
27	31	-4	16
19	21	-2	4
36	47	-11	121
46	4	5	25
08	05	3	9

17	12	5	25
-	-	-09	253

$$\bar{d} = \frac{\sum d}{n} = \frac{-9}{10} = -0.9$$

$$s_d = \sqrt{\frac{\sum d^2}{n} - \left[\frac{\sum d}{n}\right]^2} = \sqrt{\frac{253}{10} - \left[\frac{-9}{10}\right]^2} = 4.95$$

$$t_{0bs} = \frac{\bar{d}}{s_d / \sqrt{n-1}} = 4.95 \frac{-0.9}{1.55/\sqrt{10-1}} = 0.55$$

Since $t_{0bs} = -0.55$ is not less than -1.83, H_0 is accepted.

Conclusion: Coaching is not effective.

Exercise:

The following are the weights in kgs. Of 7 husbands and their wives.

Couple	:	1	2	3	4	5	6	7
Husband	:	62	56	59	73	49	54	67
Wife	:	55	61	62	68	52	51	62

Test the hypothesis that the mean weight of husbands and the mean weight of wives are equal.

Solution:

Here we have n =7 pairs of observations. And so, we use a paired t-test. Let x be the weight of the wife. Let d=x-y n be the deviations.

The null hypothesis is –

$H_0 : \mu_1 = \mu_2$ (*mean weights are equal*)

Under H_0, the test statistic $t = \dfrac{\bar{d}}{s_d / \sqrt{n-1}}$ is a student's

t-variate with $(n-1) = (7-1) = 6\ d.f.$.

The alternative hypothesis is –

$H_1 : \mu_1 \neq \mu_2$ (Mean are not equal)

The test is lower–tailed.

At a 5% level of significance, for 6 d.f., the critical value is $-t_{\alpha/2} = -2.45$ *and* $t_{\alpha/2} = 2.45$

x	y	$d = x - y$	d^2
62	55	7	49
56	61	-5	25
59	62	-3	9
73	68	5	25
49	52	-3	9
54	51	3	9
67	62	5	25
-	-	9	151

$$\bar{d} = \frac{\sum d}{n} = \frac{9}{17} = 1.29$$

$$s_d = \sqrt{\frac{\sum d^2}{n} - \left[\frac{\sum d}{n}\right]^2} = \sqrt{\frac{151}{7} - \left[\frac{9}{7}\right]^2} = 4.46$$

$$t_{0bs} = \frac{\overline{d}}{s_d / \sqrt{n-1}} = \frac{+1.29}{4.46/\sqrt{7-1}} = 0.71$$

Since $t_{0bs} = -0.71$ is within the interval $(-2.45, 2.45)$, H_0 is accepted.

Conclusion: The mean weight of husbands and mean weight of wives are equal.

10.4 CHI-SQUARE TEST

The chi-square test is applied in statistics to test the goodness of fit to verify the distribution of observed data with assumed theoretical distribution. Therefore, it is a measure to study the divergence of actual and expected frequencies. It has great use in statistics especially in sampling studies, where we expect a doubled coincidence between actual and expected frequencies and the extent to which the difference can be ignored, because of fluctuations in sampling. If there is no difference between the actual and expected frequencies χ^2 is zero. Thus, the chi-square test describes the discrepancy between theory and observation.

Characteristics of χ^2 test:

1. The test is based on events of frequencies, whereas in theoretical distribution, the test is based on mean and standard deviation.
2. To draw inferences, this test is applied especially testing the hypothesis but not useful for estimation.
3. The test can be used between the entire set of observed and expected frequencies.
4. For every increase in the number of degrees of freedom, a new χ^2 distribution is formed.
5. It is a general-purpose test and as such is highly useful in research.

Assumptions:

1. All the observations must be independent

2. All the events must be mutually exclusive

3. There must be large observations

4. For comparison purposes the data must be in original units.

Degree of Freedom

When we compare the computed value of χ^2 with the table value, the degree of freedom is evident. The degree of freedom means the number of classes to which values can be assigned at will, without,

violating restrictions. For example, we choose any four numbers, whose total is 50. Here we have a the choice to select any three numbers, say 10, 15, 20, and the fourth number is 5; [50 – (10 + 15 + 20)]. Thus, our choice of freedom is reduced by one, on the condition that the total be 50. Therefore, the restriction placed on the freedom is one and the degree of freedom is three. As the restrictions increase, the freedom is reduced.

Thus $v = n - k$

v: (nu) = Degree of freedom

k: No. of independent constraints

n : Number of frequency classes.

For a contingency tale, 2 × 2, table, the degree of freedom is

$$v = (c - 1)(r - 1)$$

$$= (2 - 1)(2 - 1)$$

$$= 1$$

Uses:

1. χ^2 **test of goodness of fit**. Through the test, we can find out the deviations between the

observed values and expected values. Here we are not concerned with the parameters but with the form of distribution. Karl Pearson has developed a method to test the difference between the theoretical value (hypothesis) and the observed value. The test is done by comparing the computed value with the table value of χ^2 for the desired degree of freedom. A Greek letter χ^2 is used to describe the magnitude of the difference between the fact and theory.

The χ^2 maybe defined as

$$\chi^2 = \Sigma \left\{ \frac{(O-E)^2}{E} \right\}$$

O = Observed frequencies

E = Expected frequencies

Steps:

1. A hypothesis is established along with the significance level
2. Compute deviations between observed value and expected value (O – E)
3. Square the deviations calculated $(O – E)^2$
4. Divide the $(O – E)^2$ by its expected frequency
5. Add all the values obtained in step 4
6. Find the value of χ^2, from χ^2 table at certain level of significance, Usually, 5% level.

If the calculated value of χ^2 is greater than the tabulated value of χ^2 at certain level of significance, we reject the hypothesis. If the computed value of χ^2 is zero then the observed value and expected values completely coincide. If the computed value of χ^2 is less than the table value at a certain degree of level of significance, it is said to be non-significant. This

implies that the discrepancy between the observed and expected frequencies may be due to fluctuations in simple sampling.

Example:

4 coins were tossed 160 times and the following results were obtained:

No. of heads	:	0	1	2	3	4
Observed frequencies:	17	52	54	31	6	

Under the assumption that coins are balanced, find the expected frequencies of getting 0, 1, 2, 3 or 4 heads and test the goodness of fit hypothesis to see whether the coins are unbiased.

x	**Expected frequency**
	$160^4 cx\ (.5)^4 = E$
0	$160\ X^4\ C0\ (.5)^4 = 10$
1	$160\ X^4\ C1\ (.5)^4 = 40$
2	$160\ X^4\ C2\ (.5)^4 = 60$
3	$160\ X^4\ C3\ (.5)^4 = 40$
4	$160\ X^4\ C4\ (.5)^4 = 10$

When applying χ^2

No. of heads	***O***	***E***	***O – E***	***(O – E)²***	$\frac{(O-E)^2}{E}$
0	17	10	7	49	4.900

1	52	40	12	144	3.600
2	54	60	– 6	36	0.600
3	31	40	– 9	81	2.025
4	6	10	– 4	16	1.600

$$\Sigma \frac{(O-E)^2}{E} = 12.725$$

d.f. = 5 – 1 = 4; χ^2 0.05 = 9.488.

The calculated value of χ^2 is 12.725, which is greater than the tabulated value of 9.488. Therefore, the fit is poor.

2.**χ^2 as a test of independence** χ^2 test can be used to find out whether one or more attributes are associated or not. For example, coaching class and successful candidate marriage and failure etc., we can find out whether they are related or independent. We take a hypothesis that the attributes are independent. If the calculated value of χ^2 is less than the tabulated value at a certain level of significance the hypothesis is correct and vice versa.

Example:

Out of a sample of 120 persons in a village, 76 were administered a new drug for preventing influenza, and out of them 24 persons were attacked by influenza. Out of those who were not administered the new drug, 12 persons were not affected by influenza.

a) Prepare 2 × 2 tables showing the actual and expected frequencies.

b) Use the chi-square test to find out whether the new drug is effective or not.

[At 5% level for one degree of freedom, the value of chi-square is 3.84]

Solution:

2 × 2 table

	A	α	
B	24	32	56 (B)
β	52	12	64
	76	44	120
	(A)		N

Let the influenza and new drug be independent. The expected frequencies are:

$\frac{76\times56}{120}=35.5$	$\frac{56\times44}{120}=20.5$	56
$\frac{76\times64}{120}=40.5$	$\frac{64\times44}{120}=23.5$	64
76	44	120

O	**E**	**O – E**	**$(O-E)^2$**	$\frac{(O-E)^2}{E}$
24	35.5	– 11.5	132.25	3.725
52	40.5	11.5	132.25	3.265
32	20.5	11.5	132.25	6.451
12	23.5	– 11.5	132.25	5.627

$$\sum \frac{(O-E)^2}{E}=19.068$$

d.f. = (2 – 1) (2 – 1) = 1, χ^2 0.05 for d.f. = 3.84

The calculated value of χ^2 is 19.068 which is much higher than the tabulated value.

Therefore, the hypothesis is rejected. Hence, we conclude that the drug is undoubtedly effective in controlling influenza.

Example:

In a certain sample of 2,000 families 1,400 families are consumers of tea. Out of 1,800 Hindu families, 1,236 families consume tea. Use χ^2 test and state whether there is any significant difference between the consumption of tea among Hindu and non-Hindu families.

Solution:

On tabulation of the information in a 2 ×2 contingency table, we get:

	Hindu	Non-Hindu	Total
Consuming tea	1236	164	1400
Non-consuming tea	564	36	600
Total	1800	200	2000

Hypothesis – The attributes are independent on the basis of the independent the expected frequencies are:

$\frac{1800 \times 1400}{2000}$ 1440 – 1260 1400

= 1260 = 140

1800 – 1260 200 – 140

= 540 (600 – 540) = 60 600

1800 200 2000

Calculation of χ^2

O	**E**	**O – E**	**$(O-E)^2$**	**$\frac{(O-E)^2}{E}$**
1236	1260	– 24	576	0.457
564	540	+ 24	576	1.068
164	140	+ 24	576	4.114
36	60	– 24	576	9.600

$$\sum \frac{(O-E)^2}{E} = 15.239$$

d.f. is 1, tabulated value of χ^2 *0.05* for *1d.f. = 3.841*.

The calculated value of χ^2 *15.239* is higher than the tabulated value (i.e.) *3.841*. Therefore, the null hypothesis is rejected. Hence, the two communities differ significantly as far as the consumption of tea is concerned.

Example:

A dice is tossed 120 times with the following results:

No. of turned up:	1	2	3	4	5	6	Total
Frequency:	30	25	18	10	22	15	120

Test the hypothesis that the dice is unbiased.

Solution:

The hypothesis is that the dice is an unbiased one.

The expected frequency is $\left[120 \times \frac{1}{6}\right] = 20$

Applying the χ^2test.

O	**E**	**O – E**	**$(O - E)^2$**	$\frac{(O-E)^2}{E}$
30	20	10	100	5.00
25	20	5	25	1.25
18	20	– 2	4	0.20
10	20	– 10	100	5.00
22	20	2	4	0.20
15	20	– 5	25	1.25

$$\sum \frac{(O-E)^2}{E} = 12.90$$

d.f. = n – 1 = 6 – 1 = 5

For 5 d.f. at 5% level of significance on the basis of tabulated value is *11.07* which is less than the calculated value of $\chi^2 = 12.90$. Therefore, the hypothesis which is an unbiased one, is rejected at a 5% level of significance.

Example:

A certain drug was administered to 456 males out of a total of 720 in a certain locality to test its efficacy against typhoid. The incidence of typhoid is shown below. Find out the effectiveness of the drug against the disease. (The table value of $\chi^2$for 1 d.f. at 5% level of significance is 3.84)

	Infection	Non-infection	Total
Administering the drug	144	312	456
Without administering the drug	192	72	264
Total	336	384	720

Solution:

2 × 2 contingency table

	Infection	Non-infection	Total
Administration of drug	144	312	456
No administration of drug	192	72	264
Total	336	384	720

Hypothesis: The drug is independent, not effective on the basis of the independent the expected frequencies are:

$\frac{336 \times 456}{720}$	243.2	456
= 212.8		
123.2	140.8	264
336	384	720

O	**E**	**O – E**	**$(O-E)^2$**	$\frac{(O-E)^2}{E}$
144	212.8	– 68.8	4733.44	22.24
192	123.2	+ 68.8	4733.44	38.42
312	243.2	+ 68.8	4733.44	19.46
72	140.8	– 68.8	4733.44	33.62

$$\sum \frac{(O-E)^2}{E} = 113.74$$

The computed value of $\chi^2 = 113.74$ which is much greater than the table value of χ^2 *0.05* at *1 d.f. = 3.841*. Therefore, it is highly significant. The null hypothesis is wrong. Therefore, the drug is effective in controlling typhoid.

Example:

Out of 8,000 graduates in a town 800 are females; out of 1600 graduate employees 120 are female. Use $\chi^2$to determine if any distinction is made in an appointment based on sex. The value of χ^2 for 5% level for one degree of freedom is 3.84

Solution:

The information given in the question can be tabulated in a 2×2 table

	Employed	Unemployed	Total
Male	1480	5720	7200
Female	120	680	800
Total	1600	6400	8000

We take the hypothesis that there is no distinction in appointment on the basis of sex.

The expected frequencies are:

$\frac{7200 \times 1600}{8000}$	5760	7200

= 1440

160	640	800
1600	6400	8000

Applying χ^2 test

O	**E**	**O – E**	**$(O - E)^2$**	$\frac{(O-E)^2}{E}$
1480	1440	40	1600	1.111
120	160	– 40	1600	10.000
5720	5760	– 40	1600	0.278
680	640	40	1600	2.500

$$\sum \frac{(O-E)^2}{E} = 13.889$$

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 13.889$$

d.f. = (r – 1) (c – 1)

= 1 × 1 = 1

For d.f. 1, χ^2 0.05 = 3.84

The calculated value of χ^2 is 13.889 which is more than the table value (3.84) therefore the hypothesis is respected. It means that a distinction is made in appointment based on sex.

SELF CHECK QUESTIONS:

1. A filling machine is expected to fill 5 kg of powder into bags. A sample of 10 bags gave the weights 4.7, 4.9, 5.0, 5.1, 5.4, 5.2, 4.6, 5.1, 4.6 and 4.7. Test whether the machine is working properly.

2. A Company has been producing steel tubes with a mean inner diameter of 2.00 cm. A sample of 10 tubes gives an inner diameter of 2.01 cm and a variance of 0.004 cm^2. Is the difference in the value of the mean significant? Value of t for 9 df at 5% level = 2.262.

10.5 SUMMARY:

The significance level is usually denoted by the Greek symbol α (lowercase alpha). Popular levels of significance are 10% (0.1), 5% (0.05), 1% (0.01), 0.5% (0.005), and 0.1% (0.001). If a test of significance gives a p-value lower than the significance level α, the null hypothesis is rejected. Such results are informally referred to as 'statistically significant'. For example, if someone argues that "there's only one chance in a thousand this could have happened by coincidence," a 0.001 level of statistical significance is implied. The lower the significance level, the stronger the evidence required. Choosing the level of significance is a somewhat arbitrary task, but for many applications, a level of 5% is chosen, for no better reason than that it is conventional.

In some situations, it is convenient to express the statistical significance as $1 - \alpha$. In general, when interpreting a stated significance, one must be careful to note what, precisely, is being tested statistically.

Different levels of α trade-off countervailing effects. Smaller levels of α increase confidence in the determination of significance, but run an increased risk of failing to reject a false null hypothesis (a Type II error, or "false negative determination"), and so have less statistical power. The selection of the level α thus inevitably involves a compromise between significance and power, and consequently between the Type I error and the Type II error. More powerful experiments – usually experiments with more subjects or replications – can obviate this choice to an arbitrary degree.

10.6 GLOSSARY

- **Degree of freedom:** means the number of classes to which values can be assigned at will, without, violating restrictions.
- **Chi-square test:** is applied in statistics to test the goodness of fit

10.7 CHECK YOUR PROGRESS

.1. The sample size n = 10, μ = 5 kg

Let us first calculate $\bar{x}$ and s from the sample data given

									Total
x	: 4.7	4.9	5.0	5.1	5.4	5.2	4.6	4.7	**49.3**
x^2	: 22.09	24.01	25.00	26.01	29.16	27.04	21.16	22.09	**243.73**

$$\bar{x} = \frac{49.3}{10} = 4.93$$

$$S = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2} = \sqrt{\frac{243.73}{10} - (4.93)^2}$$

$$= \sqrt{2.4373 - 24.30}$$

$$= \sqrt{0.073} = 0.27$$

$H_0 = \mu = 5$ kg

$H_1 = \mu \neq 5$ kg

The test statistic is $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}} = \frac{4.93 - 5}{\frac{.27}{\sqrt{9}}}$

$$= \frac{-0.07 \times 3}{.27} = -0.78$$

ndf = n – 1 = 10 – 1 = 9

Table value of t at 5% level for 9 d.f. = 2.262.

$\mu = 2.00$ cm

n = 10 tubes

$\bar{x} = 2.01$ cm

$\delta = \sqrt{0.004}$ cm

Since n < 30, the sample is small. Let us therefore apply the test for testing the mean.

H_0: $\mu = 2.00$ cm

H_1: $\mu \neq 2.00$ cm

The test statistic is

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}} = \frac{2.01 - 2.00}{\frac{\sqrt{0.004}}{\sqrt{9}}} = \frac{.01 \times 3}{\sqrt{.004}}$$

$$= \frac{.03}{.0632}$$

$$= 0.475$$

n d.f. (number of degree of freedom) = 9

Table value of f for 9 d.f. at 5% level = 2.262

2. $\mu = 2.00$ cm

n = 10 tubes

$\bar{x} = 2.01$ cm

$\delta = \sqrt{0.004}$ cm

Since n < 30, the sample is a small sample. Let us therefore apply t test for testing the mean.

H_0: $\mu = 2.00$ cm

H_1: $\mu \neq 2.00$ cm

The test statistic is

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}} = \frac{2.01 - 2.00}{\frac{\sqrt{0.004}}{\sqrt{9}}} = \frac{.01 \times 3}{\sqrt{.004}}$$

$$= \frac{.03}{.0632}$$

$$= 0.475$$

n d.f. (number of degree of freedom) = 9

Table value of f for 9 d.f. at 5% level = 2.262

3. $p_1 = \frac{16}{500} = 0.032$ (in the first sample)

$p_2 = \frac{3}{100} = 0.03$ (in the second sample)

Let us assume that the machine has not improved after overhauling.

$$S.E.(p_1 - p_2) = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$p = \frac{500 \times 0.032 + 100 \times 0.3}{500 + 100}$$

$$= \frac{16 + 3}{600} = 0.03$$

$$q = 1 - 0.03 = 0.97$$

$$S.E.(p_1 - p_2) = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$= \sqrt{(0.03)(0.97)\left(\frac{1}{500}\right) + \left(\frac{1}{100}\right)}$$

$$= \sqrt{(0.03)(0.97)(0.002+0.01)}$$

$$= 0.0187$$

$$z = \frac{0.032 - 0.03}{0.0187} = \frac{0.002}{0.0187} = 0.106$$

4.Let $X_1$and X_2denote the hourly wages (in Rs.) of workers in plant A and plant B respectively.

Then we are given

$n_1 = 150$ $\bar{x}_1 = 2.56$ $s_1 = 1.08 = \bar{\sigma}_1$

$n_2 = 200$ $\bar{x}_2 = 2.87$ $s_2 = 1.28 = \bar{\sigma}_2$

Null Hypothesis H_0

$\mu_1 = \mu_2$ is there is no significant difference between the mean level of wages of workers in plant A and plant B.

Alternative Hypothesis H_1

$\mu_2 > \mu_1$ i.e. $\mu_1 < \mu_2$ (left-tailed test)

Test statistic

Under H_0, the test statistic (for large samples) as:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$\therefore z = \frac{2.56 - 2.87}{\sqrt{\frac{(1.08)^2}{150} + \frac{(1.28)^2}{200}}}$$

$$z = \frac{-0.31}{\sqrt{0.016}} = \frac{-0.31}{0.126}$$

$= -2.46$

Critical region

For a one-tailed test, the critical value of z at 5% level of significance is 1.645. The critical region for the left-tailed test consists of all values of z.

10.8 ANSWERS TO CHECK YOUR PROGRESS

1. Conclusion: H_0 is accepted at 5% level Hence the machine is working properly.

2. Conclusion: H_0 is accepted at a 5% level since the calculated value of t is less than the table value of t. Therefore, the difference between the means of the population and the sample is not significant.

3. Since the difference is less than 2.58 S.E. (1% level), the result of the experiment supports the hypothesis. Therefore, we conclude that the machine has not improved after overhauling.

4. **Conclusion:** Since the calculated value of z(–2.46) is less than the critical value (–1.645) it is significant at a 5% level of significance. Hence the null hypothesis is rejected at a 5% level of significance and we conclude that the average hourly wages paid by plant 'B' are certainly higher than those paid by plant 'A'.

10.9 TERMINAL QUESTIONS

I. A filling machine is expected to fill 5 kg of powder into bags. A sample of 10 bags gave the weights 4.7, 4.9, 5.0, 5.1, 5.4, 5.2, 4.6, 5.1, 4.6 and 4.7. Test whether the machine is working properly.

II. A Company has been producing steel tubes of a mean inner diameter of 2.00 cm. A sample of 10 tubes gives an inner diameter of 2.01 cm and a variance of 0.004 cm^2. Is the difference in the value of the mean significant? Value of t for 9 df at 5% level = 2.262.

1. A machine puts out 16 imperfect articles in a sample of 500, after the machine is overhauled, it puts out 3 imperfect articles in a batch of 100. Has the machine improved?

2. The average hourly wage of a sample of 150 workers in a plant 'A' was Rs. 2.56 with a standard deviation of Rs. 1.08. The average wages of a sample of 200 workers in plant 'B' was Rs. 2.87 with a standard deviation of Rs. 1.28 can an applied safely assume that the hourly wages paid by plant 'B' are higher than those paid by Plant 'A'?

10.10 SUGGESTED READINGS

1. Basic Statistics – Goon, Guptha and Dasguptha – World Press Limited – Calcutta.

2. Fundamentals of Business Statistics – Sanchethi and Kappor.

3. Quantitative Methods in Management – Srivastava, Shenoy and Guptha.

4. Business Statistics – Guptha and Guptha.

PAPER CODE: MCM-02
BLOCK: 4

UNIT 11 : PARTIAL AND MULTIPLE CORRELATION

Structure

OBJECTIVES

- After studying this unit, you will be able to understand the multiple correlation and partial correlation.

11.1 INTRODUCTION

In statistics, dependence refers to any statistical relationship between two random variables or two sets of data. Correlation refers to any of a broad class of statistical relationships involving dependence. Familiar examples of dependent phenomena include the correlation between the physical statures of parents and their offspring and the correlation between the demand for a product and its price. Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. For example, an electrical utility may produce less power on a mild day based on the correlation between electricity demand and weather. In this example there is a causal relationship because extreme weather causes people to use more electricity for heating or cooling; however, statistical dependence is not sufficient to demonstrate the presence of such a causal relationship. Formally, dependence refers to any situation in which random variables do not satisfy a mathematical condition of probabilistic independence. In loose usage, correlation can refer to any departure of two or more random variables from independence, but technically it refers to any of several more specialized types of relationship between mean values. There are several correlation coefficients, often denoted ρ or r, measuring the degree of correlation.

To find out the relationship and dependence among the variable, partial and multiple correlations are studied. These are an extension of the technique of simple correlation under which you may study the interrelationship between three or more variables.

11.2 MULTIPLE CORRELATION

The coefficient of multiple correlation is a measure of how well a given variable can be predicted using a linear function of a set of other variables. It is measured by the coefficient of determination, but under the particular assumption that that the best possible linear predictors are used, whereas the coefficient of determination is defined for more general cases. The coefficient of multiple determination takes values between zero and one; a higher value indicates a better predictability of the dependent variables from the independent variables, with a value of one indicating that the predictions are exact and a value of zero indicating that no linear combination of dependent variables is better than the simpler predictor which consists of mean of the target variable.

Multiple correlation is the study of the relationship among three or more variables. It measures the combined influence of two or more independent variables on a single dependent variable. For example, if you study the combined influence of amount of fertilizer (x_2) and rainfall (x_3) on the yield of wheat (x_1), then it is called problem of multiple correlation. You shall denote the multiple correlation coefficient between x_1, the dependent variables x_2 and x_3 independent variables by $R_{1.23}$. Similarly, you can denote the other Multiple correlation coefficients by $R_{2.13}$ and $R_{3.12}$.

Calculation of Coefficient of Multiple Correlation: The formulae for calculating the multiple correlation coefficients $R_{1.23}, R_{2.13}$ and $R_{3.12}$ are as follows:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}.r_{13}.r_{23}}{1 - r_{23}^2}}$$

Where, $R_{1.23}$ = Multiple correlation coefficient r_{12}, r_{13}, r_{23} = Simple or zero order correlation coefficient.

$$R_{2.13} = \sqrt{\frac{r_{21}^2 + r_{23}^2 - 2r_{21}.r_{23}.r_{13}}{1 - r_{13}^2}}$$

$$R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{31}.r_{32}.r_{12}}{1 - r_{12}^2}}$$

Limits of Multiple Correlation Coefficients: The value of multiple correlation coefficient ($R_{1.23}$) lies between 0 and 1. It can never be negative.

$$0 \geq R_{1.23} \leq 1$$

Example 11.1: Calculate $R_{1.23}, R_{2.13}$ and $R_{3.12}$ for the following data:

$$r_{12} = 0.60, r_{13} = 0.70, r_{23} = 0.65$$

Solution: Given, $r_{12} = 0.60, r_{13} = 0.70, r_{23} = 0.65$

(i)

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}.r_{13}.r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.6)^2 + (0.7)^2 - 2\,(0.6)\,(0.7)\,(0.65)}{1 - (0.65)^2}}$$

$$= \sqrt{\frac{0.36 + 0.49 - 0.546}{0.5775}}$$

$$= \sqrt{0.526}$$

$$= 0.725$$

(ii)

$$R_{2.13} = \sqrt{\frac{r_{21}^2 + r_{23}^2 - 2r_{21}.r_{23}.r_{13}}{1 - r_{13}^2}}$$

$$= \sqrt{\frac{(0.6)^2 + (0.65)^2 - 2(0.6)(0.65)(0.7)}{1 - (0.70)^2}}$$

$$= \sqrt{\frac{0.36 + 0.4225 - 0.546}{1 - 0.49}}$$

$$= \sqrt{0.4638}$$

$$= 0.6809$$

(iii)

$$R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{31}.r_{32}.r_{12}}{1 - r_{12}^2}}$$

$$= \sqrt{\frac{(0.70)^2 + (0.65)^2 - 2\,(0.70)\,(0.65)\,(0.60)}{1 - (0.60)^2}}$$

$$= \sqrt{\frac{0.49 + 0.4225 - 0.546}{1 - 0.36}}$$

$$= \sqrt{0.5726} = 0.756.$$

Example 11.2: For a large group of students x_1 = Score in Economics, x_2 = Score in Maths, x_3 = Score in Statistics, $r_{12} = 0.69, r_{13} = 0.45, r_{23} = 0.58$. Determine the coefficient of multiple correlation.

Solution:

$$R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{31}.r_{32}.r_{12}}{1 - r_{12}^2}}$$

$$\sqrt{\frac{(0.45)^2 + (0.58)^2 - 2\,(0.45)\,(0.58)\,(0.69)}{1 - (0.69)^2}}$$

$$= \sqrt{\frac{0.2025 + 0.3364 - 0.3601}{1 - 0.4761}}$$
$$= \sqrt{0.3412} = 0.584$$

Example 11.3: The following zero order correlation coefficient are given:

$$r_{12} = 0.98, r_{13} = 0.44 \; and \; r_{23} = 0.54.$$

Calculate multiple correlation coefficient treating the first variable as dependent and second and third variables as independent.

Solution: You have to calculate the multiple correlation coefficient treating first variable as dependent and second and third variables as independent *i.e.*, you have to find $R_{1.23}$.

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}.r_{13}.r_{23}}{1 - r_{23}^2}}$$

Substituting the given values,

$$R_{1.23} = \sqrt{\frac{(0.98)^2 + (0.44)^2 - 2(0.98)(0.44)(0.54)}{1 - (0.54)^2}}$$
$$= \sqrt{\frac{0.9604 + 0.1936 - 0.4657}{1 - 0.2916}}$$
$$= \sqrt{\frac{.6883}{.7084}}$$
$$= \sqrt{0.9716} = 0.985$$

Example 11.4: If $R_{1.23} = 1$, prove that $R_{2.13} = 1$, .

Solution:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}.r_{13}.r_{23}}{1 - r_{23}^2}}$$

and

$$R_{2.13} = \sqrt{\frac{r_{21}^2 + r_{23}^2 - 2r_{21}.r_{23}.r_{13}}{1 - r_{13}^2}}$$

putting $R_{2.13} = 1$ and squaring both sides,

$$1 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}.r_{13}.r_{23}}{1 - r_{23}^2}$$
$$\Longrightarrow r_{12}^2 + r_{13}^2 - 2r_{12}.r_{13}.r_{23} = 1 - r_{23}^2$$
$$\Longrightarrow r_{12}^2 + r_{23}^2 - 2r_{12}.r_{13}.r_{23} = 1 - r_{13}^2$$
$$\Longrightarrow \frac{r_{12}^2 + r_{13}^2 - 2r_{12}.r_{23}.r_{23}}{1 - r_{13}^2} = 1$$
$$\Longrightarrow R_{2.13}^2 = 1 \; or \; R_{2.13} = 1$$

Since, the coefficient of multiple correlation is considered non-negative.

11.3 PARTIAL CORRELATION

According to Guilford (1973), the partial correlation between *X* and *Y* given a set of *n* controlling variables **Z** = {Z_1, Z_2, …, Z_n}, written $\rho_{XY \cdot \mathbf{Z}}$, is the correlation between the residuals R_X and R_Y resulting from the linear correlation of *X* with **Z** and of *Y* with **Z**, respectively. In fact, the first-order partial correlation is nothing else than a difference between a correlation and the product of the removable correlations divided by the product of the coefficients of alienation of the removable correlations.

Partial Correlation is the simple correlation between two variables after eliminating the influence of the third variable from them. For example, if you measure the relationship between yield of wheat (x_1) and the amount of fertilizer (x_2), eliminating the effect of climate (x_3) from both (having the same climate), then it is called the problem of partial correlation. For three variables ($x_1, x_2 and x_3$), there are three partial correlation coefficients. They are denoted by $r_{12.3}, r_{13.2}$ and $r_{23.1}$. The partial correlation coefficient $r_{12.3}$ indicates the relationship between x_1 and x_2 when the effect of x_3 is eliminated from both.

Calculation of Partial Correlation Coefficients: The formulae for calculating the partial correlation coefficients $r_{12.3}, r_{13.2}$ and $r_{23.1}$ are as follows:

$$r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

Where, $r_{12.3}$ = Partial correlation between x_1 and x_2 r_{12}, r_{13} and r_{23} = Simple or zero order correlation coefficient. Similarly, you have

$$r_{23.1} = \frac{r_{23} - r_{21}.r_{31}}{\sqrt{1 - r_{21}^2}\sqrt{1 - r_{31}^2}}$$

$$r_{13.2} = \frac{r_{13} - r_{12}.r_{32}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{32}^2}}$$

Limits of Partial Correlation Coefficient: The value of $r_{12.3}$ lies between -1 and +1.

$$-1 \leq r_{12.3} \leq 1$$

Example 11.5 : Given that $r_{12} = 0.7, r_{13} = 0.61, r_{23} = 0.4$. Find the value of $r_{12.3}, r_{13.2}, r_{23.1}$.

Solution:

$$r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

Substituting the values,

$$r_{12.3} = \frac{0.7 - (0.61)(0.4)}{\sqrt{1-(0.61)^2}\sqrt{1-(0.4)^2}}$$

$$= \frac{0.456}{0.792 \times 0.916} = 0.629$$

$$r_{13.2} = \frac{r_{13} - r_{12}.r_{32}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{32}^2}}$$

$$= \frac{0.61 - (0.7)(0.4)}{\sqrt{1-(0.7)^2}\sqrt{1-(0.4)^2}}$$

$$= \frac{0.61 - 0.28}{\sqrt{1-.49}\sqrt{1-.16}}$$

$$= \frac{0.33}{0.714 \times 0.916}$$

$$= \frac{0.33}{0.654}$$

$$= 0.505$$

$$r_{23.1} = \frac{r_{23} - r_{21}.r_{31}}{\sqrt{1-r_{21}^2}\sqrt{1-r_{31}^2}}$$

$$= \frac{0.4 - (0.7)(0.61)}{\sqrt{1-(0.7)^2}\sqrt{1-(0.61)^2}}$$

$$= \frac{0.4 - 0.427}{\sqrt{1-(0.49)}\sqrt{1-.3721}}$$

$$= \frac{-0.027}{0.714 \times 0.792}$$

$$= -0.048$$

Example 11.6: On the basis of observations made on 30 cotton plants the total correlation of yield of cotton (x_1) the number of balls *i.e.*, seed vessel (x_1) and height (x_3) are found to be :

$$r_{12} = 0.8, r_{13} = 0.65, r_{23} = 0.7$$

Compute the partial correlation between yield of cotton and number of balls, eliminating the effect of height.

Solution: You have to find the partial correlation between yield of cotton x_1 and the number of balls x_2, eliminating the effect of height x_3 *i.e.*, you have to find $r_{12.3}$.

$$r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

Substituting the given values,

$$r_{12.3} = \frac{0.8 - (0.65)(0.7)}{\sqrt{1 - (0.65)^2}\sqrt{1 - (0.7)^2}}$$

$$= \frac{0.8 - 0.45}{\sqrt{1 - 0.4225}\sqrt{1 - 0.49}}$$

$$= \frac{0.345}{0.76 \times 0.714}$$

$$= \frac{0.345}{0.543}$$

$$= 0.635$$

Example 11.**7:** For a large group of students x_1 = Score in theory, x_2 = Score in method, x_3 = Score in field work. The following results were found:

$$r_{12} = 0.69, r_{13} = 0.45, r_{23} = 0.58$$

Determine the partial correlation coefficient between score in field work and score in theory keeping the score in method constant and interpret the result.

Solution: You have to find partial correlation coefficient between score in field (x_3) and score in theory (x_1) keeping the scores in method constant i.e., you have to find $r_{31.2}$.

$$r_{31.2} = \frac{r_{31} - r_{32}.r_{12}}{\sqrt{1 - r_{32}^2}\sqrt{1 - r_{12}^2}}$$

$$= \frac{0.45 - (0.58)(0.69)}{\sqrt{1 - (0.58)^2}\sqrt{1 - (0.69)^2}}$$

$$= \frac{0.45 - 0.4002}{\sqrt{1 - 0.3364}\sqrt{1 - 0.4761}}$$

$$= \frac{0.0498}{\sqrt{0.6636}\sqrt{0.5239}}$$

$$= \frac{0.0498}{0.81 \times 0.72}$$

$$= \frac{0.0498}{0.5832}$$

$$= 0.085$$

Thus, there is low degree of correlation between score in field work and score in theory.

Example 11.8: Is it possible to have the following set of experimental data:

$$r_{12} = 0.6, r_{13} = 0.8, r_{23} = -0.5$$

Solution: In order to see whether there is inconsistency in the given data, you should calculate $r_{12.3}$. If the value of $r_{12.3}$ exceeds one, there is inconsistency, otherwise not.

$$r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

Substituting the given values,

$$= \frac{0.6 - (-0.5)(0.8)}{\sqrt{1 - (0.5)^2}\sqrt{1 - (0.8)^2}}$$

$$= \frac{0.6 + 0.4}{\sqrt{1 - 0.25}\sqrt{1 - 0.64}}$$

$$= \frac{1}{\sqrt{0.75}\sqrt{0.36}}$$

$$= \frac{1}{0.866 \times 0.6}$$

$$= \frac{1}{0.52}$$

$$= 1.92$$

Since, the value of $r_{12.3}$ is greater than one, there is some inconsistency in the given date.

Alternatively: You can also check the inconsistency in the data by calculating $R_{1.23}$. If the value of $R_{1.23}$ exceeds 1, there is some inconsistency otherwise not

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}.r_{13}.r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.6)^2 + (-0.5)^2 - 2(0.6)(-0.5)(0.8)}{1-(0.8)^2}}$$

$$= \sqrt{\frac{0.36 + 0.25 + 0.48}{1 - .64}}$$

$$= \sqrt{\frac{1.09}{0.36}}$$

$$= \sqrt{3.0277}$$
$$= 1.74$$

Since, the value of $R_{1.23}$ is greater than one, there is some inconsistency in the given data.

Example 11.9: Suppose a computer has found, for a given set of values of x_1, x_2 and x_3 :

$$r_{12} = 0.96, r_{13} = 0.36, r_{23} = 0.78$$

Explain whether these computed values may be said to be free from errors.

Solution: For determining whether the given computed values are free from errors or not, you compute the values of $r_{12.3}$. If $r_{12.3}$ comes out to be greater than one, the computed values cannot be regarded as free from errors.

$$r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

Substituting the given values,

$$r_{12.3} = \frac{0.96 - (0.36)(0.78)}{\sqrt{1-(0.36)^2}\sqrt{1-(0.78)^2}}$$

$$= \frac{0.96 - 0.2808}{\sqrt{0.8704}\sqrt{0.3916}}$$

$$= \frac{0.6792}{0.9329 \times 0.6258}$$

$$= \frac{0.6792}{0.5838}$$

$$= 1.163$$

Since, $r_{12.3}$ is greater than one, the given computed values do contain some errors.

11.4 RELATIONSHIP BETWEEN SIMPLE, PARTIAL AND MULTIPLE CORRELATION COEFFICIENTS.

There exists relationship between simple, partial and multiple correlation coefficients which is clear from the following equation:

(i) $1 - R_{1.23}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2)$
(ii) $1 - R_{2.13}^2 = (1 - r_{21}^2)(1 - r_{23.1}^2)$ and
(iii) $1 - R_{3.12}^2 = (1 - r_{31}^2)(1 - r_{32.1}^2)$

Example 11.10: In a trivariate distribution, $r_{12} = 0.60, r_{13} = 0.70, r_{23} = 0.65$.
find $R_{1.23}^2$ from r_{12} and $r_{13.2}$.

Solution: Given : $r_{12} = 0.60, r_{13} = 0.70, r_{23} = 0.65$
Multiple, Simple and Partial Correlation Coefficient are related as:

$$R_{1.23}^2 = 1 - (1 - r_{12}^2)(1 - r_{13.2}^2)$$

$$r_{13.2} = \frac{r_{13} - r_{12}.r_{32}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{32}^2}}$$

$$= \frac{0.70 - 0.60 \times 0.65}{\sqrt{1 - (0.60)^2}\sqrt{1 - (0.65)^2}}$$

$$= \frac{0.70 - 0.39}{0.8 \times 0.760}$$

$$= 0.509$$

$$\therefore \quad r_{13.2}^2 = 0.259,$$

$$r_{12}^2 = 0.36$$

Substituting values r_{12}^2 and $r_{13.2}^2$ for $R_{1.23}^2$, you have

$$R_{1.23}^2 = 1 - (1 - 0.36)(1 - 0.259)$$

$$= 1(0.64)(0.74) = 0.526.$$

Example 11.11: x_1, x_2 and x_3 are measured from their means with:

$$N = 10, \sum x_1^2 = 90, \sum x_2^2 = 160, \sum x_3^2 = 40$$
$$\sum x_1 x_2 = 60, \sum x_2 x_3 = 60, \sum x_3 x_1 = 40$$

Calculate $r_{12.3}$ and $R_{2.31}$.

Solution:

$$r_{12} = \frac{\sum x_1 x_2}{\sqrt{\sum x_1^2}\sqrt{\sum x_2^2}}$$

$$= \frac{60}{\sqrt{\sqrt{90 \times 160}}}$$

$$= \frac{60}{120}$$

$$= 0.5$$

$$r_{13} = \frac{\sum x_1 x_3}{\sqrt{\sum x_1^2 \times \sum x_3^2}}$$

$$= \frac{40}{\sqrt{\sqrt{90 \times 40}}}$$

$$= \frac{40}{60}$$

$$= 0.67$$

$$r_{23} = \frac{\sum x_2 x_3}{\sqrt{\sum x_2^2 \times \sum x_3^2}}$$

$$= \frac{60}{\sqrt{\sqrt{160 \times 40}}}$$

$$= \frac{60}{80}$$

$$= 0.75$$

Now,

$$r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

Substituting the values, you have

$$r_{12.3} = \frac{0.5 - 0.67 \times 0.75}{\sqrt{1 - (0.67)^2}\sqrt{1 - (0.75)^2}}$$

$$= -\frac{0.0025}{0.4910}$$

$$= -0.0051$$

$$R_{2.31} = \sqrt{\frac{r_{23}^2 + r_{21}^2 - 2r_{23}.r_{21}.r_{31}}{1 - r_{31}^2}}$$

$$= \sqrt{\frac{(0.75)^2 + (0.5)^2 - 2(0.75)(0.5)(0.67)}{1 - (0.67)^2}}$$

$$= \sqrt{\frac{0.5625 + 0.25 + 0.5025}{0.5511}}$$

$$= \sqrt{\frac{0.31}{0.5511}}$$

$$= 0.75$$

Example 11.12: Calculate $r_{12.3}$ and $R_{1.23}$ from the following data:

X:	3	4	5	6	7	8	9
Y:	2	5	6	4	3	2	4
Z:	5	6	4	5	6	5	8

Solution:

Calculate $r_{12.3}$ and $R_{1.23}$

X	X^2	Y	Y^2	Z	Z^2	XY	XZ	YZ
3	9	2	4	5	25	6	15	10
4	16	5	25	6	36	20	24	30
5	25	6	36	4	16	30	20	24
6	36	4	16	5	25	24	30	20
7	49	3	9	6	36	21	42	18
8	64	2	4	5	25	16	40	10
9	81	4	16	8	64	36	72	32
$N = 7$ $\sum X = 42$	$\sum X^2 = 42$	$\sum Y = 26$	$\sum Y^2 = 110$	$\sum Z = 39$	$\sum Z^2 = 227$	$\sum XY = 153$	$\sum XZ = 243$	$\sum YZ = 144$

$$r_{12} = \frac{N.\sum XY - \sum X.\sum Y}{\sqrt{\left[\sum X^2.N - (\sum X)^2\right]\left[\sum Y^2.N - (\sum Y)^2\right]}}$$

$$= \frac{7 \times 153 - (42 \times 26)}{\sqrt{[280 \times 7 - (42)^2][110 \times 7 - (26)^2]}}$$

$$= -0.155$$

$$r_{13} = \frac{N.\sum XZ - \sum X.\sum Z}{\sqrt{\left[\sum X^2.N - (\sum X)^2\right]\left[\sum Z^2.N - (\sum Z)^2\right]}}$$

$$= \frac{7 \times 243 - (42 \times 39)}{\sqrt{[280 \times 7 - (42)^2][227 \times 7 - (39)^2]}}$$

$$= 0.546$$

$$r_{23} = \frac{N.\sum YZ - \sum Y.\sum Z}{\sqrt{\left[\sum Y^2.N - (\sum Y)^2\right]\left[\sum Z^2.N - (\sum Z)^2\right]}}$$

$$= \frac{144 \times 7 - 26 \times 39}{\sqrt{[110 \times 7 - (26)^2][227 \times 7 - (39)^2]}}$$

$$= -0.075$$

Partial Correlation Coefficient

$$r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

$$= \frac{-0.155 - (0.546 \times -0.075)}{\sqrt{1 - (0.546)^2}\sqrt{1 - (-0.075)^2}}$$

$$= -0.1366$$

Multiple Correlation Coefficient:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}.r_{13}.r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(-0.155)^2 + (0.556)^2 - (2 \times -0.155 \times 0.546 \times -0.075)}{1 - (-0.075)^2}}$$

$$= \sqrt{\frac{0.024 + 0.298 - (.01269)}{1 - .006}}$$

$$= \sqrt{\frac{0.1951}{0.994}}$$

$$= \sqrt{0.1962}$$

$$= 0.443$$

Example 11.13: In a Trivariate Distribution, $r_{12} = 0.80, r_{23} = -0.56, r_{31} = -0.40$ compute $r_{23.1}$ and $R_{1.23}$.

Solution: (I)

$$r_{23.1} = \frac{r_{23} - r_{21}.r_{31}}{\sqrt{1 - r_{21}^2}\sqrt{1 - r_{31}^2}}$$

$$= \frac{-0.56 - (0.8)(-0.40)}{\sqrt{1 - (0.8)^2}\sqrt{1 - (-0.4)^2}}$$

$$= \frac{-0.56 + 0.32}{\sqrt{1 - 0.64}\sqrt{1 - 0.16}}$$

$$= \frac{-0.24}{\sqrt{0.36 \times 0.84}}$$

$$= \frac{-0.24}{0.5499}$$

$$= -0.436$$

(II)

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}.r_{13}.r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.8)^2 + (-0.4)^2 - 2(0.8)(-0.4)(-0.56)}{1 - (-0.56)^2}}$$

$$= \sqrt{\frac{0.64 + 0.16 - 0.3584}{1 - 0.3136}}$$

$$= \sqrt{\frac{0.4416}{0.6864}}$$

$$= 0.802$$

Example 11.14: The linear correlation coefficient between x_1 (Yield), x_2 (Irrigation) and x_3 (Fertiliser) are as follows:

$$r_{12} = 0.81, r_{13} = 0.90, r_{23} = 0.65$$

Calculate the partial correlation coefficient of: (I) yield with irrigation

(II) yield with fertilizer**.**

Solution: (I) You have to find

$$r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

Substituting the given values,

$$r_{12.3} = \frac{(0.81) - (0.90)(0.65)}{\sqrt{1-(0.90)^2}\sqrt{1-(0.65)^2}}$$

$$= \frac{0.81 - 0.585}{0.4358 \times 0.7599}$$

$$= \frac{0.225}{0.3311}$$

$$= 0.679$$

(ii) You have to find $r_{13.2}$

$$r_{13.2} = \frac{r_{13} - r_{12}.r_{32}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{32}^2}}$$

$$= \frac{(0.90) - (0.81)(0.65)}{\sqrt{1-(0.81)^2}\sqrt{1-(0.65)^2}}$$

$$= \frac{0.3735}{\sqrt{0.3439}\sqrt{0.5775}}$$

$$= \frac{0.3735}{0.5864 \times 0.7599}$$

$$= \frac{0.3735}{0.4456}$$

$$= 0.838$$

Example 11.15: Given the following zero order correlation coefficient, find (i) partial correlation coefficient between x_2 and x_3 and (ii) multiple correlation taking x_1 as dependent on x_2 and x_3.

$$r_{12} = 0.98, r_{13} = 0.44, r_{23} = 0.54$$

Solution: (I)

$$r_{23.1} = \frac{r_{23} - r_{21}.r_{31}}{\sqrt{1-r_{21}^2}\sqrt{1-r_{31}^2}}$$

$$= \frac{0.54 - (0.98)(0.44)}{\sqrt{1-(0.98)^2}\sqrt{1-(0.44)^2}}$$

$$= \frac{0.54 - 0.4312}{\sqrt{1-0.9604}\sqrt{1-0.1936}}$$

$$= \frac{0.1088}{\sqrt{0.0396}\sqrt{0.8064}}$$

$$= \frac{0.1088}{0.1786}$$

$$= 0.6091$$

(II)

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}.r_{13}.r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.98)^2 + (0.44)^2 - 2(0.98)(-0.44)(0.54)}{1 - (0.54)^2}}$$

$$= \sqrt{\frac{0.9604 + 0.1936 - 0.4656}{1 - (0.2916)}}$$

$$= \sqrt{\frac{0.6884}{0.7084}}$$

$$= \sqrt{0.9717}$$

$$= 0.985$$

Example 11.16: Is it possible to get the following from a set of experimental data:

(I) $r_{23} = 0.8, r_{31} = 0.5, r_{12} = 0.6$

(II) $r_{23} = 0.7, r_{31} = -0.4, r_{12} = 0.6$

Solution: (I) In order to see whether there is any inconsistency, you should calculate $r_{12.3}$, if its value exceed one, there is inconsistency, otherwise not.

$$r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

$$= \frac{0.6 - (-0.5)(0.8)}{\sqrt{1 - (0.5)^2}\sqrt{1 - (0.8)^2}}$$

$$= \frac{0.20}{\sqrt{0.75}\sqrt{0.36}}$$

$$= \frac{0.20}{0.52}$$

$$= 0.384$$

Since, the value of $r_{12.3}$ is less than one, the data is consistent.

(II)

$$r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

$$= \frac{(0.6) - (-0.4)(0.7)}{\sqrt{1-(-0.4)^2}\sqrt{1-(0.7)^2}}$$

$$= \frac{0.6 + 0.28}{\sqrt{0.84}\sqrt{0.51}}$$

$$= \frac{0.88}{0.655}$$

$$= 1.344$$

Since, the value of $r_{12.3}$ is greater than 1 there is some onconsistency in the given data.

Example 11.17: Test the consistency of the following data:

$$r_{12} = 0.8, r_{13} = 0.4, r_{23} = -0.56.$$

Solution: For testing whether the given computations are consistent or not, you compute the value of $r_{13.2}$. If $r_{13.2}$ comes out to be greater than one, the computations cannot regarded as consistent.

$$r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

$$= \frac{(0.8) - (0.4)(-0.56)}{\sqrt{1-(-0.4)^2}\sqrt{1-(-0.56)^2}}$$

$$= \frac{0.8 + 0.224}{\sqrt{0.84}\sqrt{0.6864}}$$

$$= \frac{1.024}{0.7593}$$

$$= 1.349$$

Since, $r_{12.3}$ is greater than one, the given computations of r_{12}, r_{13} and r_{23} are not consistent.

Example 11.18: If $r_{12} = 0.77, r_{13} = 0.72, r_{23} = 0.52$, find the partial correlation coefficient $r_{12.3}$ and multiple correlation coefficient $R_{1.23}$.

Solution: Given : $r_{12} = 0.77, r_{13} = 0.72, r_{23} = 0.52,$

$$r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}}$$

$$= \frac{.77 - .72 \times .52}{\sqrt{1-(.72)^2}\sqrt{1-(.52)^2}}$$

$$= \frac{.77 - .37}{\sqrt{1 - .5184}\sqrt{1.2704}}$$

$$= \frac{.40}{\sqrt{.4816} \times \sqrt{.7296}}$$

$$= \frac{.4}{.593}$$

$$= 0.6745$$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}.r_{13}.r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(.77)^2 + (.72)^2 - 2(.77)(.72)(.52)}{1 - (0.54)^2}}$$

$$= \sqrt{\frac{0.9604 + 0.1936 - 0.4656}{1 - (.52)^2}}$$

$$= \sqrt{\frac{.5929 + .5184 - .5766}{1 - .2704}}$$

$$= \sqrt{\frac{.5347}{.7296}}$$

$$= 0.856 \ .$$

Example 11.19: If $r_{12} = 0.60, r_{13} = 0.70, r_{23} = 0.65$, find the partial correlation between x_1 and x_2 and multiple correlation between x_1 dependent on x_2 and x_3.

Solution: Given: $r_{12} = 0.60, r_{13} = 0.70, r_{23} = 0.65$,

$$r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

$$= \frac{0.6 - 0.7 \times 0.65}{\sqrt{1 - (0.7)^2}\sqrt{1 - (0.65)^2}}$$

$$= \frac{0.6 - 0.455}{\sqrt{.4816 \times 0.5775}}$$

$$= \frac{0.145}{0.543}$$

$$= 0.2670$$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}.r_{13}.r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(.6)^2 + (.7)^2 - 2(.6)(.7)(.65)}{1 - (.65)^2}}$$

$$= \sqrt{\frac{0.9604 + 0.1936 - 0.4656}{1 - (.52)^2}}$$

$$= \sqrt{\frac{.304}{.5775}}$$

$$= 0.726 \ .$$

Example 11.20: Following table shows the correlation matrix of three variable x_1 (Height), x_2 (Weight) and x_3 (Diameter of Chest) of 10 randomly selected players:

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{bmatrix} 1.0000 & 0.8630 & 0.6480 \\ & 1.0000 & 0.7090 \\ & & 1.0000 \end{bmatrix} \end{array}$$

Calculate $r_{12.3}$ and $R_{1.23}$.

Solution: Given: $r_{12} = 0.60, r_{13} = 0.70, r_{23} = 0.65,$

$$r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

$$= \frac{0.863 - 0.648 \times 0.709}{\sqrt{1 - (0.648)^2}\sqrt{1 - (0.709)^2}}$$

$$= \frac{.863 - .4594}{\sqrt{.580} \times \sqrt{.497}}$$

$$= \frac{.4036}{\sqrt{.2883}}$$

$$= \frac{.4036}{.537}$$

$$= 0.752$$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}.r_{13}.r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(.863)^2 + (.648)^2 - 2(.863)(.648)(.709)}{1 - (.709)^2}}$$

$$= \sqrt{\frac{.745 + .42 - .793}{.497}}$$

$$= 0.865 \ .$$

11.5 SUMMARY

In total, multiple correlation is the study of the relationship among three or more variables and it measures the combined influence of two or more independent variables on a single dependent variable. The coefficient of multiple correlation is a measure of how well a given variable can be predicted using a linear function of a set of other variables. It is measured by the coefficient of determination, but under the particular assumption that that the best possible linear predictors are used, whereas the coefficient of determination is defined for more general cases. The coefficient of multiple determination takes values between zero and one; a higher value indicates a better predictability of the dependent variables from the independent variables, with a value of one indicating that the predictions are exact and a value of zero indicating that no linear combination of dependent variables is better than the simpler predictor which consists of mean of the target variable. Further, the partial correlation is the simple correlation between two variables after eliminating the influence of the third variable among them.

11.6 GLOSSARY

Correlation- is the process of establishing a relationship or connection between two or more things.

11.7 CHECK YOUR PROGRESS

1. In a trivariate distribution, it is found that
$$r_{12} = 0.41, r_{13} = 0.71, r_{23} = 0.5$$
Find that value of $r_{23.1}$ and $r_{13.2}$.

2. If $r_{12} = 0.7, r_{13} = 0.61, r_{23} = 0.4,$, find the value of $r_{12.3}, r_{13.2}$ and $r_{23.1}$.

3. Is it possible to have the following experimental data:
$$r_{12} = 0.6, r_{23} = 0.8, r_{31} = -0.5$$

4. In a trivariate distribution $r_{23} = .2, r_{13} = .5, r_{12} = .6$.Compute $r_{12.3}$ and $R_{1.23}$.

5. Suppose a computer has found for a given set of values of x_1, x_2, x_3: $r_{12} = 0.91, r_{13} = 0.33$ and $r_{23} = 0.81$. Explain whether these computations may be said to be free from errors.

11.8 ANSWERS TO CHECK YOUR PROGRESS

1. $[r_{23.1} = 0.325, r_{13.2} = 0.639]$
2. $[\ r_{12.3} = 0.629, r_{13.2} = 0.505, r_{23.1} = 0.048]$
3. $[\ r_{12.3} = 1.92,$ Inconsistency]
4. $[r_{12.3} = 0.47, R_{1.23} = 0.714]$
5. $[r_{12.3} = 1.161;$ Not free from errors]

11.9 TERMINAL QUESTIONS

1. Explain multiple correlation.
2. Explain partial correlation.
3. Explain the relation found between simple, partial, and multiple correlation coefficients.

11.10 SUGGESTED READINGS

1. Basic Statistics – Goon, Guptha and Dasguptha – World Press Limited – Calcutta.
2. Fundamentals of Business Statistics – Sanchethi and Kappor.
3. Quantitative Methods in Management – Srivastava, Shenoy and Guptha.
4. Business Statistics – Guptha and Guptha

UNIT: 12 MULTIPLE REGRESSION ANALYSIS

Structure

OBJECTIVES

After studying this unit, you will be able to:

(i) Define the concept of multiple regression.
(ii) Prepare multiple Regression equations using Normal Equations;
(iii) prepare a Multiple Regression Equation in terms of Simple Correlation Coefficients; and
(iv) to understand the concept of the standard error of estimate for multiple regression.

12.1 INTRODUCTION

In the previous unit, you have studied that partial and multiple correlation are extension of the technique of simple correlation. Similarly, multiple regression is also an extension of the technique of simple regression, under which you will study the interrelationship between three or more variables. It is also used to estimate the most probable value of the dependent variable for given values of the independent variables.

12.2 MULTIPLE REGRESSION

In multiple regression, you will study three variables, and you may consider one variable as the dependent variable and the other two as independent variables.

The multiple regression equations can be worked out by the following methods:

12.2.1 Multiple Regression Equation using Normal Equations: This method is also called as Least Square Method. Under this method, computation of regression equations is done by solving three normal equations, e.g.

Multiple regression equation of X_1 on X_2 and X_3 is given by:

$$X_1 = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3$$

Where, X_1 = Dependent variable, X_2 and X_3 = Independent variable.
$b_{12.3}$ and $b_{13.2}$ = Partial regression coefficients.

Using least square method, the values of constants $a_{1.23}, b_{12.3}$ and $b_{13.2}$ are obtained by solving the following three normal equations:

$$\Sigma X_1 = N.a_{1.23} + b_{12.3}\Sigma X_2 + b_{13.2}\Sigma X_3$$
$$\Sigma X_1X_2 = a_{1.23}\Sigma X_2 + b_{12.3}\Sigma X_2^2 + b_{13.2}\Sigma X_2X_3$$
$$\Sigma X_1X_3 = a_{1.23}\Sigma X_3 + b_{12.3}\Sigma X_2X_3 + b_{13.2}\Sigma X_3^2$$

Similarly, the multiple regression equations of X_2 on X_1 and X_3 and X_3 ; on X_1 and X_2 and their normal equations can also be written as:

Multiple Regression Equation of X_2 on X_1 and X_3 is given by:

$$X_2 = a_{2.13} + b_{21.3}X_1 + b_{23.1}X_3$$

Three Normal Equations are:

$$\Sigma X_2 = Na_{2.13} + b_{21.3}\Sigma X_1 + b_{23.1}\Sigma X_3$$
$$\Sigma X_2X_1 = a_{2.13}\Sigma X_1 + b_{21.3}\Sigma X_1^2 + b_{23.1}\Sigma X_3X_1$$
$$\Sigma X_2X_3 = a_{2.13}\Sigma X_3 + b_{21.3}\Sigma X_1X_3 + b_{23.1}\Sigma X_3^2$$

Multiple Regression Equation of X_3 on X_1 and X_2 is given by:

$$X_3 = a_{3.12} + b_{31.2}X_1 + b_{32.1}X_2$$

The Three Normal Equations are :

$$\Sigma X_3 = Na_{3.12} + b_{31.2}\Sigma X_1 + b_{32.1}\Sigma X_2$$
$$\Sigma X_3X_1 = a_{3.12}\Sigma X_1 + b_{31.2}\Sigma X_1^2 + b_{32.1}\Sigma X_2X_1$$
$$\Sigma X_3X_2 = a_{3.12}\Sigma X_2 + b_{31.2}\Sigma X_1X_2 + b_{32.1}\Sigma X_2^2$$

Example 12.1: For the following set of data, calculate the multiple regression equation of X_1 on X_2 and X_3 :

X_1:	4	6	7	9	13	15
X_2:	15	12	8	6	4	3
X_3:	30	24	20	14	10	4

Solution: The regression equation of X_1 on X_2 and X_3 is

$$X_1 = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3$$

The three normal equations are:

$$\Sigma X_1 = Na_{1.23} + b_{12.3}\Sigma X_2 + b_{13.2}\Sigma X_3$$
$$\Sigma X_1 X_2 = a_{1.23}\Sigma X_2 + b_{12.3}\Sigma X_2^2 + b_{13.2}\Sigma X_3 X_2$$
$$\Sigma X_1 X_3 = a_{1.23}\Sigma X_3 + b_{12.3}\Sigma X_2 X_3 + b_{13.2}\Sigma X_3^2$$

X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	X_2^2	X_3^2
4	15	30	60	120	450	225	900
6	12	24	72	144	288	144	576
7	8	20	56	140	160	64	400
9	6	14	54	126	84	36	196
13	4	10	52	130	40	16	100
15	3	4	45	60	12	9	16
$\sum X_1 = 54$	$\sum X_2 = 48$	$\sum X_3 = 102$	$\sum X_1X_2 = 339$	$\sum X_1X_3 = 720$	$\sum X_2X_3 = 1034$	$\sum X_2^2 = 494$	$\sum X_3^2 = 2188$

Substituting the values in the normal equations:

$$54 = 6a_{1.23} + 48b_{12.3} + 102b_{13.2}$$
$$339 = 48a_{1.23} + 494b_{12.3} + 1034b_{13.2}$$
$$720 = 102a_{1.33} + 1034b_{12.3} + 2188b_{13.2}$$

Multiplying (i) by 8, you get

$$432 = 48a_{1.23} + 384b_{12.3} + 816b_{13.2}$$

Subtracting (ii) from (iv), you get

$$-93 = 110b_{12.3} + 218b_{13.2}$$

Multiplying (i) by 17, you get

$$918 = 102a_{1.23} + 816b_{12.3} + 1734b_{13.2}$$

Subtracting (iii) from (vi), you get

$$-198 = 218b_{12.3} + 454b_{13.2}$$

Multiplying (v) by 109, you get

$$-10137 = 11990b_{12.3} + 23762b_{13.2}$$

Multiplying (vii) by 55, you get

$$-10890 = 11990b_{12.3} + 24970b_{13.2}$$

Subtracting (viii) from (ix), you get

$$753 = -1208b_{13.2}$$
$$b_{13.2} = -\frac{753}{1208}$$
$$= -0.623$$

Subtracting the value of $b_{13.2}$ in equation (v), you get

$$-93 = 110b_{12.3} + 218(-0.623)$$

$$135.814 - 93 = 110b_{12.3}$$

$$b_{12.3} = \frac{42.814}{110}$$
$$= 0.389$$

Subtracting the value of $b_{12.3}$ and $b_{13.2}$ in equation (i), you get

$$6a_{1.23} + 48(0.389) + 102(-0.623) = 54$$
$$6a_{1.23} + 18.672 - 63.546 = 54$$
$$6a_{1.23} = 54 - 18.672 + 63.546$$
$$6a_{1.23} = 98.874$$

$$a_{1.23} = \frac{98.874}{6}$$
$$= 16.479$$

Short-Cut Method: If the sizes of the values of variables are very large, then the above system of solving normal equations becomes a very tedious process. In such a case, in place of actual values, deviations from the means of the variables are used to simplify the computational procedure.

Multiple Regression Equation of X_1 on X_2 and X_3 in deviation is given by:

$$X_1 - \bar{X}_1 = b_{12.3}(X_2 - \bar{X}_2) + b_{13.2}(X_3 - \bar{X}_3)$$

or

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3$$

where

$$x_1 = X_1 - \bar{X}_1, x_2 = X_2 - \bar{X}_2, x_3 = X_3 - \bar{X}_3$$

The values of the partial regression coefficients ($b_{12.3}$ and $b_{13.2}$) can be obtained by solving the following two normal equations:

$$\Sigma x_1 x_2 = b_{12.3}\Sigma x_2^2 + b_{13.2}\Sigma x_3 x_2$$
$$\Sigma x_1 x_3 = b_{12.3}\Sigma x_2 x_3 + b_{13.2}\Sigma x_3^2$$

Further solved, you have

$$b_{12.3} = \frac{(\Sigma x_1 x_2)(\Sigma x_3^2) - (\Sigma x_1 x_3)(\Sigma x_2 x_3)}{(\Sigma x_2^2)(\Sigma x_3^2) - (\Sigma x_2 x_3)^2}$$

$$b_{13.2} = \frac{(\Sigma x_1 x_3)(\Sigma x_2^2) - (\Sigma x_1 x_2)(\Sigma x_3 x_2)}{(\Sigma x_3^2)(\Sigma x_2^2) - (\Sigma x_3 x_2)^2}$$

Similarly, the multiple regression equations of X_2 on X_1 and X_3 ; and X_3 on X_1 and X_2 and their normal equations can also be written.

Multiple Regression Equation of X_2 on X_1 and X_3 in deviation from is given by:

$$X_2 - \bar{X}_2 = b_{21.3}(X_1 - \bar{X}_1) + b_{23.1}(X_3 - \bar{X}_3)$$

or

$$x_2 = b_{21.3}x_1 + b_{23.1}x_3$$

Two normal equations are:

$$\Sigma x_2 x_1 = b_{21.3}\Sigma x_1^2 + b_{23.1}\Sigma x_1 x_3$$
$$\Sigma x_2 x_3 = b_{21.3}\Sigma x_1 x_3 + b_{23.1}\Sigma x_3^2$$

Further solved, you have

$$b_{21.3} = \frac{(\Sigma x_2 x_1)(\Sigma x_3^2) - (\Sigma x_2 x_3)(\Sigma x_1 x_3)}{(\Sigma x_1^2)(\Sigma x_3^2) - (\Sigma x_1 x_3)^2}$$

$$b_{23.1} = \frac{(\Sigma x_2 x_3)(\Sigma x_1^2) - (\Sigma x_2 x_1)(\Sigma x_3 x_1)}{(\Sigma x_3^2)(\Sigma x_1^2) - (\Sigma x_3 x_1)^2}$$

Multiple Regression Equation of X_3 on X_1 and X_2 in deviation from is given by:

$$X_3 - \bar{X}_3 = b_{31.2}(X_1 - \bar{X}_1) + b_{32.1}(X_2 - \bar{X}_2)$$

or

$$x_3 = b_{31.2} x_1 + b_{32.1} x_2$$

Two normal equations are:

$$\Sigma x_1 x_3 = b_{31.2} \Sigma x_1^2 + b_{23.1} \Sigma x_1 x_2$$
$$\Sigma x_2 x_3 = b_{31.2} \Sigma x_1 x_2 + b_{32.1} \Sigma x_2^2$$

Further solved, you have

$$b_{31.2} = \frac{(\Sigma x_3 x_1)(\Sigma x_2^2) - (\Sigma x_3 x_2)(\Sigma x_1 x_2)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$$b_{32.1} = \frac{(\Sigma x_3 x_2)(\Sigma x_1^2) - (\Sigma x_3 x_1)(\Sigma x_2 x_1)}{(\Sigma x_2^2)(\Sigma x_1^2) - (\Sigma x_2 x_1)^2}$$

Example 12.2: From the following data, find the least square regression of X_3 on X_1 and X_2 using actual mean method. Also estimate X_3 when $X_1 = 10$ and $X_2 = 6$.

X_1:	3	5	6	8	12	14
X_2:	16	10	7	4	3	2
X_3:	90	72	54	42	30	12

Solution:

X_1	$x_1 = (X_1 - \bar{X}_1)$	x_1^2	X_2	$x_2 = (X_2 - \bar{X}_2)$	x_2^2	X_3	$x_3 = (X_3 - \bar{X}_3)$	x_3^2	x_1x_2	x_1x_3	x_2x_3
3	-5	25	16	+9	81	90	+40	1600	-45	-200	+360
5	-3	9	10	+3	9	72	+22	484	-9	-66	+66
6	-2	4	7	0	0	54	+4	16	0	-8	0
8	0	0	4	-3	9	42	-8	64	0	8	+24
12	+4	16	3	-4	16	30	-20	400	-16	-80	+80
14	+6	36	2	-5	25	12	-38	1444	-30	-228	+190
$\Sigma X_1 = 48$	$\Sigma x_1 = 0$	$\Sigma x_1^2 = 90$	$\Sigma X_2 = 42$	$\Sigma x_2 = 0$	$\Sigma x_2^2 = 140$	$\Sigma X_3 = 300$	$\Sigma x_3 = 0$	$\Sigma x_3^2 = 4008$	$\Sigma x_1x_2 = 100$	$\Sigma x_1x_3 = -582$	$\Sigma x_2x_3 = 720$

$$\bar{X}_1 = \frac{48}{6} = 8, \bar{X}_2 = \frac{42}{6} = 7, \bar{X}_3 = \frac{300}{6} = 50,$$

Regression Equation of X_3 on X_1 and X_2 is:

$$X_3 - \bar{X}_3 = b_{31.2}(X_1 - \bar{X}_1) + b_{32.1}(X_2 - \bar{X}_2)$$

$$b_{31.2} = \frac{(\Sigma x_3 x_1)(\Sigma x_2^2) - (\Sigma x_3 x_2)(\Sigma x_1 x_2)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$$= \frac{(-582)(140) - (720)(-100)}{(90)(140) - (-100)^2}$$

$$= \frac{-81480 + 72000}{12600 - 10000} = \frac{-9480}{2600} = -3.646$$

$$b_{32.1} = \frac{(\Sigma x_3 x_2)(\Sigma x_1^2) - (\Sigma x_3 x_1)(\Sigma x_2 x_1)}{(\Sigma x_2^2)(\Sigma x_1^2) - (\Sigma x_2 x_1)^2}$$

$$= \frac{(720)(90) - (-582)(-100)}{(90)(140) - (-100)^2}$$

$$= \frac{64800 + 58200}{12600 - 10000} = \frac{6600}{2600} = 2.538$$

Substituting the values in the above equations, you get

$$X_3 - 50 = -3.646(X_1 - 8) + 2.538(X_2 - 7)$$

$$X_3 - 50 = -3.646X_1 + 29.168 + 2.538X_2 - 17.766$$

$$X_3 = -3.646X_1 + 2.538X_2 + 61.402$$

When $X_1 = 10$ and $X_2 = 6$, So, $X_3 = -3.646(10) + 2.538(6) + 61.402$

$$= -36.46 + 15.228 + 61.402 = 40.17 \ \ or \ \ 40.$$

Example 12.3: Given the following information (variables are measured from their respective means):

$$\Sigma x_1 x_2 = 720, \Sigma x_2 x_3 = -582, \Sigma x_1 x_3 = -100$$

$$\Sigma x_2^2 = 4008, \Sigma x_3^2 = 90, \Sigma x_1^2 = 140$$

$$\overline{X}_1 = 7, \overline{X}_2 = 50, \overline{X}_3 = 8$$

Find the multiple regression equation of X_1 on X_2 and X_3. Estimate X_1 when $X_2 = 10$ and $X_3 = 95$.

Solution: Regression Equation of X_1 on X_2 and X_3 is given by.

$$X_1 - \overline{X}_1 = b_{12.3}(X_2 - \overline{X}_2) + b_{13.2}(X_3 - \overline{X}_3)$$

$$b_{12.3} = \frac{(\Sigma x_1 x_2)(\Sigma x_3^2) - (\Sigma x_1 x_3)(\Sigma x_2 x_3)}{(\Sigma x_2^2)(\Sigma x_3^2) - (\Sigma x_2 x_3)^2}$$

$$= \frac{(720)(90) - (-100)(-582)}{(4008)(90) - (-582)^2}$$

$$= \frac{64800 + 58200}{360720 - 338724}$$

$$= \frac{6600}{21996}$$

$$= 0.30$$

$$b_{13.2} = \frac{(\Sigma x_1 x_3)(\Sigma x_2^2) - (\Sigma x_1 x_2)(\Sigma x_3 x_2)}{(\Sigma x_3^2)(\Sigma x_2^2) - (\Sigma x_3 x_2)^2}$$

$$= \frac{(-100)(4008) - (720)(-582)}{(4008)(90) - (-582)^2}$$

$$= \frac{-400800 + 419040}{360720 - 338724}$$

$$= \frac{18240}{21996}$$

$$= 0.83$$

You are given: $\bar{X}_1 = 7, \bar{X}_2 = 50, \bar{X}_3 = 8$

Substituting the values in the above equation, you get

$$X_1 - 7 = 0.30(X_2 - 50) + 0.83(X_3 - 8)$$

or

$$X_1 - 7 = 0.30X_2 - 15 + 0.83X_3 - 6.64$$

$\therefore X_1 = 0.30X_2 + 0.83X_3 - 14.64$ is the required equation

When $X_2 = 20$ and $X_3 = 30$

$$X_1 = 0.30(20) + 0.83(30) - 14.64 = 6 + 24.9 - 14.64 = 16.26$$

Example 12.4: The following data for the three variables X_1, X_2 and X_3 are given below:

$$\Sigma x_1x_2 = 218, \quad \Sigma x_1x_3 = -198, \quad \Sigma x_2x_3 = -93$$

$$\Sigma x_1^2 = 454, \qquad \Sigma x_2^2 = 110, \qquad \Sigma x_3^2 = 90$$

x_1, x_2 and x_3 are measured from their means. Find the two partial regression coefficients($b_{12.3}$ *and* $b_{13.2}$).

Solution:

$$b_{12.3} = \frac{(\Sigma x_1x_2)(\Sigma x_3^2) - (\Sigma x_1x_3)(\Sigma x_2x_3)}{(\Sigma x_2^2)(\Sigma x_3^2) - (\Sigma x_2x_3)^2}$$

$$= \frac{(218)(90) - (-198)(-93)}{(90)(110) - (-93)^2}$$

$$= \frac{19620 - 18414}{9900 - 8649}$$

$$= \frac{1206}{1251}$$

$$= 0.964$$

$$b_{13.2} = \frac{(\Sigma x_1x_3)(\Sigma x_2^2) - (\Sigma x_1x_2)(\Sigma x_3x_2)}{(\Sigma x_3^2)(\Sigma x_2^2) - (\Sigma x_3x_2)^2}$$

$$= \frac{(-198)(110) - (218)(-93)}{(90)(110) - (-93)^2}$$

$$= \frac{-21780 + 20274}{9900 - 8649}$$

$$= \frac{-1506}{1251}$$

$$= -1.203$$

12.2.2 Multiple Regression Equation in terms of Simple Correlation Coefficient: When the values of $\bar{X}_1, \bar{X}_2$ and $\bar{X}_3, ó_1, ó_2$ and $ó_3$ and r_{12}, r_{13} and r_{23} are given, then the multiple regression equation is expressed in the following manner:

Multiple Regression Equation of X_1 on X_2 and X_3

$$X_1 - \bar{X}_1 = b_{12.3}(X_2 - \bar{X}_2) + b_{13.2}(X_3 - \bar{X}_3)$$

or

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3$$

where

$$x_1 = X_1 - \bar{X}_1, x_2 = X_2 - \bar{X}_2, x_3 = X_3 - \bar{X}_3$$

The values of the partial regression coefficients ($b_{12.3}$ and $b_{13.2}$) are determined by using the following formulae:

$$b_{12.3} = \left[\frac{\sigma_1}{\sigma_2}\right].\left[\frac{r_{12} - r_{13}.r_{23}}{1 - r_{23}^2}\right]$$

$$b_{13.2} = \left[\frac{\sigma_1}{\sigma_3}\right].\left[\frac{r_{13} - r_{12}.r_{32}}{1 - r_{32}^2}\right]$$

Multiple Regression Equation of X_1 on X_2 and X_3 can also be written as :

$$x_1 = \left[\frac{\sigma_1}{\sigma_2}\right].\left[\frac{r_{12} - r_{13}.r_{23}}{1 - r_{23}^2}\right]x_2 + \left[\frac{\sigma_1}{\sigma_3}\right].\left[\frac{r_{13} - r_{12}.r_{32}}{1 - r_{32}^2}\right]x_3$$

or

$$X_1 - \bar{X}_1 = \left[\frac{\sigma_1}{\sigma_2}\right].\left[\frac{r_{12} - r_{13}.r_{23}}{1 - r_{23}^2}\right](X_2 - \bar{X}_2) + \left[\frac{\sigma_1}{\sigma_3}\right].\left[\frac{r_{13} - r_{12}.r_{32}}{1 - r_{23}^2}\right](X_3 - \bar{X}_3)$$

Multiple Regression Equation of X_2 on X_1 and X_3

$$X_2 - \bar{X}_2 = b_{21.3}(X_1 - \bar{X}_1) + b_{23.1}(X_3 - \bar{X}_3)$$

or

$$x_2 = b_{21.3}x_1 + b_{23.1}x_3$$

Where,

$$b_{21.3} = \left[\frac{\sigma_2}{\sigma_1}\right].\left[\frac{r_{21} - r_{23}.r_{13}}{1 - r_{13}^2}\right]$$

$$b_{23.1} = \left[\frac{\sigma_2}{\sigma_3}\right].\left[\frac{r_{23} - r_{21}.r_{31}}{1 - r_{31}^2}\right]$$

Multiple Regression Equation of X_2 on X_1 and X_3 can also be written as :

$$x_2 = \left[\frac{\sigma_2}{\sigma_1}\right].\left[\frac{r_{21} - r_{23}.r_{13}}{1 - r_{13}^2}\right]x_1 + \left[\frac{\sigma_2}{\sigma_3}\right].\left[\frac{r_{23} - r_{21}.r_{31}}{1 - r_{31}^2}\right]x_3$$

or

$$X_2 - \bar{X}_1 = \left[\frac{\sigma_2}{\sigma_1}\right].\left[\frac{r_{21} - r_{23}.r_{13}}{1 - r_{13}^2}\right](X_1 - \bar{X}_1) + \left[\frac{\sigma_2}{\sigma_3}\right].\left[\frac{r_{23} - r_{21}.r_{31}}{1 - r_{31}^2}\right](X_3 - \bar{X}_3)$$

Multiple Regression Equation of X_3 on X_1 and X_2

$$X_3 - \bar{X}_3 = b_{31.2}(X_1 - \bar{X}_1) + b_{32.1}(X_2 - \bar{X}_2)$$

or

$$x_3 = b_{31.2}x_1 + b_{32.1}x_2$$

Where,

$$b_{31.2} = \left[\frac{\sigma_3}{\sigma_1}\right].\left[\frac{r_{31} - r_{32}.r_{12}}{1 - r_{12}^2}\right]$$

$$b_{32.1} = \left[\frac{\sigma_3}{\sigma_2}\right].\left[\frac{r_{32} - r_{31}.r_{21}}{1 - r_{21}^2}\right]$$

Multiple regression on X_3 on X_1 and X_2 can also be written as:

$$x_3 = \left[\frac{\sigma_3}{\sigma_1}\right].\left[\frac{r_{31} - r_{32}.r_{12}}{1 - r_{12}^2}\right]x_1 + \left[\frac{\sigma_3}{\sigma_2}\right].\left[\frac{r_{32} - r_{31}.r_{21}}{1 - r_{21}^2}\right]x_2$$

or

$$X_3 - \bar{X}_3 = \left[\frac{\sigma_3}{\sigma_1}\right].\left[\frac{r_{31} - r_{32}.r_{12}}{1 - r_{12}^2}\right](X_1 - \bar{X}_1) + \left[\frac{\sigma_3}{\sigma_2}\right].\left[\frac{r_{32} - r_{31}.r_{21}}{1 - r_{21}^2}\right](X_2 - \bar{X}_2)$$

Note: $r_{12} = r_{21}, r_{23} = r_{32}, r_{13} = r_{31}$

Example 12.5: A teacher in mathematics wishes to determine the relationship of marks in the final examination to those in two tests given during the semester. Calling X_1, X_2 and X_3, the marks of a student on the 1st, 2nd, and final examinations, respectively, he made the following computations from a total of 120 students:

$$\bar{X}_1 = 6.8 \quad \bar{X}_2 = 7.0 \quad \bar{X}_3 = 74$$
$$\sigma_1 = 1.0 \quad \sigma_2 = 0.80 \quad \sigma_3 = 9.0$$
$$r_{12} = 0.60 \quad r_{13} = 0.70 \quad r_{23} = 0.65$$

(i) Find the relevant regression equation.

(II) Estimate the final marks of two students who secured respectively 9 and 7,4 and 8 on the two tests.

Solution: The relevant least square regression equation will be X_3 on X_1 and X_2 which is given by:

$$X_3 - \bar{X}_3 = b_{31.2}(X_1 - \bar{X}_1) + b_{32.1}(X_2 - \bar{X}_2)$$

$$b_{31.2} = \frac{\sigma_3}{\sigma_1}.\left[\frac{r_{31} - r_{32}.r_{12}}{1 - r_{12}^2}\right]$$

$$= \frac{9}{1} \times \left[\frac{(.70) - (.65)(.60)}{1 - (.60)^2}\right]$$

$$= 9 \times \left[\frac{(.70) - (.39)}{1 - (.60)^2}\right] = 9 \times \left[\frac{.31}{.64}\right] = \frac{2.79}{.64} = 4.36$$

$$b_{32.1} = \frac{\sigma_3}{\sigma_2}.\left[\frac{r_{32} - r_{31}.r_{21}}{1 - r_{21}^2}\right]$$

$$= \frac{9}{.80} \times \left[\frac{(.65) - (.70)(.60)}{1 - (.60)^2}\right]$$

$$= \frac{9}{.80} \times \left[\frac{.65 - .42}{0.64}\right]$$

$$= 4.04$$

Thus, the regression equation of X_3 on X_1 and X_2 is

$$X_3 - 74 = 4.36(X_1 - 6.8) + 4.04(X_2 - 7)$$
$$\therefore \quad X_3 = 16.07 + 4.36X_1 + 4.04X_2$$

Final marks of students who scored 9 and 7 marks:

When $X_1 = 9$ and $X_2 = 7$

$$X_3 = 16.07 + 4.36(9) + 4.04(7)$$
$$= 16.07 + 39.24 + 28.28 = 83.59 \ or\ 84$$

Final marks of students who scored 4 and 8 marks:

When $X_1 = 4$ and $X_2 = 8$

$$X_3 = 16.07 + 4.36(4) + 4.04(8)$$
$$= 16.07 + 17.44 + 32.32 = 65.8 \ or\ 66$$

Example 12.6: Given the following, determine the regression equations of:

(I) x_1 on x_2 and x_3 and

(II) x_2 on x_1 and x_3 when the variates are measured from their means:

$$r_{12} = 0.8 \quad r_{13} = 0.6 \quad r_{23} = 0.5$$
$$\sigma_1 = 10 \quad \sigma_2 = 8 \quad \sigma_3 = 5$$

Solution: (I) The regression equation of x_1 on x_2 and x_3 when the variates are measured from their means is given by:

$$x_1 = b_{12.3}x_2 + b_{13.2}x_2$$

where

$$x_1 = X_1 - \bar{X}_1, x_2 = X_2 - \bar{X}_2, x_3 = X_3 - \bar{X}_3$$

$$b_{12.3} = \left[\frac{\sigma_1}{\sigma_2}\right] \times \left[\frac{r_{12} - r_{13}.r_{23}}{1 - r_{23}^2}\right]$$

$$= \left[\frac{10}{8}\right] \times \left[\frac{(0.8) - (0.6)(0.5)}{1 - (0.5)^2}\right] = 0.833$$

$$b_{13.2} = \left[\frac{\sigma_1}{\sigma_3}\right] \times \left[\frac{r_{13} - r_{12}.r_{32}}{1 - r_{32}^2}\right]$$

$$= \left[\frac{10}{5}\right] \times \left[\frac{(0.6) - (0.8)(0.5)}{1 - (0.5)^2}\right] = 0.533$$

$\therefore$ The required regression equation is :

$$x_1 = .833x_2 + .533x_3$$

(II) The regression equation of x_2 on x_1 and x_3 when the variates are measured from their means is given by :

$$x_2 = b_{21.3}x_1 + b_{23.1}x_3$$

$$b_{21.3} = \left[\frac{\sigma_2}{\sigma_1}\right] \times \left[\frac{r_{21} - r_{23}.r_{13}}{1 - r_{13}^2}\right]$$

$$= \left[\frac{8}{10}\right] \times \left[\frac{(0.8) - (.5)(.6)}{1 - (.6)^2}\right] = .625$$

$$b_{23.1} = \left[\frac{\sigma_2}{\sigma_3}\right] \times \left[\frac{r_{23} - r_{21}.r_{31}}{1 - r_{31}^2}\right]$$

$$= \left[\frac{8}{5}\right] \times \left[\frac{(.5) - (0.8)(0.6)}{1 - (.6)^2}\right] = 0.05$$

∴ The required regression equation is :

$$x_2 = .625x_1 + .05\, x_3$$

Example 12.7: A random sample of 15 students of the Basic Statistics course when observed for weights (X_1), age (X_2) and height (X_3) offered the following information:

$$r_{12} = 0.8, r_{23} = 0.3, r_{13} = 0.5, S_1 = 8.5, S_2 = 4.5, S_3 = 2.1$$
$$\bar{X}_1 = 70kg, \bar{X}_2 = 22\ years\ and \bar{X}_3 = 160cms.$$

Obtain:

(I) Multiple and partial correlation coefficients $R_{1.23}$ and $r_{13.2}$.

(II) Multiple regression of X_1 on X_2 and X_3 and estimate the value of X_1 for $X_2 = 25\ yrs$ and $X_3 = 140cms$.

Solution: (I)

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}.r_{13}.r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.8)^2 + (0.5)^2 - 2(0.8)(0.5)(0.3)}{1 - (0.3)^2}}$$

$$= \sqrt{\frac{0.64 + 0.25 - 0.24}{0.91}} = \sqrt{\frac{0.65}{0.91}} = 0.8452$$

$$r_{12.3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

$$= \frac{(0.8) - (0.5)(0.3)}{\sqrt{1 - (0.5)^2}\sqrt{1 - (0.3)^2}}$$

$$= \frac{.8 - 0.15}{\sqrt{.75} \times \sqrt{0.91}}$$

$$= \frac{0.65}{0.8261} = 0.7868$$

(II) Multiple Regression on X_1 on X_2 and X_3 is given by:

$$X_1 - \bar{X}_1 = b_{12.3}(X_2 - \bar{X}_2) + b_{13.2}(X_3 - \bar{X}_3)$$

$$b_{12.3} = \left[\frac{S_1}{S_2}\right] \times \left[\frac{r_{12} - r_{13}.r_{23}}{1 - r_{23}^2}\right]$$

$$= \left[\frac{8.5}{4.5}\right] \times \left[\frac{(0.8) - (0.5)(0.3)}{1 - (.3)^2}\right] = 1.349$$

$$b_{13.2} = \left[\frac{S_1}{S_3}\right] . \left[\frac{r_{13} - r_{12} . r_{32}}{1 - r_{32}^2}\right]$$

$$= \left[\frac{8.5}{2.1}\right] \times \left[\frac{(0.5) - (0.8)(0.3)}{1 - (.3)^2}\right] = 1.156$$

Substituting the values in the above equation, you get

$$X_1 - 70 = 1.349(X_2 - 22) + 1.156(X_3 - 160)$$
$$X_1 - 70 = 1.349X_2 - 29.678 + 1.156X_3 - 184.96$$
$$\therefore X_1 = 1.349X_2 + 1.156X_3 - 144.638$$

Estimation of X_1 for $X_2 = 25$ and $X_3 = 140$:

When $X_2 = 25$ and $X_3 = 140$,

$$X_1 = 1.349(25) + 1.156(140) - 144.638$$
$$= 33.725 + 161.84 - 144.638 = 50.927$$

Example 12.8: In a trivarite distribution:

$$\sigma_1 = 3, \sigma_2 = 4, \sigma_3 = 5$$
$$r_{23} = 0.4, r_{31} = 0.6, r_{12} = 0.7$$

(I) Compute $r_{23.1}$ and $R_{1.23}$

(II) Determine the regression equation of x_1 on x_2 and x_3 if the variates are measured from their means:

Solution: (I)

$$r_{23.1} = \frac{r_{23} - r_{21} . r_{31}}{\sqrt{1 - r_{21}^2}\sqrt{1 - r_{31}^2}}$$

$$= \frac{(0.4) - (0.7)(0.6)}{\sqrt{1 - (0.7)^2}\sqrt{1 - (0.6)^2}}$$

$$= \frac{0.4 - (.42)}{\sqrt{0.51}\sqrt{0.64}}$$

$$= \frac{-0.02}{0.5713} = 0.035$$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12} . r_{13} . r_{23}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.7)^2 + (0.6)^2 - 2(0.7)(0.6)(0.4)}{1 - (0.4)^2}}$$

$$= \sqrt{\frac{0.49 + 0.36 - 0.336}{0.84}}$$

$$= \sqrt{\frac{0.514}{0.84}} = 0.782$$

(II) The regression equation of x_1 on x_2 and x_3 when the variates are measured from their means are given by:

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3$$

where

$$x_1 = X_1 - \bar{X}_1, x_2 = X_2 - \bar{X}_2, x_3 = X_3 - \bar{X}_3$$

$$b_{12.3} = \left[\frac{\sigma_1}{\sigma_2}\right] \times \left[\frac{r_{12} - r_{13}.r_{23}}{1 - r_{23}^2}\right]$$

$$= \left[\frac{3}{4}\right] \times \left[\frac{(0.7) - (0.6)(0.4)}{1 - (0.4)^2}\right]$$

$$= \frac{0.75 \times 0.46}{0.84} = \frac{0.345}{0.84} = 0.41$$

$$b_{13.2} = \left[\frac{\sigma_1}{\sigma_3}\right] \times \left[\frac{r_{13} - r_{12}.r_{32}}{1 - r_{32}^2}\right]$$

$$= \left[\frac{3}{5}\right] \times \left[\frac{(0.6) - (0.7)(0.4)}{1 - (0.4)^2}\right]$$

$$= \frac{0.6 \times 0.32}{0.84} = \frac{0.192}{0.84} = 0.229$$

Thus, the required regression equation is:

$$x_1 = 0.41x_2 + 0.229x_3$$

12.3 STANDARD ERROR OF ESTIMATE (OR RELIABILITY OF ESTIMATES) FOR MULTIPLE REGRESSION

The standard error of estimate measures the reliability of the estimates given by the multiple regression equation. It shows to what extent the estimated values given by the regression equations are closer to the actual values.

For three regression equations, there are three standard errors or estimates:

(I) Standard Error of Estimate of X_1 on X_2 and X_3 ($S_{1.23}$)
(II) Standard Error of Estimate of X_2 on X_1 and X_3 ($S_{2.13}$)
(III) Standard Error of Estimate of X_3 on X_1 and X_2 ($S_{3.12}$)

The formulae for calculating the standard error of estimates are given as follows:

$$S_{1.23} = \sigma_1 . \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}.r_{13}.r_{23}}{1 - r_{23}^2}}$$

$$S_{2.13} = \sigma_2 . \sqrt{\frac{1 - r_{21}^2 - r_{23}^2 - r_{13}^2 + 2r_{21}.r_{23}.r_{13}}{1 - r_{13}^2}}$$

$$S_{3.12} = \sigma_3 . \sqrt{\frac{1 - r_{31}^2 - r_{32}^2 - r_{13}^2 + 2r_{31}.r_{32}.r_{12}}{1 - r_{12}^2}}$$

Example 12.9: If $r_{12} = 0.8, r_{13} = 0.5, r_{23} = 0.3$ and $S_1 = 8.5$, compute the standard error of estimate of X_1 on X_2 and X_3 .

Solution: Standard Error of Estimate of X_1 on X_2 and X_3 is given by :

$$S_{1.23} = \sigma_1 . \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}.r_{13}.r_{23}}{1 - r_{23}^2}}$$

$$= 8.5 \sqrt{\frac{1 - (0.8)^2 - (0.5)^2 - (0.3)^2 + 2(0.8)(0.5)(0.3)}{1 - (0.3)^2}}$$

$$= 4.543$$

12.4 COEFFICIENT OF MULTIPLE DETERMINATION (R^2)

The coefficient of determination in multiple regression is denoted by $R_{1.23}^2$ is similar to the coefficient of determination r^2 in the simple linear regression. It represents the proportion (fraction) of the total variation in the dependent variable X_1 that has been explained by the independent variables (X_1 *and* X_2) in the multiple regression equation.

For example, if $R_{1.23} = 0.7252$, then $R_{1.23}^2 = 0.5259 = 0.526$

The value of $R_{1.23}^2 = 0.526$ indicates that 52.6% variation in the dependent variable X_1 are explained by the independent variables X_2 and X_3 in the multiple regression equation of X_1 on X_2 and X_3.

Example 12.10: A random sample of 15 students of an advanced course in statistics when observed for weight (X_1), age (X_2) and height (X_3) offered the following information:

$r_{12} = 0.8, r_{13} = 0.5, r_{23} = 0.3$

$S_1 = 8.5, S_1 = 4.5$ *and* $S_3 = 2.1$

Find the following:

(a) Partial regression coefficient $b_{1.23}$ and $b_{13.2}$.

(b) Standard error of estimate $S_{1.23}$.

(c) Correlation Coefficients $R_{1.23}$ and $r_{12.3}$.

(d) Multiple regression of X_1 on X_2 and X_3 when $\bar{X}_1 = 70kg, \bar{X}_2 = 22$ years and $\bar{X}_3 = 150cm$.

(e) Weight of a student (X_1) of 25 years of age and 140 cm in height.

Solution: Given:

$$r_{12} = 0.8, r_{13} = 0.5, r_{23} = 0.3$$

$$S_1 = 8.5, S_1 = 4.5 \; and \; S_3 = 2.1$$

(a)

$$b_{12.3} = \left[\frac{S_1}{S_2}\right].\left[\frac{r_{12}.r_{13}.r_{23}}{1-r_{23}^2}\right]$$

$$= \left[\frac{8.5}{4.5}\right] \times \left[\frac{0.8-(0.5)(0.3)}{1-(0.3)^2}\right] = 1.349$$

$$b_{13.2} = \left[\frac{S_1}{S_3}\right].\left[\frac{r_{13}.r_{12}.r_{32}}{1-r_{32}^2}\right]$$

$$= \left[\frac{8.5}{2.1}\right] \times \left[\frac{(0.5)-(0.8)(0.3)}{1-(0.3)^2}\right] = 1.156$$

(b) $S_{1.23} = S_1 \sqrt{\frac{1-r_{12}^2-r_{13}^2-r_{23}^2+2r_{12}.r_{13}.r_{23}}{1-r_{23}^2}}$

$$= 8.5 \times \sqrt{\frac{1-(0.8)^2-(0.5)^2-(0.3)^2+2(0.8)(0.5)(0.3)}{1-(0.3)^2}}$$

$$= 8.5 \times \sqrt{\frac{1-.64-.25-0.09+0.24}{0.91}}$$

$$= 4.543$$

(c)

$$R_{1.23} = \sqrt{\frac{r_{12}^2+r_{13}^2-2r_{12}.r_{13}.r_{23}}{1-r_{23}^2}}$$

$$= \sqrt{\frac{(0.8)^2+(0.5)^2-2(0.8)(0.5)(0.3)}{1-(0.3)^2}}$$

$$= \sqrt{\frac{0.64+.25-0.24}{0.91}}$$

$$= \sqrt{\frac{0.65}{0.91}}$$

$$= 0.8452$$

$$r_{12.3} = \frac{r_{12}.r_{13}.r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

$$= \frac{(0.8)-(0.5).(0.3)}{\sqrt{1-(0.5)^2}\sqrt{1-(0.3)^2}}$$

$$= \frac{0.65}{0.8261} = 0.7868$$

(d) Multiple Regression Equation of X_1 on X_2 and X_3 :

$$X_1 - \bar{X}_1 = b_{12.3}(X_2 - \bar{X}_2) + b_{13.2}(X_3 - \bar{X}_3)$$

Substituting the values, you have

$$X_1 - 70 = 1.349(X_2 - 22) + 1.156(X_3 - 150)$$

$$X_1 = -133.078 + 1.349X_2 + 1.156X_3$$

(e) For $X_2 = 25$ and $X_3 = 140$,

$$X_1 = -133.078 + 1.349(25) + 1.156(140)$$

$$= -133.078 + 33.725 + 161.84 = 62.487$$

Example 12.11: Given the following data, determine the regression equation of x_1 on x_2 and x_3 if the variates are measured from means:

$$r_{12} = 0.8, \quad r_{13} = 0.6, \quad r_{23} = 0.5$$

$$\sigma_1 = 10, \quad \sigma_2 = 8, \quad \sigma_3 = 15$$

Also, find the standard error of the estimate of x_1 on x_2 and x_3.

Solution: The regression equation of x_1 on x_2 and x_3 is :

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3$$

where,

$$x_1 = X_1 - \bar{X}_1, x_2 = X_2 - \bar{X}_2 \ and \ x_3 = X_3 - \bar{X}_3$$

Here,

$$b_{12.3} = \left[\frac{\sigma_1}{\sigma_2}\right].\left[\frac{r_{12}.r_{13}.r_{23}}{1 - r_{23}^2}\right]$$

$$= \left[\frac{10}{8}\right] \times \left[\frac{0.8 - (0.6)(0.5)}{1 - (0.5)^2}\right]$$

$$= \left[\frac{10}{8}\right] \times \left[\frac{0.8 - .30}{1 - .25}\right]$$

$$= \frac{10}{8} \times \frac{0.50}{0.75} = 0.833$$

$$b_{13.2} = \left[\frac{\sigma_1}{\sigma_3}\right].\left[\frac{r_{13}.r_{12}.r_{32}}{1 - r_{32}^2}\right]$$

$$= \frac{10}{5} \times \left[\frac{(0.6) - (0.8)(0.5)}{1 - (0.5)^2}\right]$$

$$= \frac{10}{5} \times \frac{0.20}{0.75}$$

$$= \frac{2}{3.75} = 0.53$$

Thus, regression equation of x_1 on x_2 and x_3 is:

$$x_1 = 0.833x_2 + 0.53x_3$$

Standard Error of Estimate of X_1 on X_2 and X_3

$$S_{1.23} = \sigma_1 \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}.r_{13}.r_{23}}{1 - r_{23}^2}}$$

$$= 10. \sqrt{\frac{1 - (0.8)^2 - (0.6)^2 - (0.5)^2 + 2(0.8)(0.6)(0.5)}{1 - (0.5)^2}}$$

$$= 10.\sqrt{\frac{1 - .64 - .36 - .25 + .48}{0.75}}$$

$$= 10.\sqrt{\frac{.23}{0.75}}$$

$$= 10 \times .5537 = 5.537$$

Example 12.12: The following values have been obtained from the measurement of three variables x_1, x_2 and x_3 :

$$\bar{X}_1 = 6.8 \quad \bar{X}_2 = 7.0 \quad \bar{X}_3 = 7.4$$
$$S_1 = 1.0 \quad S_2 = 0.80 \quad S_3 = 0.90$$
$$r_{12} = 0.60 \quad r_{13} = 0.70 \quad r_{23} = 0.65$$

(I) Obtain the regression equation of X_1 on X_2 and X_3.
(II) Estimate the value of X_1 for $X_2 = 10$ and $X_3 = 9$.
(III) Find the coefficients of multiple determination $R^2_{1.23}$ from r_{12} and $r_{12.3}$.

Solution: The regression equation of X_1 for X_2 and X_3 is given by :

$$X_1 - \bar{X}_1 = b_{12.3}(X_2 - \bar{X}_2) + b_{13.2}(X_3 - \bar{X}_3) \ldots.(i)$$

where,

$$b_{12.3} = \frac{S_1}{S_2}\left[\frac{r_{12} - r_{13}.r_{23}}{1 - r^2_{23}}\right]$$

$$= \frac{1}{0.80}\left[\frac{0.60 - 0.70 \times 0.65}{1 - (0.65)^2}\right]$$

or

$$b_{12.3} = (1.25)\left[\frac{0.60 - 0.455}{0.578}\right] = 0.313$$

$$b_{13.2} = \frac{S_1}{S_2}.\left[\frac{r_{13} - r_{12}.r_{32}}{1 - r^2_{32}}\right]$$

$$= \frac{1}{0.90}\left[\frac{0.70 - 0.60 \times 0.65}{1 - (0.65)^2}\right]$$

$$= (1.111)\left[\frac{0.70 - 0.39}{0.578}\right] = 0.595$$

Substituting the values in equation (I), you have

$$X_1 - 6.8 = 0.313(X_2 - 7.0) + 0.595(X_3 - 7.4)$$
$$X_1 = 0.206 + 0.313X_2 + 0.595X_3$$

(II) Substituting for $X_2 = 10$ and $X_3 = 9$ in the above regression and solving for x_1.

$$X_1 = 0.206 + 0.313(10) + 0.595(9) = 8.691$$

(III) Multiple and partial correlation coefficients are related as:

$$R^2_{12.3} = 1 - (1 - r^2_{12})(1 - r^2_{13.2})$$

$$r_{13.2} = \frac{r_{13}.r_{12}.r_{32}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{32}^2}}$$

$$= \frac{0.70 - 0.60 \times 0.65}{\sqrt{1-(0.60)^2}\sqrt{1-(0.65)^2}}$$

$$= \frac{0.70 - 0.39}{0.8 \times 0.760} = 0.509$$

or,

$$r_{13.2}^2 = 0.259$$

Substituting the values of r_{12}^2 and $r_{13.2}^2$ for $R_{1.23}^2$ you have

$$R_{1.23}^2 = 1 - (1 - 0.36)(1 - 0.259) = 0.526$$

Example 12.13: In a trivariate distribution:

$$\overline{X}_1 = 28.02, \overline{X}_2 = 4.91, \overline{X}_3 = 594, S_1 = 4.4, S_2 = 1.1, S_3 = 80$$

$$r_{12} = 0.80, r_{23} = -0.56, r_{31} = -0.40$$

(I) Find the correlation coefficient $r_{23.1}$ and $R_{1.23}$.

(II) Estimate the value of X_1 when $X_2 = 6.0$ and $X_3 = 650$.

Solution:

(I)

$$r_{23.1} = \frac{r_{23} - r_{21}.r_{31}}{\sqrt{1-r_{21}^2}\sqrt{1-r_{31}^2}}$$

Substituting the values, you get

$$r_{23.1} = \frac{-0.56 - (0.80).(-0.40)}{\sqrt{1-(0.80)^2}\sqrt{1-(-0.40)^2}}$$

$$= \frac{-0.56 + .32}{\sqrt{1-.64}\sqrt{1-.16}}$$

$$= \frac{-0.24}{0.6 \times 0.916}$$

$$= -0.436$$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}.r_{13}.r_{23}}{1-r_{23}^2}}$$

Substituting the values, you get

$$R_{1.23} = \sqrt{\frac{(0.80)^2 + (-0.40)^2 + 2(0.80)(-0.40)(-0.56)}{1-(-0.56)^2}}$$

$$= \sqrt{\frac{0.64 + .16 - .3584}{1 - .3136}}$$

$$= \sqrt{\frac{0.4416}{0.6864}} = 0.802$$

(II) The liner regression equation of X_1 for X_2 and X_3 is given by:

$$X_1 - \bar{X}_1 = b_{12.3}(X_2 - \bar{X}_2) + b_{13.2}(X_3 - \bar{X}_3)$$

$$b_{12.3} = \frac{S_1}{S_2}\left[\frac{r_{12} - r_{13}.r_{23}}{1 - r_{23}^2}\right]$$

$$= \frac{4.4}{1.1}\left[\frac{0.80 - 0.224}{0.6864}\right]$$

$$= \frac{4.4}{1.1}\left[\frac{0.576}{0.6864}\right] = 3.357$$

$$b_{13.2} = \frac{S_1}{S_2}.\left[\frac{r_{13} - r_{12}.r_{32}}{1 - r_{32}^2}\right]$$

$$= \frac{4.4}{80}.\left[\frac{-0.40 - (0.80)(-0.56)}{1 - (-.56)^2}\right]$$

$$= \frac{4.4}{80}\left[\frac{-0.40 + 0.448}{0.6884}\right]$$

$$= \frac{4.4}{80}\left[\frac{0.048}{0.6864}\right]$$

$$= 0.0038 \; or \; 0.004$$

(III) Substituting the values in equation, you get

$$X_1 - 28.02 = 3.357(X_2 - 4.91) + .004(X_3 - 594)$$

or

$$X_1 - 28.02 = 3.357X_2 - 16.4828 + .004X_3 - 2.376$$

$$X_1 = 3.357X_2 + .004X_3 + 9.1612$$

(IV) Estimation of X_1 for X_2 and X_3:

When $X_2 = 6.00, X_2 = 650, \quad X_1 = 3.357(6.00) + .004(650) + 9.1612$

$$= 20.142 + 2.6 + 9.1612$$

$$= 31.9032$$

12.5 SUMMARY

In total, multiple regression is an extension of the technique of simple regression under which the interrelationship between three or more variables is studied to estimate the most probable value of the dependent variable for given values of the independent variables. The multiple regression equations can be worked out by ways of multiple regression equations using the normal equation, and the multiple regression equation in terms of the simple correlation coefficient. Further, the standard error of estimate measures the reliability of the estimates given by the multiple regression equation. It shows to what extent the estimated values given by the regression equations are closer to the actual values. As far as the coefficient of determination in multiple regression ($R_{1.23}^2$) is concerned, it represents the

proportion (fraction) of the total variation in the dependent variable X_1 that has been explained by the independent variables (X_1 *and* X_2) in the multiple regression equation.

12.6 GLOSSARY

Multiple regression is an extension of the technique of simple regression under which the interrelationship between three or more variables is studied.

12.7 CHECK YOUR PROGRESS

1. From the following data, find the least square regression of X_1, X_2 and X_3 and estimate The value of X_1 for given values of $X_2 = 16$ and $X_3 = 4$:

X_1:	10	5	10	4	8
X_2:	16	13	21	10	13
X_3:	3	6	4	5	3

2. Compute the values of b_0, b_1 and b_2 for the equation $Y = b_0 + b_1X_1 + b_2X_2$ from the following data:

Y:	3	5	6	8	12	14
X_1:	16	10	7	4	3	2
X_2:	90	72	54	42	30	12

3. Obtain the parameters of the multiple linear regression model: $Y = \beta_1 + \beta_2X_2 + \beta_3X_3$ from the following data:

$$N = 6, \Sigma Y = 54, \Sigma X_2 = 48, \Sigma X_3 = 102$$
$$\Sigma YX_2 = 339, \Sigma YX_3 = 720, \Sigma X_2X_3 = 1034, \Sigma X_2^2 = 494, \Sigma X_3^2 = 2188$$

4. For the following set of data, find the multiple regression of X_1 on X_2 and X_3 using actual mean method. Also, predict the value of X_1 when $X_2 = 5$ and $X_3 = 7$:

X_1:	12	24	32	28
X_2:	6	12	16	22
X_3:	4	6	12	18

5. From the data given below, find the multiple regression equation of X_1 on X_2 and X_3 using the actual mean method.

X_1:	4	6	7	9	13	15
X_2:	15	12	8	6	4	3
X_3:	30	24	20	14	10	4

6. From the data given below, find the multiple linear regression of X_1 on X_2 and X_3 using actual mean method.

X_1:	18	20	17	14	21
X_2:	38	40	25	28	44
X_3:	20	15	5	12	18

7. Given the following information (variables are measured from their respective means):

$$\Sigma x_1x_2 = 1900, \quad \Sigma x_1x_3 = -20, \quad \Sigma x_2x_3 = -50$$
$$\Sigma x_1^2 = 1350, \quad \Sigma x_2^2 = 2800, \quad \Sigma x_3^2 = 24$$
$$\bar{X}_1 = 65, \quad \bar{X}_2 = 55, \quad \bar{X}_3 = 30$$

Obtain the partial regression coefficients ($b_{12.3}$ *and* $b_{13.2}$)
Also estimate the value of X_1 when $X_2 = 60$ and $X_3 = 25$.

8. Given the following information (variables are measured from their respective means):

$$\Sigma x_1^2 = 1350, \quad \Sigma x_2^2 = 2800, \quad \Sigma x_3^2 = 24$$
$$\Sigma x_1 x_2 = 1900, \quad \Sigma x_1 x_3 = -20, \quad \Sigma x_2 x_3 = -50$$

Determine the regression equation of X_1 on X_2 and X_3.

12.8 ANSWERS TO CHECK YOUR PROGRESS

1. [$X_1 = 4.753 + 0.502\, X_2 - 1.115\, X_3, 8.325$]
2. [$Y = 16.1067 + .426\, X_1 - 0.221\, X_2$]
3. [$Y = 16.479 + 0.389\, X_2 - 0.623\, X_3$]
4. [$X_1 = 2.577 + 1.661\, X_2 + 0.0169\, X_3, X_1 = 110$]
5. [$X_1 = 16.479\, X_2 + 0.623\, X_3$]
6. [$X_1 = 0.5\, X_2 - 0.36\, X_3 + 5.54$]
7. [$b_{12.3} = 0.689, b_{13.2} = 0.603, X_1 = 65.43$]
8. [$X_1 = 0.689\, X_2 + .603\, X_3$]

12.9 TERMINAL QUESTIONS

1. The following constants are obtained from measurements of length in mm(x_1), volume in c.c.(x_2) and weight in gm(x_3) of 300 eggs :

$$\bar{X}_1 = 55.95 \quad S_1 = 2.26 \quad r_{12} = 0.578$$
$$\bar{X}_2 = 51.48 \quad S_2 = 4.39 \quad r_{13} = 0.581$$
$$\bar{X}_3 = 56.03 \quad S_3 = 4.41 \quad r_{23} = 0.974$$

Obtain the linear regression equation of egg weight on egg length and egg volume. Hence, estimate the weight of an egg whose length is 58 mm and volume is 52.5 c.c.

2. In a trivariate distribution:

$$\sigma_1 = 2.7, \quad \sigma_2 = 2.4, \quad \sigma_3 = 2.7$$
$$r_{12} = 0.28, \quad r_{23} = 0.49, \quad r_{31} = 0.51$$

Determine the regression equation of x_3 on x_1 and x_2 if the variates are measured from their means.

3. Given the following data:

$$\bar{X}_1 = 6, \quad \bar{X}_2 = 7, \quad \bar{X}_3 = 8$$
$$\sigma_1 = 1, \quad \sigma_2 = 2, \quad \sigma_3 = 3$$
$$r_{12} = 0.6, \quad r_{13} = 0.7, \quad r_{23} = 0.8$$

Obtain the linear regression equation of X_3 on X_1 and X_3. Hence estimate X_3 when $X_1 = 4$ and $X_2 = 5$.

4. If $r_{12} = 0.926, r_{13} = 0.891, r_{23} = 0.955$ and $S_1 = 1.51$Compute the standard error of an estimate of X_1 on X_2 and X_3($S_{1.23}$).

5. Given the following data:

$$\bar{X}_1 = 55 \quad S_1 = 5 \quad r_{12} = 0.57$$
$$\bar{X}_2 = 51 \quad S_2 = 7 \quad r_{13} = 0.58$$
$$\bar{X}_3 = 56 \quad S_3 = 9 \quad r_{23} = 0.97$$

Calculate:
(I) Multiple Regression of X_3 on X_1 and X_2.
(II) Multiple Correlation Coefficient $R_{1.23}$

6. From the following data:

$$r_{12} = 0.8, r_{13} = 0.5, r_{23} = 0.3$$
$$S_1 = 8.5, S_2 = 4.5 \; and \; S_3 = 2.1$$

(I) Obtain the regression equation of X_1 on X_2 and $X_3$with $\bar{X}_1 = 70kg, \bar{X}_2 =$ 22 years and $\bar{X}_3 = 150cm$.

(II) Estimate the value of X_1 on $X_2 = 25$ years and $X_3 = 140cm$, and

(III) Find the coefficient of multiple determination $R^2_{1.23}$ from r_{12} and $r_{13.2}$. What does R^2 indicate?

ANSWERS

1. [$X_3 = 3.54 + 0.052X_1 + 0.963X_2, X_3 = 57.11 \; gms.$]

2. [$x_3 = 0.405x_1 + 0.424x_2$]

3. [$x_3 = -4.41 + 1.03x_1 + 0.89x_2, x_3 = 4.16$]

4. [$S_{1.23} = 0.5702$]

5. [$X_3 = 0.08X_1 + 1.21X_2 - 10.11, R_{1.23} = 0.97$]

6. [$X_1 = 1.349X_2 + 1.156X_3 - 133.078, X_1 = 62.487 R^2_{1.23} = 0.7443, R^2$ indicates the 74.43% variation in X_1 are explained by the multiple regression equation.

12.10 SUGGESTED READINGS

1. Basic Statistics – Goon, Guptha and Dasgupta – World Press Limited Calcutta.

2. Fundamentals of Business Statistics – Sanchethi and Kappor.

3. Quantitative Methods in Management – Srivastava, Shenoy and Guptha.

4. Business Statistics – Guptha and Guptha

PAPER CODE: MCM-02
BLOCK: 4

UNIT: 13 TYPES AND TECHNIQUES OF STATISTICAL QUALITY CONTROL

Structure

OBJECTIVES

After studying this unit, you will be able to understand:

- The meaning of statistical quality control;
- Advantages of quality control;
- Limitations of quality control
- Causes of variation in quality characteristics;
- Methods of statistical quality control;
- Techniques of statistical quality control; and
- Control charts for variables;

13.1 INTRODUCTION

In this era of competition, it is a dire need for manufacturer to produce qualitative products to enter and stay in the market, and they have to keep a continuous watch over the quality of the goods produced by them. But, due to large-scale production level, it is not possible for a producer to check the quality of each and every item produced. Therefore, to control the quality of the manufactured goods, the study of statistical quality control (SQC) becomes very relevant and useful.

13.2 MEANING OF STATISTICAL QUALITY CONTROL

Statistical Quality Control refers to the use of statistical techniques in controlling the quality of manufactured goods. It is the way of establishing and achieving the quality specifications that requires use of tools and techniques of statistics. It is an important application of the theory of probability and the theory of sampling for the maintenance of uniform quality in a continuous flow of manufactured products. One of the major tools of SQC is the control chart, which was firstly introduced by W.A. Shewhart through the application of normal distribution.

Some important definitions of statistical quality control are given below:

(i) "Statistical quality control can be simply defined as an economic and effective system of maintaining and improving the quality of outputs throughout the whole operating process of specification, production and inspection based on continuous testing with random samples."

-Ya Lun Chou

(ii) **"**Statistical quality control should be viewed as a kit of tools which may influence decisions to the functions of specification, production or inspection". **-Eugene L. Grant**

From the above definitions, the essential characteristics of SQC may be brought about like: (i) it is designed to control quality standard of goods produced for marketing, (ii) it is exercised by the producers during the production process to assess the quality of the goods, (iii) it is carried out with the help of certain statistical tools like Mean chart, Range chart, P-chart, C-chart, Sampling Inspection Plans, etc. and (iv) it is designed to determine the variations in quality of the goods and limits of tolerance.

13.3 ADVANTAGES OF STATISTICAL QUALITY CONTROL

The following are the advantages of statistical quality control:

(i) It provides an objective method of controlling the quality of the product during the production process. It tells the production manager at a glance whether the quality of the product is under control or not.

(ii) It provides a quick method to eliminate assignable causes of variation. By using the technique of statistical quality control, we can detect assignable causes of variations may be detected and take necessary remedial measures may be taken to avoid them.

(iii) It provides better quality assurance at a lower inspection cost. Sampling inspection is always cheaper vis-a-vis 100 per cent inspection. Control charts are simple to construct and easy to interpret and economical.

(iv) The acceptance sampling protects the interest of the consumers by helping them to reject a lot of bad quality. This is also helpful to the producers because they become aware of the probability of a good lot being rejected.

(v) The very presence of statistical quality control in a manufacturing plant has a healthy influence on the psychology of workers and makes them quality conscious. They came to know that the quality is being checked.

(vi) A quality-conscious manufacturing unit is able to earn the goodwill from the consumers of its product, which is of immense long-run value.

(vii) Past data on quality control may serve as a guide for the choice of a new plant and machinery, as well as technical staff.

(viii) It is possible to defend the quality of output before any government agency based on quality control records.

13.4 LIMITATIONS OF STATISTICAL QUALITY CONTROL

Despite the great significance of significance of statistical quality control, the technique of SQC suffers from certain limitations, as under:

(i) It cannot be applied indiscriminately as a panacea for all quality evils.

(ii) It cannot be used mechanically in all production process without studying their peculiar environment.

(iii) It involves mathematical and statistical problems in the process of analysis and interpretation of variations in quality.

(iv) It provides only an information service.

13.5 CAUSES OF VARIATION IN QUALITY CHARACTERISTICS

Every manufacturer produces the product according to pre-determined standards. Though the product is carried out with the most sophisticated technology, some variations in the quality of products are bound to take place, e.g., it is not possible that all pins, nuts, or bolts produced in a factory would be exactly of the same quality. There must be some variation, however minor it might be, in the quality of the various items produced. There may be various causes of this variation. These causes are classified into the following two groups:

(i) **Assignable Causes:** These causes, as the name suggests, refer to those changes in the quality of the products which can be assigned or attributed to any particular cause, like defective materials, defective labour, defective machine, etc. However, the effect of such variations can be eliminated with a better system of control, like SQC.

(ii) **Chance Causes:** These causes, as the name suggests, take place as per chance or in a random fashion as a result of the cumulative effect of a multiplicity of several minor production and hence it is accepted as an allowable variation in any scheme of production.

Out of these two types of causes, nothing can be done about the chance causes. However, assignable variations can be detected and corrected.

13.6 METHODS OF STATISTICAL QUALITY CONTROL

There are two types of methods of statistical quality control, which are applied in two distinct phases of plant operation. They are as under:

13.6.1 Process control: Under the process control, the quality of the products is controlled while the products are in the process of production. The process control is secured with the technique of control charts. Control charts are used as a measure of quality control not only in the production process but also in the areas of advertising, packing, airline reservations, etc. Control charts ensure that whether the products confirm to the specified quality standard or not.

13.6.2 Product Control: Under the product control, the quality of the products is controlled while the product is ready for sale and dispatch to the customers. The product control is secured with the technique of acceptance sampling. In acceptance sampling, the manufactured articles are formed into lots, a few items are chosen randomly, and lot is either accepted based on a certain set of rules, usually called sampling inspection plans.

Thus, process control is concerned with controlling of quality of the goods during the process of manufacturing, whereas product control is concerned with the inspection of finished goods, when they are ready for delivery.

13.7 CONTROL CHARTS

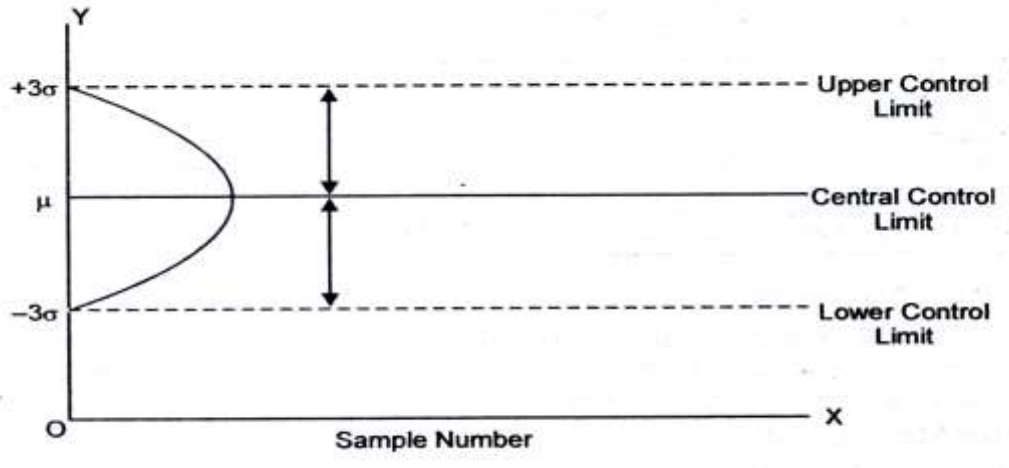
The control charts are the graphic devices developed by Walter A. Shewhart for detecting unnatural pattern of variation in the production process and determining the permissible limits of probability and sampling. Control charts are the core of statistical quality control. These are based on the theory of probability and sampling. Control charts are simple to construct and easy to interpret and they tell the production manager at a glance whether or not the process is in control i.e., within the tolerance limits. A control chart consists of three horizontal lines: (i) Central Line (CL), (ii) Upper Control Limit (UCL), and (iii) Lower Control Limit (LCL).

13.7.1 Central Line (CL): The central line is the middle line of the chart. It indicates the grand average of the measurements of the samples. It shows the desired standard or level of the process. The central line is generally drawn as bold line.

13.7.2 Upper Control Limit (UCL): The upper control limit is usually obtained by adding 3 sigma (3σ) to the grand average. It is denoted by Mean-3σ. The upper control limit is generally drawn as dotted line.

13.7.3 Lower Control Limit (LCL): The lower control limit is usually obtained by subtracting 3 sigma (3σ) to the grand average. It is denoted by Mean-3σ. The upper control limit is generally drawn as a dotted line.

On the basis of these three lines, a control chart is constructed. The general format of a control chart is given in the diagram below:



In the control chart, the mean values of the statistics T (i.e., Mean, Range, S.D., etc.) for successive samples are plotted and often joined by broken lines to provide a visual clarity. So long as the sample points fall within the upper and lower control limits, there is nothing to worry and in such a case the variation between the samples is attributed to chance causes.

13.8 LOGIC OF SETTING OF CONTROL LIMITS AT ± 3 σ

Dr. Shewhart has proposed the 3σ limits for the control charts. From the probability, if a variable X is normally distributed, the probability that a random variable will be between $\overline{X} \pm 3\sigma$ where, $\overline{X}$ is the mean and σ is the standard deviation is 0.9973, which is extremely high. Thus, the probability of a random variable fall outside these limits is 0.0027, which is very low. In other words, the occurrence of events beyond the limits of $\overline{X} \pm 3\sigma$, provided the events lie on a normal curve, is on the whole nearly 3 out of 1000 events that are extremely remote chances under normal circumstances. Thus, if $\pm 3\sigma$ limits are employed and the variable quality characteristic is assumed to be normally distributed; then, the probability of sample points falling outside these limits when the process is in control is very small.

13.9 PURPOSE AND USES OF CONTROL CHARTS

The control charts are useful in the following situations:

(i) It helps in determining the quality standard of the products when they are in process.

(ii) It helps in detecting the chance and assignable variations in the quality standards of the products by setting two control limit lines.

(iii) It reveals variations in the quality standards of the products from the desired level.

(iv) It indicates whether the production process is in control or not to take necessary steps for its correction.

(v) Control charts are simple to construct and easy to interpret.

(vi) It ensures less inspection cost and time in the process control.

(vii) Control charts tell the production manager at a glance whether or not the process is in control.

13.10 TYPES OF CONTROL CHARTS

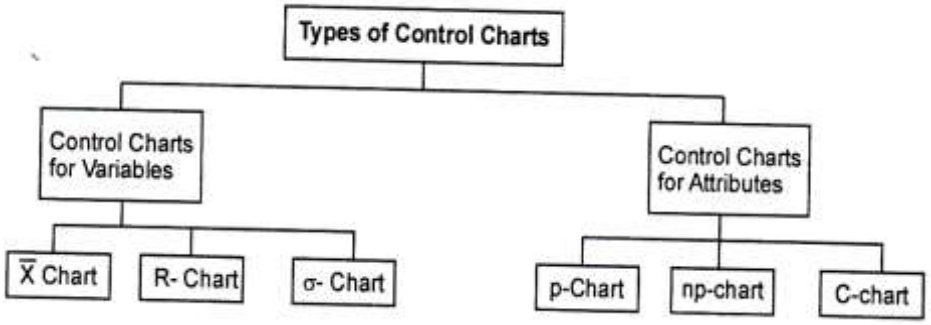
Control charts are of two types, depending on whether a given quality or characteristic of a product is measurable or not. These are:

(A) Control Charts for Variables

(i) $\overline{X}$-Chart (ii) R-Chart (iii) $\sigma - Chart$

(B) Control Charts for Attributes (you will study in the next Unit)

(i) $p - chart$ (ii) $np - chart$ (iii) $C - chart$



13.10.1 CONTROL CHARTS FOR VARIABLES

These charts are used when the quality or characteristics of a product is capable of being measured in quantitative terms such as gauge of a steel box, diameter of a screw, tensile strength of a steel pipe, resistance of a wire, etc. Such charts are of three types:

(i) $\bar{X} - Chart$: This chart is constructed for controlling the variations in the average quality standard of the products in a production process.

Procedure: The construction of $\bar{X} - chart$ involves the following steps:

(a) Compute the mean of each sample i.e.,

$$\bar{X}_1, \bar{X}_2, \bar{X}_3 \ldots \bar{X}_k$$

(b) Compute the mean of the samples mean samples mean by dividing the sum of the sample means by the number of samples i.e.,

$$\bar{\bar{X}} = \frac{\bar{X}_1, \bar{X}_2, \bar{X}_3 \ldots \bar{X}_k}{No.of\ samples} = \frac{\Sigma \bar{X}}{k}$$

Where, k = No. of samples

This grand mean $\bar{\bar{X}}$ represents the Central Line (CL)

(i) Determine the control limits by using the following formula:

(a) On the Basis of the standard deviation of the population (σ)

$$Control\ Limits = \bar{\bar{X}} \pm \frac{3\sigma}{\sqrt{n}}$$

Therefore,

$$UCL = \bar{\bar{X}} + \frac{3\sigma}{\sqrt{n}}$$

and

$$LCL = \bar{\bar{X}} - \frac{3\sigma}{\sqrt{n}}$$

These control limits represent the upper control line and lower control line.

(b) On the basis of the Quality Control Factors A_2 and $\bar{R}$

Control Limits = $\bar{\bar{X}} \pm A_2\bar{R}$, where, $\bar{R}$ = Mean of the ranges

$$UCL = \bar{\bar{X}} + A_2\bar{R}$$

$$LCL = \bar{\bar{X}} - A_2\bar{R}$$

Where, A_2 is a quality control factor whose value is obtained from the control chart table with reference to the size of the sample, and

$\bar{R}$=Mean of Ranges = $\frac{\Sigma R}{N}$

(iv) Construct the mean chart ($\bar{X} - chart$) by plotting the sample number on x-axis and sample mean, UCL, LCL and Central Line on the y-axis.

(v) Interpret the $\bar{X} - Chart$. If all the sample means ($\bar{X}$) fall within the control limits, the production process is in a state of control; otherwise, it is beyond control.

(ii) R-Chart

The Range (R-chart) is constructed for controlling the variation in the dispersion or variability of the quality standard of the products in a production process.

Procedure:

The construction of the R-chart involves the following steps:

(i) Compute the range (R) of each sample using the formula:

R = L - S L = Largest value S = Smallest value

(ii) Compute the mean of ranges by dividing the sum of the samples ranges (ΣR) by the number of samples, i.e.,

$$\bar{R} = \frac{R_1 + R_2 + \cdots R_k}{k} = \frac{\Sigma R}{k}$$

Where, k = No. of samples

The mean of ranges ($\bar{R}$) represents the Central line (CL) for the R-Chart

(iii) Determine the control limits by using the following formula:

(a) On the basis of Quality Control Factors D_3 and D_4 and $\bar{R}$:

Upper control limit (UCL) = $D_4\bar{R}$

Lower control limit (LCL) = $D_3\bar{R}$

Where D_3 and D_4 are the quality control factors and their values are obtained from the control chart table with reference to the size of the sample.

$\bar{R} = Mean\ of\ Range$

(b) On the Basis of Quality Control Factors D_1, D_2 and population standard deviation (σ)

UCL = $D_2\sigma$

LCL = $D_1\sigma$

Where D_1 and D_2 are the quality control factors.

The value of LCL cannot be negative and in such a case it would be reduced to zero.

(iv) Construct the R-Chart (Range Chart) by plotting the sample number on the x-axis and sample ranges (R), UCL, LCL, and Central Line (CL) on the y-axis.

(v) Interpret the R-chart. If all the sample ranges (R) fall within the control limits, the production process is in a state of control otherwise, it is beyond control.

Example 13.1: Construct $\bar{X} - Chart$ and Range Chart for the following data of 5 samples with each set of 5 items:

Sample No.	Weights				
1	20	15	10	11	14
2	12	18	10	8	22
3	21	19	17	10	13
4	15	12	19	14	20
5	20	19	26	12	23

(Conversion factors for n = 5, A_2 = 0.577, D_3 = 0, D_4 = 2.115)

Solution: Construction of $\bar{X}$ and R Charts

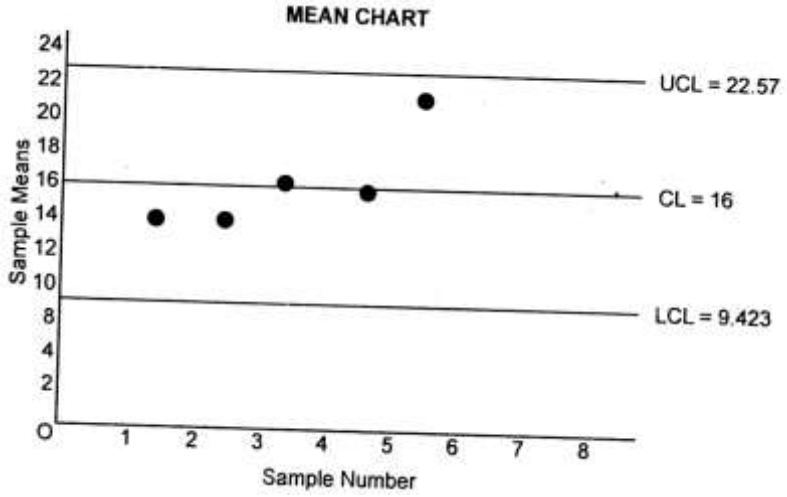
Sample No.	Weights of Item in each sample (X)					Total Weights (ΣX)	$\bar{X} = (\Sigma X \div 5)$	Range $R = (L - S)$
1	20	15	10	11	14	70	14	10

2	12	18	10	8	22	70	14	14
3	21	19	17	10	13	80	16	11
4	15	12	19	14	20	80	16	8
5	20	19	26	12	23	100	20	14
$K = 5$							$\Sigma\bar{X} = 80$	$\Sigma R = 57$

$$\bar{\bar{X}} = \frac{\Sigma\bar{X}}{k} = \frac{80}{5} = 16 \qquad \bar{R} = \frac{\Sigma R}{k} = \frac{57}{5} = 11.4$$

$\bar{X}$ Chart

$\bar{\bar{X}}$ = 16 (Central Line)



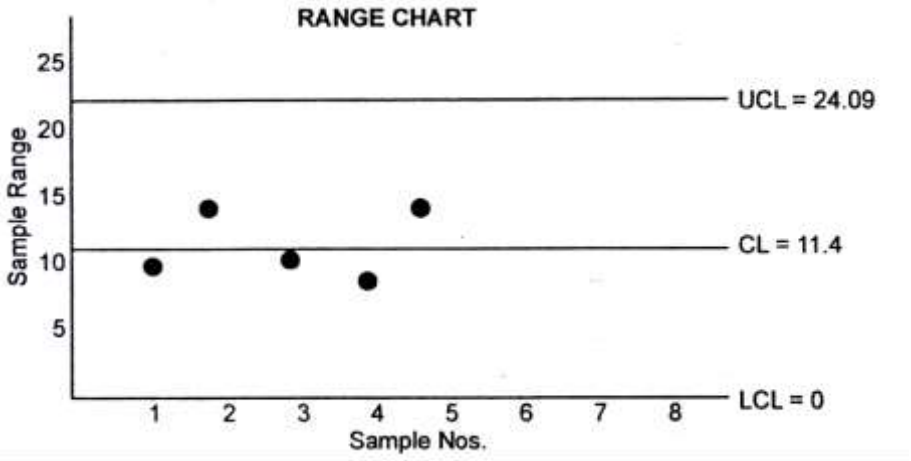
Control Limits

UCL $= \bar{\bar{X}} + A_2\bar{R}$
$= 16 + 0.577 \times 11.4$
$= 16 + 6.577$
$= 22.577$

LCL $= \bar{\bar{X}} - A_2\bar{R}$
$= 16 - 0.577 \times 11.4$
$= 16 - 6.577$
$= 9.423$

As the entire sample mean values fall within the control limits, the chart shows that the given process is in statistical control.

Range Chart:



$\overline{R}$ = 11.4 (Central Line)

Control Limits:

UCL = $D_2\overline{R} = 2.115 \times 11.4 = 24.09$ LCL = $D_3\overline{R} = 0 \times 11.4 = 0$

As all the range points fall within the control limits, so $R - chart$ shows that the given process is in statistical control.

Example 13.2: A machine is set to deliver packet of a given weight. 10 sample of size 5 each were recorded in the data given below:

Sample No.:	1	2	3	4	5	6	7	8	9	10	Total
Mean $\overline{X}$:	15	17	15	18	17	14	18	15	17	16	162
Range:	7	7	4	9	8	7	12	4	11	5	74

Construct the Mean Chart and Range chart and comment on state of control. (Conversion Factors for n = 5 are A_2 = .577, D_3 = 0, D_4 = 2.115).

Solution:

Sample No.:	1	2	3	4	5	6	7	8	9	10	Total
Mean $\overline{X}$:	15	17	15	18	17	14	18	15	17	16	162
Range:	7	7	4	9	8	7	12	4	11	5	74

$$\overline{\overline{X}} = \frac{\Sigma\overline{X}}{N} = \frac{162}{10} = 16.2 \qquad \overline{R} = \frac{\Sigma R}{N} = \frac{74}{10} = 7.4$$

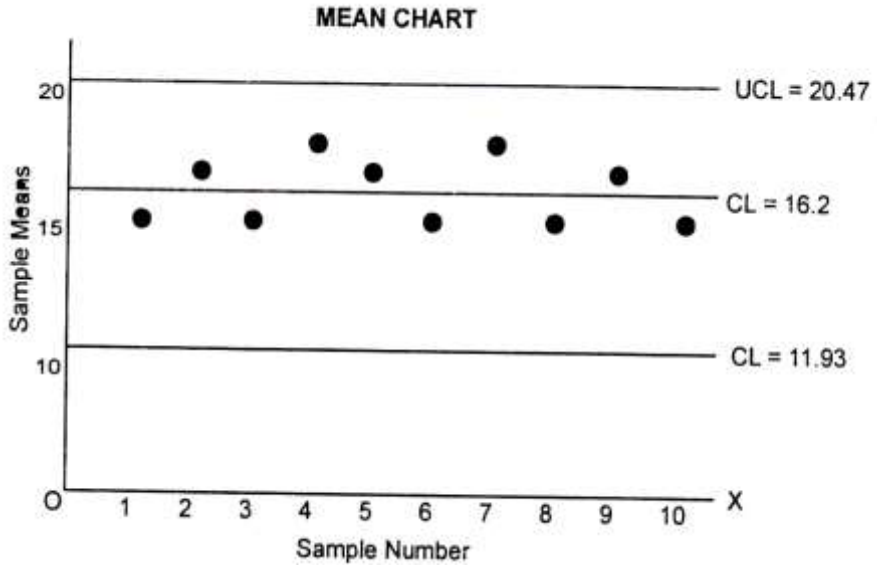
Mean Chart $(\overline{X}\ Chart)$

$\overline{\overline{X}}$ = 16.2 (Central Line)

Control Limits

UCL $= \overline{X} + A_2\overline{R}$
$= 16.2 + 0.577 \times 7.4$
$= 20.47$

LCL $= \overline{X} - A_2\overline{R}$
$= 16.2 - 0.577 \times 7.4$
$= 11.93$

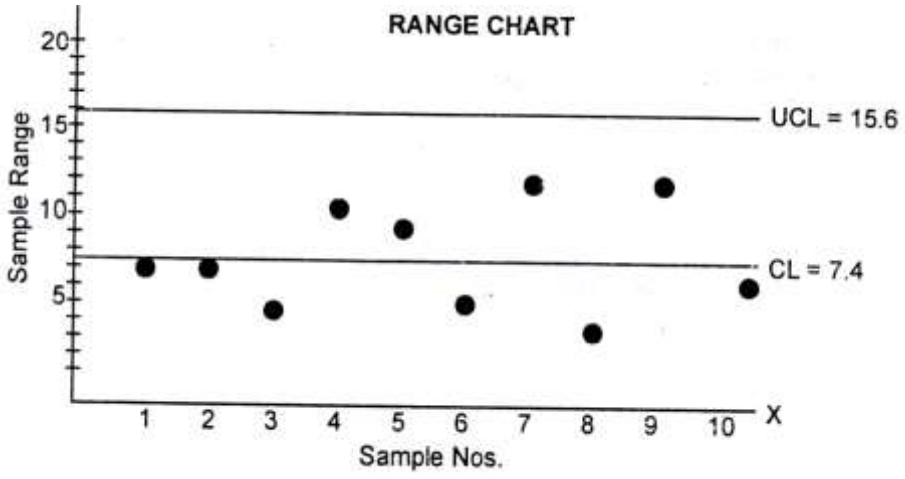


As the entire sample mean values fall within the control limits, the chart shows that the given process is in statistical control.

Range Chart ($R - Chart$)
$\overline{R}$ = 7.4 (Central Line)
Control Limits

$$\text{UCL} = D_4.\overline{R}$$
$$= 2.115 \times 7.4$$
$$= 15.65$$

$$\text{LCL} = D_3\overline{R}$$
$$= 0 \times 7.45 = 0$$



As all the range points fall within the control limits, so $R - chart$ shows that the given process is in statistical control.

Example 13.3: The following are the lengths and ranges of lengths of a finished product from 10 samples, each of size 5. The specification limits for length are 200±5 cm. Construct $\overline{X}$ and R charts, and examine whether the process is under control, and state your recommendations.

Sample No.:	1	2	3	4	5	6	7	8	9	10
Mean $\bar{X}$:	201	198	202	200	203	204	199	196	199	201
Range:	5	0	7	3	4	7	2	8	5	6

Assume for n = 5, A_2 = 0.577, D_3 = 0, D_4 = 2.115.

Solution: The specification limits for length are given to be 200±5 cm. Hence, mean is known where as standard deviation is unknown.

Control limits for $\bar{x}$ chart

Central limit,

$$CL = \bar{\bar{X}} = 200.$$

$$UCL = \bar{\bar{X}} + A_2\bar{R}, where\ \bar{R} = \frac{\Sigma R_i}{10} = \frac{47}{10} = 4.7.$$

$$UCL = 200 + 0.577 \times 4.7 = 202.712$$

$$LCL = 200 - 0.577 \times 4.7 = 197.29.$$

Control limits for R chart

$$CL = \bar{R} = 4.7.$$

$$UCL = D_4\bar{R} = 2.115 \times 4.7 = 9.941$$

$$LCL = D_3\bar{R} = 0 \times 4.7 = 0.$$

The $\bar{X}$ and R charts are drawn in Fig. (a) and (b) respectively.

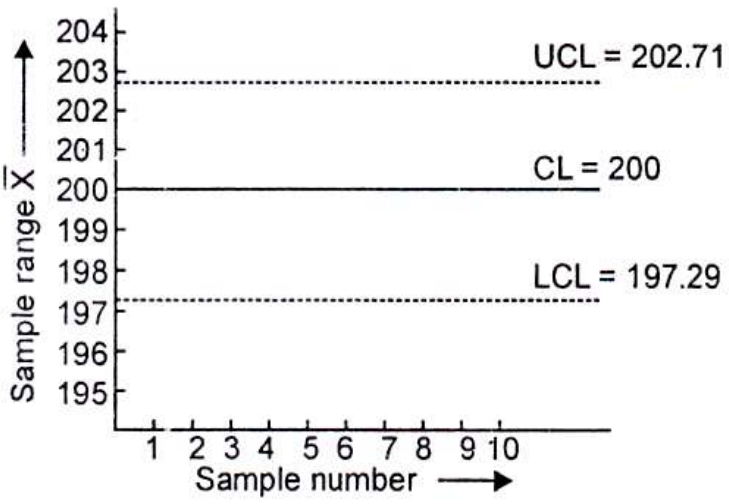


Fig. (a)

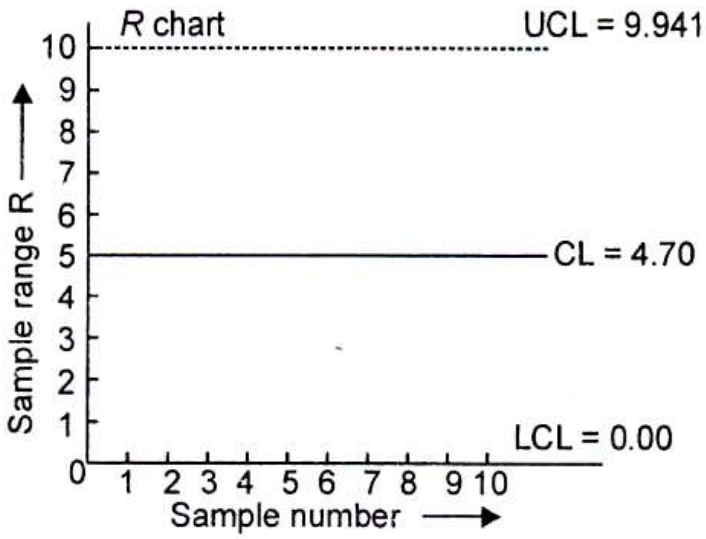


Fig. (b)

It can be seen that all points lie within the control limits of the R chart. The process variability is, therefore, under control. However, three points corresponding to sample no. 5, 6, and 8 lay outside the control limits of $\bar{X}$ chart. The process is, therefore, not in statistical control. The process needs to be checked to see whether there is any assignable cause. If they are found, then the process should be readjusted to remove them, otherwise, fluctuations are going on to be there.

Example 13.4: Twenty-five samples of six items each were related from the assembly line of a machine having mean of 25 samples is 0.81 inches and range .0025 inches. Compute the upper control limits and lower control limits of mean chart and the range chart.
(For n = 6, A_2 = 0.483, D_3 = 0, D_4 = 2.004).

Solution: Given: $\bar{\bar{X}} = 0.81, \bar{R} = 0.0025, n = 6$

$\bar{X}$ -Chart

Control Limits

$$UCL = \bar{\bar{X}} + A_2\bar{R}$$
$$= .81 + (.483)(.0025)$$
$$= .8112$$
$$LCL = \bar{\bar{X}} - A_2\bar{R}$$
$$= .81 - (.483)(.0025)$$
$$= .8088$$

Range Chart

Control Limits

$$UCL = D_4\bar{R}$$
$$= (2.004) \times (.0025)$$
$$= 0.0050.$$
$$LCL = D_3\bar{R}$$
$$= 0 \times (.0025) = 0.$$

Example 13.5: The mean life of a battery cell manufactured by a certain plant as estimated on the basis of a large sample was found to be 1500 hrs with standard deviation of 180 hrs. Compute the 3-sigma (3σ) control limits for $\bar{X} - Chart$ for a sample of size n = 9.

Solution: Given: $\bar{\bar{X}} = 1500\ hrs., \sigma = 180 hrs., n = 9$

Control Limits for $\bar{X}\ Chart.$

$$UCL = \bar{\bar{X}} + 3\frac{\sigma}{\sqrt{n}}$$

$$= 1500 + 3 \times \frac{180}{\sqrt{9}}$$

$$= 1680$$

$$LCL = \bar{\bar{X}} - 3\frac{\sigma}{\sqrt{n}}$$

$$= 1500 - 3 \times \frac{180}{\sqrt{9}}$$

$$= 1320$$

(iii) σ-Chart

This chart is constructed to get a better picture of the variations in the quality standard in a process than that is obtained from the Range chart provided the standard deviation of the various samples are readily available.

Procedure: The construction of σ-chart involves the following steps:

(i) Find the S.D. of each sample, if not given.

(ii) Compute the mean of the standard deviation by using the formula:

$$\bar{S} = \frac{\Sigma S}{k} = \frac{S_1 + S_2 + S_3 + \cdots S_k}{k}$$

The mean of S.D.s ($\bar{S}$) represents the central line (CL).

(iii) Find the upper and lower control limits by using the formula:

a) On the basis of quality control factors B_1, B_2 and population standard deviation (σ)

$UCL = B_2.\sigma$ $\qquad$ $LCL = B_1.\sigma$

b) On the basis of quality control factors B_3, B_4 and estimated population standard deviation ($\bar{S}$)

$UCL = B_4.\bar{S}$ $\qquad$ $LCL = B_3.\bar{S}$

Where, B_1, B_2, B_3 and B_4 are the quality control factors.

(iv) Construct σ-Chart by plotting the sample number on the x-axis and sample S.D. (σ), UCL, LCL and CL on the y-axis.

(v) Interpret the chart thus drawn.

Example 13.6: Quality control is maintained in a factory with the help of mean and standard deviation charts. Ten items are chosen in every sample. Eighteen samples in all were chosen whose $\Sigma\bar{X}$ was 595.8 and ΣS was 8.28. Determine the three sigma limits of $\bar{X}$ and σ charts. You may use the following factors for finding 3σ limits

n	A_1	B_3	B_4
10	0.949	0.28	1.72

Solution: Given: $\Sigma\bar{X} = 595.8, \Sigma S = 8.28, n = 18$

$$\bar{\bar{X}} = \frac{\Sigma\bar{X}}{k} = \frac{595.8}{18} = 33.1$$
$$\bar{S} = \frac{\Sigma S}{k} = \frac{8.28}{18} = 0.46$$

$\bar{X}$ *Chart*

$\bar{\bar{X}}$ 33.1 (Central Line)

Control Limits

UCL $= \bar{\bar{X}} + A_1\bar{\sigma}$
= 33.1 + (0.949) (.46)
= 33.1 + .43675
= 33.53

LCL $= \bar{\bar{X}} - A_1\bar{\sigma}$
= 33.1 - (0.949) (.46)
= 33.1 - .43675
= 32.66

σ Chart

Control Limits

$$\bar{S} = 0.46$$
$$UCL = B_4\bar{S} = 1.72 \times 0.46 = 0.7912$$
$$LCL = B_3\bar{S} = 0.28 \times 0.46 = 0.1288$$

13.11 SUMMARY

Statistical Quality Control is a process to control the quality of manufactured goods by way of establishing and achieving quality specification, which requires the use of tools and techniques of statistics. Further, the control charts for variables are used when the quality or characteristics of a product is capable of being measured in quantitative terms, and they are very much useful in SQC. SQC is an important application of the theory of probability and theory of sampling for the maintenance of uniform quality in a continuous flow of manufactured products. It has so many merits and limitations. Despite that, it is very useful to identify the causes of variation and to control them effectively.

13.12 GLOSSARY

Statistical Quality Control: It refers to the use of statistical techniques in controlling the quality of manufactured goods.

13.13 CHECK YOUR PROGRESS

1. Construct $\bar{X} - Chart$ and R-Chart for the following data of 12 samples with each set of 5 items:

42	42	19	36	42	51	60	18	15	69	64	61
65	45	24	54	51	74	60	20	30	109	90	78
75	68	80	69	57	75	72	27	39	113	93	94

78	72	81	77	59	78	95	42	62	118	109	109
87	90	81	84	78	132	138	60	84	153	112	136

(Given: n = 5, A_2 = 0.58, D_3 = 0, D_4 = 2.115)

2. A machine is set to deliver a packet of a given weight. 10 samples of size 5 each were recorded in the data given below:

Sample No.	1	2	3	4	5	6	7	8	9	10
Sample Mean $\overline{X}$	20	34	45	39	26	29	13	34	37	23
Sample Range (R)	23	29	15	5	29	17	21	11	90	10

Construct $\overline{X}$ chart and range chart, and point out whether the process is within control

(Conversion factors for : n = 5, A_2 = 0.58, D_3 = 0, D_4 = 2.115)

3. The following data provide the mean ($\overline{X}$) and range (R) of 10 samples having 15 items each, construct a mean chart and range chart, and comment on the process of quality:

Sample No.	1	2	3	4	5	6	7	8	9	10
Sample Mean $\overline{X}$	11.2	11.8	10.8	11.6	11.0	9.6	10.4	9.6	10.6	10.0
Sample Range (R)	7	4	8	5	7	4	8	4	7	9

(Conversion factors for: n = 5 are A_2 = 0.577, D_3 = 0 and D_4 = 2.115)

4. Thirty samples of 5 items each were taken from the output of a machine, and a critical dimension was measured. The mean of 30 samples was 0.6550 inches and the Range mean 0.0036 inches. Compute the control limits for $\overline{X}$ and R charts.

(Conversion factors for: n = 5, A_2 = 0.58, D_3 = 0, D_4 = 2.115)

5. A drilling machine bores holes with a mean diameter of 0.5230 cm. and a standard deviation of 0.0032 cm. calculate the 2-sigma and 3-sigma upper and lower control limits for means of samples of size 4, and prepare a control chart.

13.14 ANSWERS TO CHECK YOUR PROGRESS

1. $[\text{UCL}_{\bar{x}} = 106.2, \text{LCL}_{\bar{x}} = 37.0, \text{UCL}_{\bar{R}} = 125.9, \text{LCL}_{\bar{R}} = 0]$
2. $[\text{UCL}_{\bar{x}} = 41.658, \text{LCL}_{\bar{x}} = 18.342, \text{UCL}_{R} = 42512, \text{LCL}_{R} = 0]$
3. $[\text{UCL}_{\bar{x}} = 41.2951, \text{LCL}_{\bar{x}} = 7.0249, \text{UCL}_{R} = 13.3245, \text{LCL}_{R} = 0]$
4. [(i) .6570, .6529 (ii).0076, 0]
5. $\begin{bmatrix} 2-\text{sigma limits} : \text{UCL} = 0.5262; \text{LCL} = 0.5198 \\ 2-\text{sigma limits} : \text{UCL} = 0.5278; \text{LCL} = 0.5182 \end{bmatrix}$

13.15 TERMINAL QUESTIONS

1. Write down various types of control charts for variables.
2. What do you mean by statistical quality control? Explain.
3. Explain the procedure for the computation of control limits in 3-sigma control charts.
4. Plot a control chart ($\bar{X}$-chart) by using imaginary figures.
5. Distinguish between $\bar{X}$ and σ charts.

13.16 SUGGESTED READINGS

1. Basic Statistics – Goon, Guptha and Dasgupta – World Press Limited Calcutta.
2. Fundamentals of Business Statistics – Sanchethi and Kappor.
3. Quantitative Methods in Management – Srivastava, Shenoy and Guptha.
4. Business Statistics – Guptha and Guptha

UNIT: 14 CONTROL CHARTS FOR ATTRIBUTES

Structure

OBJECTIVES

After studying this unit, you will be able to understand:

- Control charts for attributes; and
- Acceptance sampling

14.1 INTRODUCTION

In the preceding unit, you have studied that the control charts are the graphic devices developed by Walter A. Shewhart for detecting unnatural pattern of variation in the production process and determining the permissible limits of probability and sampling. They are of two types, depending on whether a given quality or characteristics of a product is measurable or not, like control charts for variables and control charts for attributes. One of them has already been discussed in the preceding unit and the other *i. e.,* control charts for attributes, will be discussed along with the acceptance sampling in this unit.

14.2 CONTROL CHARTS FOR ATTRIBUTES

These charts are used when the quality or characteristics of a product cannot be measured in quantitative terms, and the data is studied based on totality of attributes, like defective and non-defective. Such charts are of three types:

14.2.1 p-chart (Fraction Defective Chart)

This chart is constructed for controlling the quality standard in the average fraction defective of the products in a process when the observed sample items are classified into defectives and non-defectives.

Procedure: The construction of a p-chart involves the following steps:

(i) Find the fraction defective or proportion of defective in each sample, i.e.,

$$p_1, p_2, p_3 \dots p_k.$$

(ii) Find the mean of the fraction defectives by using the formula:

$$\bar{p} = \frac{Total\ No.of\ Defectives}{Total\ No.of\ Units\ Inspected},$$

$$\bar{q} = 1 - \bar{p}$$

Alternatively, the value of $\bar{p}$ can also be calculated as:

$$\bar{p} = \frac{p_1, p_2, p_3 \dots p_k}{k}$$

$$where, k = No.of\ samples$$

The value of $\bar{p}$ represents the central line of the p-chart

(iii) Determine the control limits by using the formula:

$$Control\ Limits = \bar{p} \pm 3\sqrt{\frac{\bar{p}.\bar{q}}{n}}$$

$$\therefore \qquad LCL = \bar{p} - 3\sqrt{\frac{\bar{p}.\bar{q}}{n}}$$

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}.\bar{q}}{n}}$$

The value of LCL cannot be negative, and in such a case, it would be reduced to zero.

(iv) Construct the p-chart by plotting the sample number on x-axis and sample fraction defectives, UCL, LCL, and central line on the y-axis.

(v) Interpret the p-chart. If all the sample fraction defective (p) fall within the control limits, the process is in a state of control otherwise it is beyond the control.

Note:

(1) If the number of defectives is small, then p-chart should be constructed by finding the percentage of defectives.

(2) This chart is especially useful when the size of the sample (n) is unequal. In such a case the value of n can be obtained by dividing the defective units in all the samples by the numbers of samples.

(EQUAL SAMPLE SIZE)

Example 14.1: The following data refers to visual defects found during the inspection of the first 10 samples size 100 each from a lot of two-wheelers manufactured by an automobile company:

Sample No.	1	2	3	4	5	6	7	8	9	10
No. of defectives	5	3	3	6	5	6	8	10	10	4

Construct a control chart for the fraction defective. What conclusions can you draw from the control chart?

Solution: We are given: n = size of sample = 100, k = No. of samples = 10

Computation of Fraction Defectives

Sample No. (k)	Size of Sample (n)	No. of defectives (d)	Fraction defectives d/n
1	100	5	5/ 100 = 0.05
2	100	3	3/ 100 = 0.03
3	100	3	0.03
4	100	6	0.06
5	100	5	0.05
6	100	6	0.06
7	100	8	0.08
8	100	10	0.10
9	100	10	0.10
10	100	4	0.04
$k = 10$	1,000	$\Sigma d = 60$	0.60

$$\bar{p} = \frac{Total\ No.of\ Defectives}{Total\ No.of\ Units}$$
$$= \frac{60}{1000} = 0.06$$
$$\Rightarrow \quad \bar{q} = 1 - \bar{p}$$
$$= 1 - 0.06$$
$$= 0.94$$

The value of $\bar{p}$ represents the central line

Control Limits for $\bar{p}$ - chart

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}.\bar{q}}{n}}$$

$$= 0.06 + 3\sqrt{\frac{.06 \times .94}{100}}$$

$$= 0.06 + 3(0.0237)$$

$$= 0.06 + .0711$$

$$= 0.1311$$

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}.\bar{q}}{n}}$$

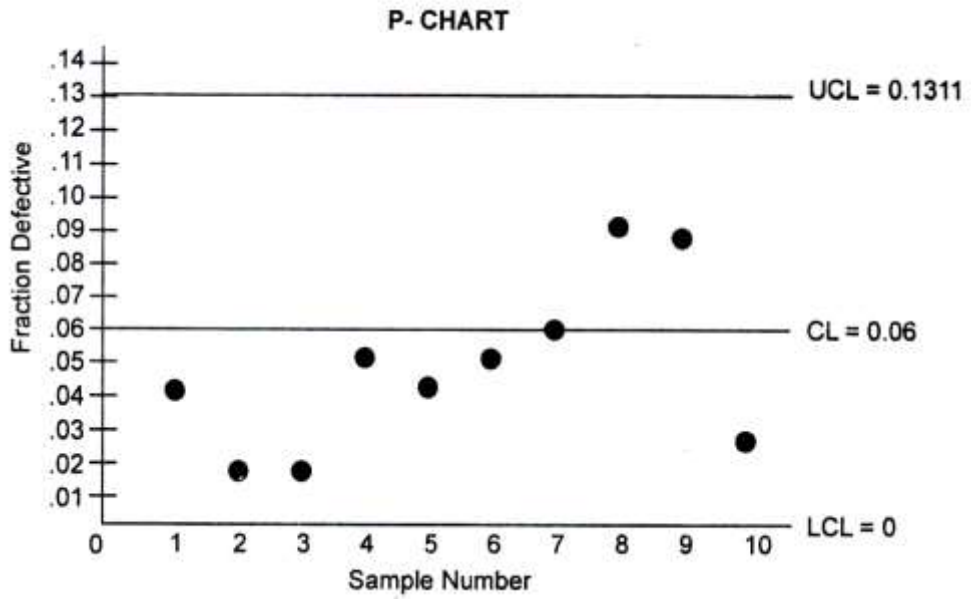
$$= 0.06 - 3\sqrt{\frac{.06 \times .94}{100}}$$

$$= 0.06 - 3(0.0237)$$

$$= 0.06 - .0711$$

$$= -0.0111 = 0.$$

Since, the fraction defective cannot be negative, LCL is taken as zero. The fraction defective chart ($\bar{p}$ - chart) is shown below:



The above chart shows that all the points lie within the control limits. This suggests that the process is in control.

(UNEQUAL SAMPLE SIZE)

Example 14.2: The number of defective needles of sewing machine has been given in the following table on the basis of daily inspection. Prepare 'p-chart' and state whether the production process is in control.

Day	1	2	3	4	5	6	7	8	9	10
No. of needles inspected	90	60	70	100	120	50	100	110	100	100
No. of defective needles	5	12	7	3	6	5	10	6	8	25

Solution: Computation of Control limits for p-chart

Day	No. of needles Inspected	No. of defectives	Percentage of defective needles
1	90	5	5.56
2	60	12	20.00
4	70	7	10.00
3	100	3	3.00
5	120	6	5.00
6	50	5	10.00
7	100	10	10.00
8	110	6	5.45
9	100	8	8.00
10	100	25	25.00
$k = 10$	900	87	

$$\bar{p} = \frac{Total\ No.of\ Defectiver\ Needles}{Total\ No.of\ Items\ Inspected}$$
$$= \frac{87}{900}$$
$$= 0.0967$$

$\bar{p}$ in the percentage form = 9.67

The value of $\bar{p}$ represents the central line.

Control Limits for $\bar{p}$ – chart

$$\bar{p} \pm 3\sqrt{\frac{\bar{p}.\bar{q}}{n}}$$

Where, $\bar{p} = 0.0967$

$\bar{q} = 1 - 0.0967 = 0.9033$

$$n = \frac{Total\ No.of\ Items\ Inspected}{k}$$

$$= \frac{900}{10} = 90$$

Substituting the values, we get

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}.\bar{q}}{n}}$$

$$= 0.0967 + 3\sqrt{\frac{0.0967 \times 0.9033}{90}}$$

$$= 0.0967 + 3 \times 0.0312$$

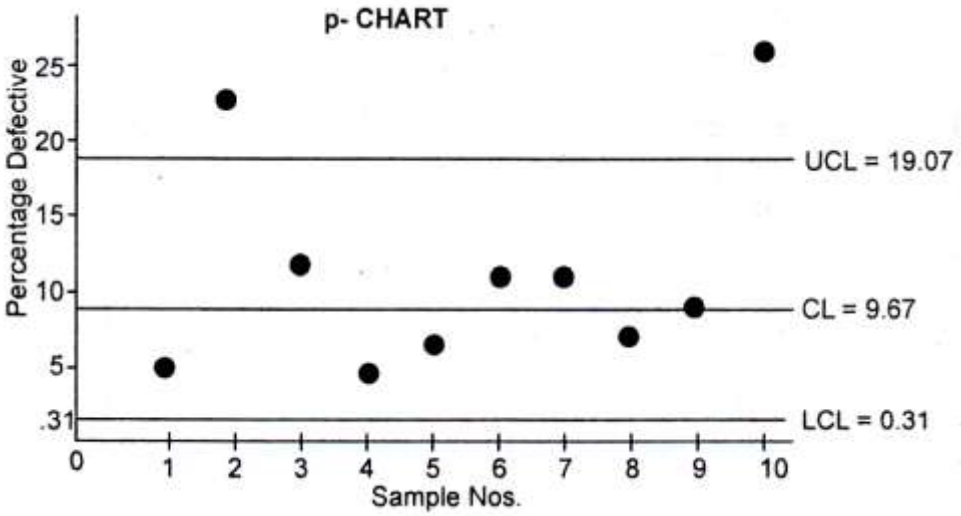
$$= 0.1903\ or\ 19.03\%$$

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}.\bar{q}}{n}}$$

$$= 0.0967 - 3\sqrt{\frac{0.0967 \times 0.9033}{90}}$$

$$= 0.0967 - 3 \times 0.0312$$

$$= 0.0031\ or\ 0.31\%$$



The above chart shows that although out of 10 points 8 points are within the control limits but the points of sample number 2 and 10 are outside the upper limit. This suggests that the process is not in control.

Example 14.3: Construct a control chart for the proportion of defectives obtained in repeated samples of size 100 from a process which is considered to be under control when the average proportion of defective p is equal to 0.20. Draw the central line and the upper and lower control limits on graph paper.

Solution: We are given:

$\bar{p}$ = Average fraction defective$= 0.2$

$$n = 100$$

$$\bar{q} = 1 - \bar{p} = 1 - 0.2 = 0.8$$

Central Line $= \bar{p} = 0.2$

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}.\bar{q}}{n}}$$

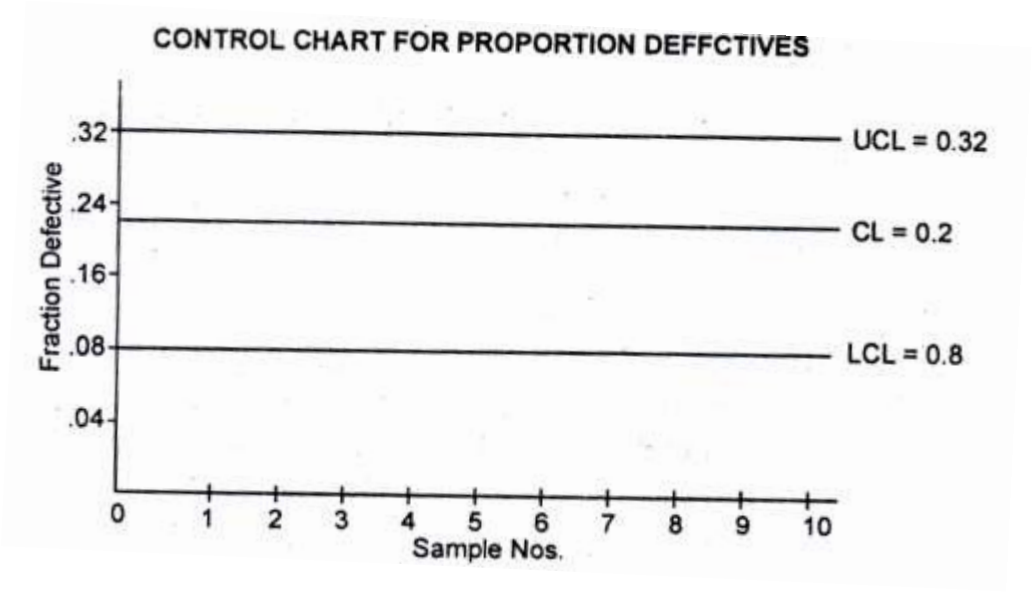
$$= 0.2 + 3\sqrt{\frac{0.20 \times 0.80}{100}}$$

$$= 0.2 + 3 \times (0.04) = 0.32$$

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}.\bar{q}}{n}}$$

$$= 0.2 - 3\sqrt{\frac{0.20 \times 0.80}{100}}$$

$$= 0.2 - 3 \times (0.04) = 0.08$$



Example 14.4: A daily sample of 30 items was taken over 14 days to establish control limits. If 21 defectives were found, what should be the upper and lower control limits for the proportion of defectives?

Solution: No. of sample $(k) = 14$
Size of the sample $(n) = 30$
Σd *i.e.*, number of defectives $= 21$

$$\bar{p} = \text{Average fraction defective}$$
$$= \frac{21}{14 \times 30} = 0.05$$
$$\bar{q} = 1 - \bar{p}$$
$$= 1 - 0.05 = 0.95$$

Control limits for p-chart:
Control Line $= \bar{p} = 0.05$

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}.\bar{q}}{n}}$$
$$= 0.05 + 3\sqrt{\frac{(.05)(.95)}{30}} = 0.17$$
$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}.\bar{q}}{n}}$$
$$= 0.05 - 3\sqrt{\frac{(.05)(.95)}{30}} = 0.69$$

The negative value of LCL is taken as zero.

14.2.2 np-Chart (Number of Defective Chart)

This chart is constructed for controlling the quality standard of attributes in a process where the sample size is equal, and it is required to plot the number of defectives (np) in samples instead of fraction defectives(p).

Procedure: The construction of np-chart involves the following steps:

(i) Find the average number of defectives $(n\bar{p})$

$$n\bar{p} = \frac{Total\ no.of\ Defectives}{Total\ no.of\ Samples} = \frac{\Sigma d}{k}$$

The value of $n\bar{p}$ represents the central line.

(ii) Find the value of $\bar{p}$ by using the formula:

$$\bar{p} = \frac{n\bar{p}}{n}$$
$$\Rightarrow \quad \bar{q} = 1 - \bar{p}$$

Alternatively, The value of $\bar{p}$ can also be calculated by using the formula:

$$\bar{p} = \frac{\Sigma d}{n \times k}$$

(iii) Determine the control limits by using the formula:

$$\therefore \quad UCL = n\bar{p} + 3\sqrt{n\bar{p}\bar{q}}$$
$$LCL = n\bar{p} - 3\sqrt{n\bar{p}\bar{q}}$$

The value of LCL cannot be negative, and in such a case, it would be reduced to zero.

(iv) Construct np-chart by plotting the sample number on x-axis and sample number of defectives, UCL, LCL, and control line (CL) on the y-axis.

(v) Interpret np-chart. If all the sample number of defectives fall within the control limits, the process is in a state of control otherwise it is beyond control.

Note: The construction and interpretation of the number of defective chart, i.e., *np* chart, is similar to that of *p*-chart. In the *np*-chart, the central line is drawn at *np* instead of *p* and the actual number of defectives (np) in samples of fixed size n is plotted instead of the fraction of defectives.

Example 14.5: An inspection of 10 samples of size 400 each from 10 lots reveal the following numbers of defectives:

17	15	14	26	9	4	19	12	9	15

Calculate the control limits for the number of defective units. Plot on the graph and state whether the process is under control or not.

Solution: We are given, n =400

k = (No. of samples) = 10,

$$\bar{p} = \text{Average fraction defectives}$$
$$= \frac{140}{10 \times 400}$$
$$= 0.035$$
$$\bar{q} = 1 - 0.035$$
$$= 0.965$$

$$Also, n = 400$$

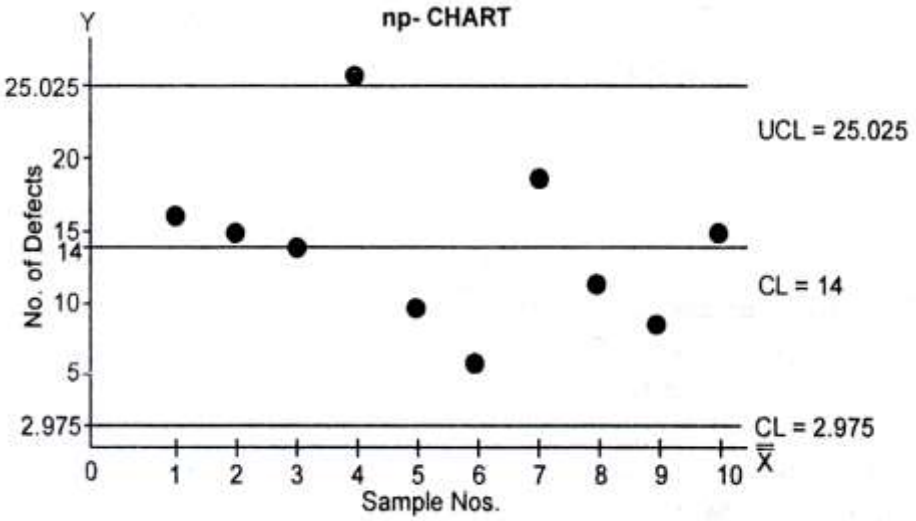
$$\therefore \quad n\bar{p} = 400 \times .035 = 14$$

The value of $n\bar{p}$ represents the central line

Control Limits for np-chart

$$UCL = n\bar{p} + 3\sqrt{n\bar{p}\bar{q}}$$
$$= 14 + 3\sqrt{400 \times 0.035 \times 0.965}$$
$$= 14 + 3(3.675)$$
$$= 14 + 11.025$$
$$= 25.025$$
$$LCL = n\bar{p} - 3\sqrt{n\bar{p}\bar{q}}$$

$$= 14 - 3\sqrt{400 \times 0.035 \times 0.965}$$
$$= 14 - 3 \times 3.675$$
$$= 2.975$$



The above chart shows that although out of 10 points 9 points are within the control limits but, the point for sample 4 is outside the UCL. This suggests that the process is not in control.

Example 14.6: In a certain sampling inspection, the numbers of defectives are found in 10 samples of 100 each are given below:

16	18	11	18	21	10	20	18	17	21

Do these indicate that the quality characteristics inspected is under statistical control?

Solution: Here, we use $n\bar{p}$ chart to find whether the quality characteristics under inspection are in a state of control or not.

We are given: n = 100k = 10

Σd= Total no. of Defectives = 170

$$\bar{p} = \frac{170}{100 \times 10} = 0.17,$$
$$\bar{q} = 1 - 0.17 = 0.83$$

$Also, n = 400$

$\therefore \quad n\bar{p} = 100 \times .17 = 17$

The value of $n\bar{p}$ represents the central line

Control Limits for $n\bar{p}$-chart

$$UCL = n\bar{p} + 3\sqrt{n\bar{p}\bar{q}}$$
$$= 17 + 3\sqrt{100 \times 0.17 \times 0.83}$$
$$= 17 + 11.268$$
$$= 28.268$$
$$LCL = n\bar{p} - 3\sqrt{n\bar{p}\bar{q}}$$
$$= 17 - 3\sqrt{100 \times 0.17 \times 0.83}$$
$$= 17 - 11.268$$

$$= 5.732$$

Since, none of the points is lying outside the lower and upper control limits, the process is in a state of statistical control.

Example 14.7: It was found that the production process is termed "Controlled" in a sample size of 10 units each when the average number of defective is 1.2. What control limits will you establish for a control chart of a sample size of 10 units each?

Solution: Here, we use $n\bar{p}$ the chart is used to find out whether quality characteristics are under inspection are in a state of control or not.

We are given: $n\bar{p} = 1.2, n = 10$

$$\bar{p} = \frac{n\bar{p}}{n}$$

$$= \frac{1.2}{10} = 0.12,$$

$$\bar{q} = 1 - \bar{p}$$

$$= 1 - 0.12 = 0.88$$

Control Limits for $n\bar{p}$-chart

$$UCL = n\bar{p} + 3\sqrt{n\bar{p}\bar{q}}$$

$$= 1.2 + 3\sqrt{10 \times 0.12 \times 0.88}$$

$$= 1.2 + 3(1.027)$$

$$= 4.281$$

$$LCL = n\bar{p} - 3\sqrt{n\bar{p}\bar{q}}$$

$$= 1.2 - 3\sqrt{10 \times 0.12 \times 0.88}$$

$$= 1.2 - 3.081$$

$$= -1.881$$

14.2.3 C-Chart (Number of Defects Per Unit Chart)

This chat is used for the control of number of defects per unit say a piece of cloth/glass/paper/bottle which may contain more than one defect. The inspection unit in this chart will be a single unit of product. The probability of occurrence of each defect tends to remain very small. Hence, the distribution of the number of defects may be assumed to be a Poisson distribution with Mean = Variance.

Procedure: The construction of a C-chart involves the following steps:

(i) Determine the number of defects per unit (C) in the samples of equal size.

(ii) Find the mean of the number of defects counted in several units by using the formula:

$$\bar{C} = \frac{\Sigma C}{K}$$

Where, K = Total No. of Units Inspected.

The value of $\bar{C}$ represents the central line of the C-chart.

(iii) Determine the control limits by using the formula:

Control Limits $= \bar{C} \pm 3\sqrt{\bar{C}}$.

$$UCL = \bar{C} + 3\sqrt{\bar{C}}$$

$$LCL = \bar{C} - 3\sqrt{\bar{C}}$$

The value of LCL cannot be negative, and in such a case it would be reduced to zero.

(iv) Construct C-chart by plotting the sample numbers on the x-axis and number of defects observed per unit, LCL, UCL, and CL on the y-axis.

(v) Interpret C-Chart. If the observed values of the number of defects per unit fall within the control limits, the process is a state of control otherwise it is beyond the control.

Although the application of C-chart is somewhat limited compared with $\bar{X}$ and R charts, yet a number of practical situations exist in many industries where C-chart is used. The following are the field of applications of C-chart.

(1) Number of defects of all kinds of aircraft assembled finally.

(2) Number of defects counted in a roll of coated paper, sheet of photographic film, bale (or pieces) of cloth, etc.

Example 14.8: Ten pieces of cloth out of different rolls of equal length contained the following number of defects:

1	2	5	0	6	0	9	4	4	3

Draw a control chart for the number of defects and state whether the process is in a state of statistical control.

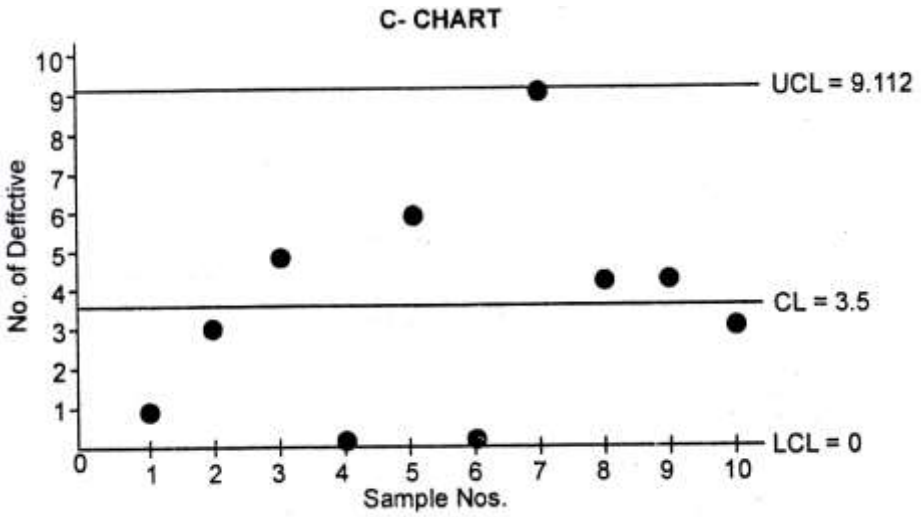
Solution: We have N=10, and C =No. of defects = 35

$$\bar{C} = \frac{\Sigma C}{N} = \frac{35}{10} = 3.5$$

The value $\bar{C}$ represents the central line.

Control Limits for C-Chart

$$UCL = \bar{C} + 3\sqrt{\bar{C}}$$
$$= 3.5 + 3\sqrt{3.5}$$
$$= 3.5 + 5.612$$
$$= 9.112$$
$$LCL = \bar{C} - 3\sqrt{\bar{C}}$$
$$= 3.5 - 3\sqrt{3.5}$$
$$= 3.5 - 5.612$$
$$= -2.112 = 0$$



The above chart shows that all the plotted points are within the two control limits. This suggests that the process is in control.

Example 14.9: The numbers of defects of 20 items are given below:

Item No.	1	2	3	4	5	6	7	8	9	10
No. of defects	2	0	4	1	0	8	0	1	2	0
Item No.	11	12	13	14	15	16	17	18	19	20
No. of defects	6	0	2	1	0	3	2	1	0	2

Devise a suitable control chart and draw your conclusion.

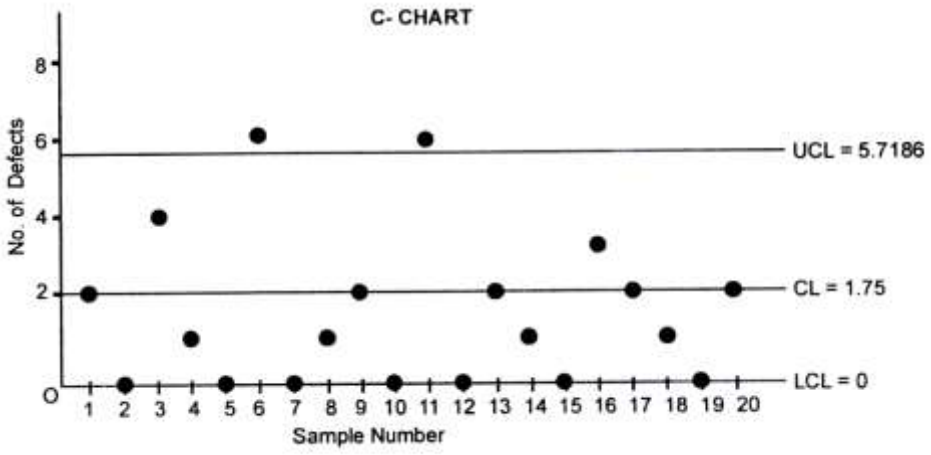
Solution: As the number of defects per unit is given, the suitable control chart is C-chart. We have N =20, and C =No. of defects = 35

$$\bar{C} = \frac{\Sigma C}{N} = \frac{35}{20} = 1.75$$

The value $\bar{C}$ represents the central line.

Control Limits for C-Chart

$$UCL = \bar{C} + 3\sqrt{\bar{C}}$$
$$= 1.75 + 3\sqrt{1.75}$$
$$= 5.7186$$
$$LCL = \bar{C} - 3\sqrt{\bar{C}}$$
$$= 1.75 - 3\sqrt{1.75}$$
$$= -2.218 = 0$$



The above chart shows that although out of 20 plotted points 18 points are within the control limits but the points for sample 6 and sample 11 are outside the UCL. This suggests that the process is not control.

14.3 ACCEPTANCE SAMPLING

Another major area of statistical quality control is product control or acceptance sampling. Product control is concerned with the inspection of manufactured products. The items are inspected to determine whether to accept a lot of items conforming to the standards of equality or to reject a lot as non-conforming. Here, the decision is arrived at through sampling. That is why product control is called acceptance sampling. According to Simpson and Kafka, "Acceptance sampling is concerned with the decision to accept a mass of non-conforming to quality. The decision is reached through sampling."

14.4 RISKS IN ACCEPTANCE SAMPLING OR PRODUCT CONTROL

There are two types of risks in acceptance sampling or product quality control which are given below:

14.4.1 Producer's Risk

Sometimes, it happens that in spite of good quality, the sample taken may show defective units as such the lot will be rejected. In spite of good quality, the lot is rejected; such a type of risk of rejection is known as producer's risk. In other words, the probability of rejecting a lot which has actually been found satisfactory by the producer according to acceptable quality level is known as producer's risk. Thus, the risk of rejecting a lot of good items is known as the producer's risk.

14.4.2 Consumer's Risk

Sometimes, it may happen that the quality of the lot is not good but the sample results show good quality units as such the consumer has to accept a defective lot. Such a risk is known as consumer's risk. In other words, the probability of accepting a lot which has actually been satisfactory by the

consumer according to pre-determined standard is known as consumer's risk. Thus, the risk of accepting a lot of bad items is known as the consumer's risk.
The consumer and producer both decide the acceptance standard of the lot. This is as known of Acceptable Quality Level (AQL) or Lot Tolerance Percentage Defective (LTPD).

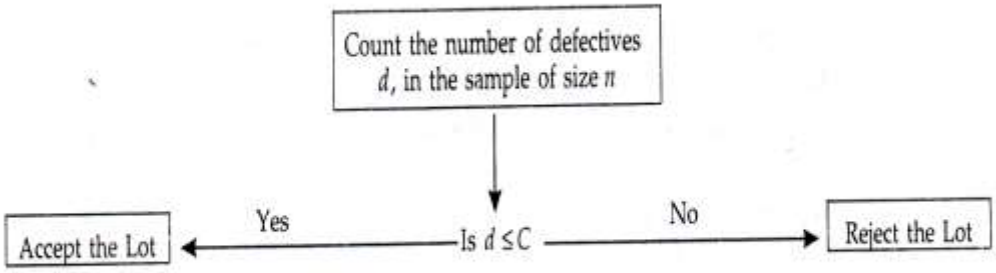
14.5 TYPES OF SAMPLING INSPECTION PLANS

Acceptance sampling is based on sampling. After the inspection of samples, the decision is made about the acceptance or rejection of a lot. In acceptance sampling, the number of samples and their order play a significant role. To frame the rules for acceptance or rejection of a lot acceptance sampling plan is prepared.

Three types of sampling plans are frequently used in acceptance sampling, as under:

14.5.1 Single Sampling Plan

Under a single sampling plan, a sample of n items are first chosen at random from a lot of N items. If the sample contains c or a few defectives, then a lot is accepted, while if it contains more than c defectives, then a lot is rejected (c is known as 'acceptance number'). The single sampling plan is shown in the following chart:



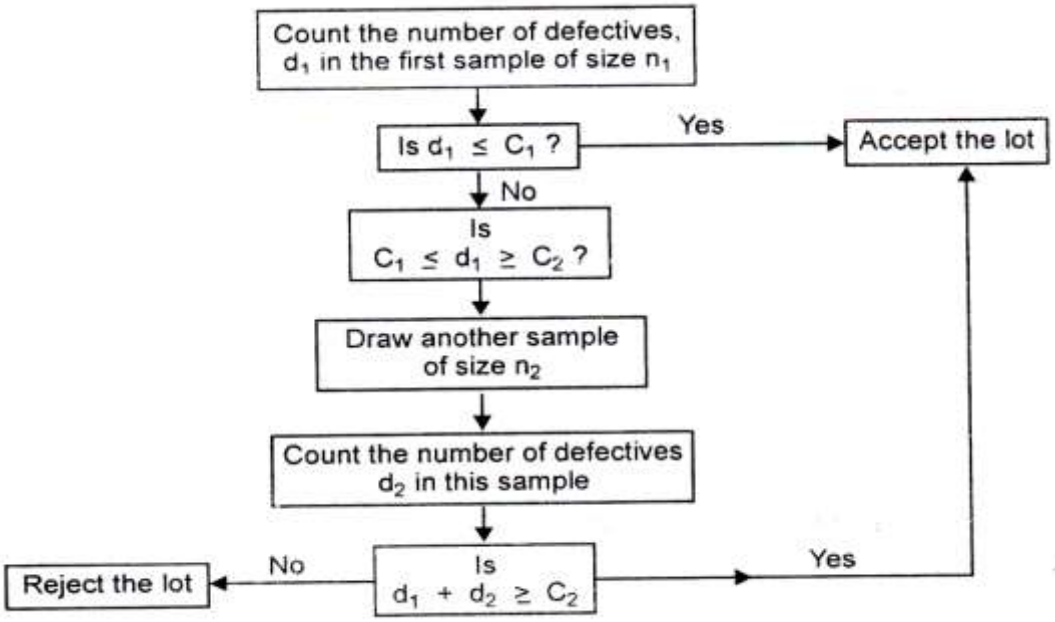
Example 14.10: A lot of goods consisting of 500 items is submitted for inspection, for which the tolerable acceptance number of defectives is 10. Take a sample of 30 items you find that there are 8 defectives. State whether the lot should be accepted or rejected for marketing.

Solution: We have c i.e., acceptance number =10
and d i.e., the number of defective observed in the sample =8.
Thus, $d < c \quad (i.e., 8 < 10)$
Since, $d < c$, the lot under consideration should be accepted.

14.5.2 Double Sampling Plan

Under this sampling plan, a sample of n_1 items are first chosen at random from the lot of size N. If the sample contains, say, c_1 or a few defectives, the lot is accepted; if it contains more than c_2 defectives, the lot is rejected. If, however, the number of defectives in the sample exceedsc_1, but it is not more than c_2, a second sample of n_2 items are taken from the same lot. Now, the total number of defectives in the two samples together does not exceed c_2, the lot is accepted; otherwise,

it is rejected. (c_1 is known as the acceptance number for the first sample and c_2, the acceptance number for both the samples taken together. The Double sampling plan is shown in the following chart:



Example 14.11: A double sampling plan has the following facts as specified:
N = 5,000, n_1 = 50, c_1 = 4, n_2 = 100 and c_2 = 6 Execute the plan.

Solution: Execution of the double sampling plan involves the following steps:

(i) Inspect all 50 items of the first sample after taking the same number at random from the lot of 5,000 items.

(ii) Accept the lot, the number of defectives observed from the sample (d_1) is less than or equal to 4 ($i.e.\, c_1$) if $d_1 > 6$ ($i.e.\, c_2$) reject the lot.

(iii) If the number of defectives in the sample thus observed $i.e.\, d_1$ is more than 4 (c_1) but not more than 6 (c_2)Inspect all 100 items of the second sample.

(iv) If now the total number of defectives ($d_1 + d_2$) observed in the combined sample of 150 items ($n_1 + n_2$) is less than 6 ($i.e.\,, c_2$), accept the lot. If it exceeds 6, then reject the lot.

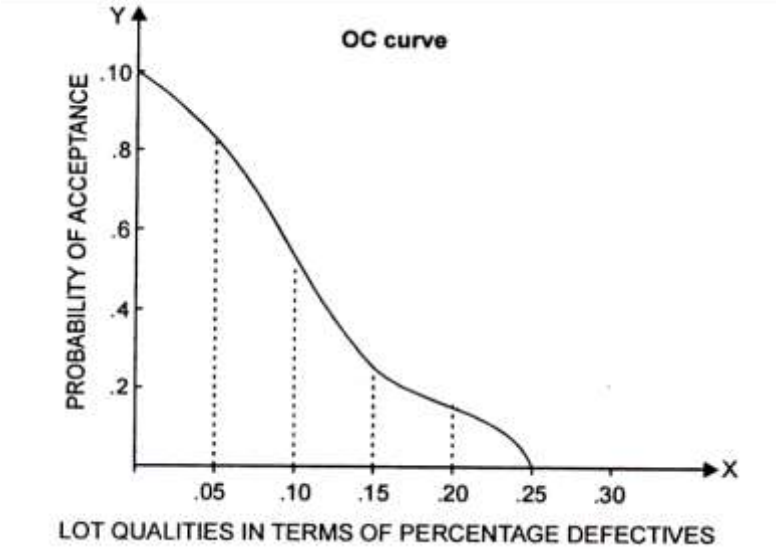
14.5.3 Multiple or Sequential Sampling Plan

Under this sampling plan, a decision to accept or reject a lot is taken after inspecting more than two samples of small size each. In this plan, units are examined one at a time, and after examining each unit decision is taken. However, such plans are very complicated and hence rarely used in practice.

14.6 OPERATING CHARACTERISTIC CURVE OF AN ACCEPTANCE SAMPLING PLAN

This is a graphic measure of assessing the ability of a sampling plan in distinguishing between good and bad items. It depicts the relationship between the probabilities of acceptance of a lot Pα (p) for different lot quality expressed in terms of percentage defectives. In the construction of the OC curve, we take $p\ i.e.$ lot of qualities in terms of percentage defectives along the x-axis and $P_2\ (p)\ i.e.$, probability of acceptance of a lot along the y-axis.

There is always an operating characteristic above (OC Curve) corresponding to any given sampling plan. A typical OC Curve has the following shape:



14.7 LOT QUALITIES IN TERMS OF PERCENTAGE DEFECTIVE

In the above figure, it has been assumed that the acceptable and rejectable qualities are measured as the proportion of items that are defective and are $P_\alpha = 0.05$; and $P_r = 0.15$. From the OC curve, it must be seen that the probability of acceptance of a lot of the quality 0.05 is a little less than 0.9, and the probability of rejection of a lot of the quality 0.15 is a little more than 0.1. This shows that the chance of rejection of good products, which is the producer's risk, is a little more than 0.1, and the chance of acceptance of bad products, which is the consumer's risk, is a little more than 0.1. Thus, the risks of both producer and consumer are more or less the same. The steepness of the curve depends on the sample size. The larger the sample, the steeper the OC curve. The position of OC curve is determined by the maximum number of defective items allowable for acceptance, called the acceptance number. If the curve is shifted to the left or right, according as the acceptance number is made smaller or larger.

Example 14.12: From the following data relating to a single sampling plan, determine the probability of acceptance at 0.5%, 7.5%, 1%, 2%, 5%, 10% and 15% defectives in the lot quality and fit an OC Curve to represent the data:

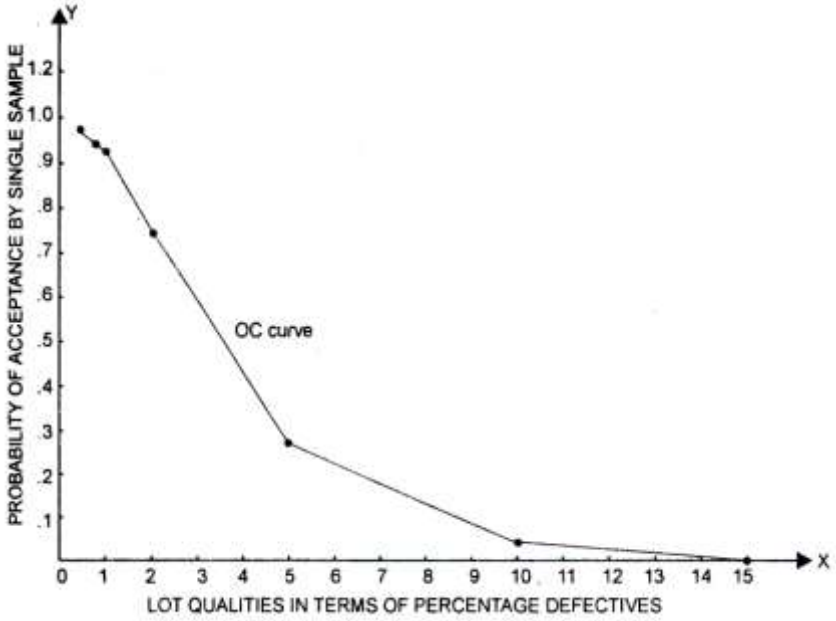
$$N = 1{,}000, n = 50, C = 1$$

Solution: It is case of single sampling plan. Since the number of tolerable defective or $c = 1$, the lot will be accepted if the sample gives 0 or 1 defective.

Computation of cumulative Probabilities using Poisson Distribution

% defective in lot	Mean defective (m)	$P(0) = e^{-m}$	$P(1) = m.e^{-m}$	$P_\alpha(p) = P(0) + P(1)$
0.50	$\frac{50}{100} \times .5 = 2.5$	0.7788	0.1947	0.9735
0.75	$\frac{50}{100} \times .75 = .38$	0.6839	0.2599	0.9438
1.00	$\frac{50}{100} \times 1 = .50$	0.6065	0.3033	0.9098
2.00	$\frac{50}{100} \times 2 = 1.00$	0.3678	0.3678	0.7358
5.00	$\frac{50}{100} \times 5 = 2.50$	0.0821	0.2053	0.2874
10.00	$\frac{50}{100} \times 10 = 5.00$	0.0070	0.0350	0.0420
15.00	$\frac{50}{100} \times 15 = 7.50$	0.0006	0.0045	0.0051

Now, you will represent the probabilities of the acceptance of the lot with the given percentage of defectives by a single sampling plan. This is drawn below:



The above OC curve indicates that out of the 1,000 items (Since N=1000) inspected 974 (.9735 x 1000) items will be accepted and 26 items rejected will be with 0.5% defectives.

14.8 SUMMARY

In total, the charts for attributes are used when the quality or characteristics of a product cannot be measured in quantitative terms and the data is studied on the basis of totality of attributes like defective and non-defectives. Further, acceptance sampling is another major area of statistical quality control. It is concerned with the inspection of manufactured products by covering the risks on the part of producers as well as consumers.

14.9 GLOSSARY

Control Chart for Attributes: Charts are used when the qualities or characteristics of a product cannot be measured in quantitative terms.

14.10 CHECK YOUR PROGRESS

1 If the average fraction defective of a large sample of products is 0.1537, calculate the control limits for fraction defectives (Given that the size of each sample is 200).

2 An inspection of 9 samples of size 100 each from 9 lots reveals the following number of defective units:

Item No. (each of 100 items)	1	2	3	4	5	6	7	8	9
No. of defectives:	12	7	9	8	10	6	7	11	8

Construct a suitable control chart and give your conclusion.

3 In a manufacturing concern of radio production lot of 250 items are inspected at a time. 20 samples taken are different in trades and defectives noted are given below. Draw a suitable control chart.

Lot No. (each of 250 items)	1	2	3	4	5	6	7	8	9	10
No. of defectives:	25	47	23	36	24	34	39	32	35	22
Lot No.	11	12	13	14	15	16	17	18	19	20
No. of defectives: (each of 250 Items)	45	40	32	35	21	40	15	28	23	42

4 In a certain sampling inspection, the number of defectives found in 21 samples of 100 each is given below:

5, 7, 9, 7, 8, 13, 8, 4, 8, 4, 3, 7, 7, 12, 15, 5, 13, 4, 3, 10, 8.

Does this indicate that the quality characteristics under inspection are under statistical

5 During an examination of equal length of cloth, the following number of defects are observed:

2, 3, 4, 0, 5, 6, 7, 4, 3, 2.

Draw a control chart for the number of defects and comment whether the process is under control or not.

14.11 ANSWERS TO CHECK YOUR PROGRESS

1. [UCL= 0.17738, LCL =0.12952 or 0]
2. [CL =8.66, UCL =9.38, LCL =7.94]
3. [CL = =31.9 UCL = 32.84, LCL =16.16]
4. [UCL = 16.48, LCL = -0.48=0]
5. [CL = $\bar{C}$ = 3.6, UCL = 9.292, LCL = 0]

14.12 TERMINAL QUESTIONS

1. Write down various types of control charts for attributes.
2. What do you mean by acceptance sampling? Explain.
3. Explain the procedure for computing fraction defectives in control charts.
4. Plot a control chart (p-chart) by using imaginary figures.
5. Distinguish between the control chart and acceptance sampling.

14.13 SUGGESTED READINGS

1. Basic Statistics – Goon, Guptha and Dasguptha – World Press Limited – Calcutta.
2. Fundamentals of Business Statistics – Sanchethi and Kappor.
3. Quantitative Methods in Management – Srivastava, Shenoy and Guptha.
4. Business Statistics – Guptha and Guptha.

UNIT 15: Chi-Square Test

Structure

OBJECTIVES

After studying this unit, you shall be able to understand:

- Concept of chi-square distribution.
- Uses of chi-square distribution.

15.1 INTRODUCTION

As you know, various statistical tools are applied for testing hypotheses. These statistical tests can be categorized into two different groups—parametric tests and non-parametric tests. Parametric tests are those tests that are based on parameters of the population. To apply parametric tests, certain assumptions regarding population distribution must be fulfilled. This becomes a limitation for implementing these tests. To avoid these limitations, we may use another category of tests known as nonparametric tests.

In this block, you will study different non-parametric tests. The chi-square test is the most popular form of non-parametric test, but can also be used as a parametric test. In this unit, you will study in detail the theoretical concept and various uses of the chi-square test.

15.2 MEANING &USES OF CHI-SQUARE TEST

The chi-square test is an important test among the several tests of significance developed by statisticians. This test was propounded by Prof. Karl Pearson in 1900. The chi-square test is represented by the symbol χ^2 (pronounced as Ki-square) and owes its origin to the Greek letter 'chi'. The chi-square test can be used as a parametric as well as a nonparametric test for comparing the variance of the population, as a test of independence, or as a test of goodness of fit. This test is mainly used in the context of sampling analysis for comparing a variance to a theoretical variance. According to Neil R. Ullman, "As a non-parametric test, it can be used to determine if categorical data shows dependency or the two classifications are independent. It can also be used to make comparisons between theoretical populations and actual data when categories are used." The quantity χ^2 describes the magnitude of discrepancy between theory and observation, i.e., with the help of the χ^2 test, we can know whether a given discrepancy between theory and observation can be attributed to chance or whether it results from the inadequacy of the theory to fit the observed facts. The χ^2 test is one of the simplest and most widely used non-parametric tests in statistical work, which can be applied in several situations. This technique is mainly used for the following purposes:

- To test the goodness of fit
- To test the significance of the association between two attributes
- To test the homogeneity or the significance of population variance

15.3 χ^2DISTRIBUTION

The probability distribution of the chi-square variable is known as the chi-square distribution. If Z_1, Z_2,……Z_K are independent random variables, each having a standard normal distribution N(0,1), then the distribution of

$\chi_k^2 = Z_1^2 + Z_2^2 +$ ……+ Z_k^2 It is called a χ^2 distribution.

Conceptually, χ^2 is a measure of discrepancy existing between the observed and expected frequencies, whose value largely depends on the number of degrees of freedom. Degrees of freedom are the number of values that we can choose freely, i.e., they are not decided by the other predetermined parameters. It means that the chi-square distribution has only one parameter, i.e., the number of degrees of freedom. For example, if it is given that the sum of three variables is equal to 50, then we are free to choose any two variables, but the third variable must be equal to 50 – (Total of two variables), because only then the sum of three variables would be equal to 50. Here degree of freedom is 2. Degrees of freedom are generally represented by v or df.

Through the χ^2 test, we can determine the extent of difference between the theory or expected value and the observed or actual value. Mathematically, it is defined as follows:

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

where: O = Observed frequencies

E = Expected frequencies

The chi-square test is particularly useful in tests involving nominal data, but can be used for higher scales also. Nominal data is the most elementary form of measurement, which partitions a set into

categories that are mutually exclusive and collectively exhaustive. For example, a population may be classified into two categories—males and females.

15.3.1 Properties of χ^2 Distribution

χ^2 distribution possesses many properties. some important properties of the distribution are as follows:

(1) χ^2 distribution is a continuous probability distribution that has the value zero at its lower limit and extends to infinity in the positive direction. The negative value of χ^2 is not possible because the differences between the observed and expected frequencies are always squared, hence, the value of χ^2 can never be negative.

(2) The exact shape of the distribution depends upon the number of degrees of freedom (v). In general, when v is small, the shape of the curve is skewed to the right, and as v gets larger, the distribution becomes more and more symmetrical and can be approximated by the normal distribution.

(3) The mean of the χ^2 distribution is given by the degrees of freedom and the variance is twice the degrees of freedom. It can be expressed as follows:

$$E(\chi^2) = \mu = v$$
$$V(\chi^2) = \sigma^2 = 2v$$

(4) The χ^2 is a sample statistic having no corresponding parameter. This makes the χ^2 distribution a non-parametric distribution.

(5) The additive property holds good for the χ^2 distribution. It means that the sum of independent χ^2 variates is also a χ^2 variate. Thus, if χ_1^2 is a χ^2 variate with v_1 d.f. and χ_2^2 is another χ^2 variate with v_2 d.f. independent of χ_1^2, then their sum $\chi_1^2 + \chi_2^2$ is also a χ^2 variate with $v_1 + v_2$ d.f.

15.3.2 Conditions for Application of χ^2 Test

A chi-square test should be used if the following conditions are satisfied:

(1) The total number of items 'N' should be large enough to guarantee a measure of similarity between the theoretically correct distribution and the sampling distribution being studied.

A generally accepted value of N is 50. Hence, the sample should contain at least 50 observations.

(2) The sample data must be drawn at random from the target population so that there is no element of bias.

(3) The experimental data or sample observation, i.e., all items or observations in the sample, should be independent of each other.

(4) The data should be expressed in original units (absolute form) for the convenience of comparison and not in relative form, like percentage or ratio or proportion, etc.

(5) There should not be less than five observations in any one cell (each data entry is known as a cell). If any group has frequencies below the accepted level, then 'pooling' of frequencies is done whereby the less frequencies are added to the preceding or succeeding frequency so that the resulting sum is more than the acceptable level. Some statisticians consider 10 as a better figure for the minimum acceptable level instead of 5.

(6) The constraints should be linear, i.e., the equations defining constraints should have no square or higher powers of frequencies.

15.4 CHI-SQUARE AS A TEST FOR COMPARING VARIANCE

The chi-square value is often used to judge the significance of population variance. It means that χ^2 test can be used to test if a random sample has been drawn from a normal population with a mean μ and variance σ_p^2. To judge the significance of population variance, it is assumed that the sample variance is equal to the population variance. Thus, the null hypothesis is taken as follows:

$$H_0 : \sigma_s^2 = \sigma_p^2$$

It is obvious that the χ^2 test is based on the χ^2 distribution, which deals with the collection of values that involve adding up squares. When we have to use chi-square as a test of population variance, we have to work out the value of χ^2 by the following formula to test the null hypothesis.

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2} (n - 1)$$

where:

σ_s^2 = variance of the sample

σ_p^2 = variance of the population

(n– 1) = degrees of freedom

n = number of items in the sample

For different degrees of freedom and significance level, critical values of χ^2 are available in the form of a table. You can find this table in the appendices of any good book on statistics. For arriving at a decision, the value of χ^2 calculated by the above-mentioned formula is compared with the tabular value of χ^2 for (n–1) degrees of freedom at a specified level of significance.

If the calculated value of χ^2 is less than the tabular value, then the null hypothesis is accepted. On the contrary, if the calculated value of χ^2 is equal to or greater than the tabular value, then the hypothesis is rejected.

One important thing to be remembered in this regard is that the chi-square distribution is not symmetrical, and all the values are positive. The use of the χ^2 test for comparing variance is based on the assumption of normal distribution of the population.

Illustration 1: The weight of 10 students is as follows:

S.No.	1	2	3	4	5	6	7	8	9	10
Weight (Kg)	38	40	45	53	47	43	55	48	52	49

Can we say that the variance of the distribution of weight of all students from which the above sample of 10 students was drawn is equal to 20 kgs? Test this at a 5 percent and 1 percent level of significance.

Solution:

First of all, we should work out the variance of the sample data, i.e. σ_s^2 which is calculated as follows:

S. No.	X_i (Weight in kgs.)	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
1	38	– 9	81
2	40	– 7	49
3	45	– 2	04
4	53	+ 6	36
5	47	0	00
6	43	– 4	16
7	55	+8	64
8	48	+1	01
9	52	+5	25
10	49	+2	04
n = 10	$\sum X_i = 470$		$\sum (X_i - \bar{X})^2 = 280$

$$\bar{X} = \frac{\sum X_i}{n} = \frac{470}{10} = 47 \text{ kgs}$$

$$\therefore \quad \sigma_s = \sqrt{\frac{(X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{280}{10-1}} = \sqrt{31.11}$$

or $\sigma_s^2 = 31.11$

$H_0 : \sigma_s^2 = \sigma_p^2$

In order to test this null hypothesis, we have to work out the value of χ^2 as follows:

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1)$$

$$= \frac{31.11}{20}(10-1) = 13.999$$

Degrees of freedom = n – 1 or 10 –1 = 9

At 5% level of significance, the tabular value of χ^2 is 16.92 and at 1% level of significance, it is 21.67 for 9 d.f., and both these values are greater than the calculated value of $\chi^{2,}$ which s 13.999. Hence, we accept the null hypothesis and conclude that the variance of the given distribution can be taken as 20 kgs at 5% as well as at 1% level of significance. In other words, the sample can be said to have been taken from a population with a variance 20 kgs.

Illustration 2: A sample of 15 bottles is randomly drawn from a certain population. The sum of squared deviation from the mean of the given sample is 55. Has this sample been drawn from a population with a variance 6?

Solution:

Given that: $n = 15, \quad \sum (X_i - \bar{X})^2 = 55, \quad \sigma_p^2 = 6$

The null hypothesis can be taken as follows:

$$H_0: \sigma_s^2 = \sigma_p^2$$

First of all, we have to calculate the sample variance as under:

$$\sigma_s^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

$$= \frac{55}{15-1} = 3.93$$

Now, we have to calculate the value of χ^2 as under:

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1)$$

$$= \frac{3.93}{6}(15-1) = 9.17$$

The tabular value of χ^2 for (n 1), i.e. (15 – 1) or 14 degrees of freedom at a 5% level of significance, is 23.7. Since the calculated value of χ^2 is less than the tabular value of χ^2, therefore null hypothesis is accepted, i.e. there is no significant difference between sample variance and population variance and the sample has been drawn from a population with variance 6.

15.5 CHI-SQUARE AS A NON-PARAMETRIC TEST

Chi-square can be used as a parametric as well as a non-parametric test. In the earlier section, you have studied that chi-square can be used to compare the sample variance with the population variance. In that situation, it is used as a parametric test because it is based on population parameter. But it is mostly used as a non-parametric test. In fact, chi-square is one of the most important and popular non-parametric tests, which is free from those assumptions that have to be fulfilled in the case of parametric tests. This test is especially useful in case of nominal data, but it can also be used for higher scales (ordinal, interval, ratio) also. As a non-parametric test, chi-square can be applied in the following situations:

15.5.1 Chi-Square as a Test of Independence

The application of the chi-square test is very useful in the field of association of attributes. In association of attributes, our objective is to find out whether the two attributes are independent or there is any association between them. For this purpose, we proceed on the null hypothesis of no association between the two attributes, i.e., the two attributes are independent. Thereafter, the value of χ^2 is calculated, and it is compared with the tabular value of χ^2. If the calculated value of χ^2 is less than or equal to its tabular value, then null hypothesis is accepted and the two attributes are assumed as independent, otherwise, if the calculated value is more than its tabular value then null hypothesis is rejected and it is assumed that there is association between the two attributes. Hence, the chi-square test is used for testing mutual independence between two attributes.

For example, suppose we are interested in knowing whether a new medicine is effective in controlling fever or not. We can take the help of the chi-square test to make this decision. In that case, first of all, we will take the null hypothesis that the two attributes, i.e. new medicine and control of fever are independent which means that new medicine is not effective in controlling fever. On this basis, we first calculate the expected frequencies and then work out the value of χ^2. If the calculated value of χ^2 is less than the tabular value at a certain level of significance for given degrees of freedom, we conclude that null hypothesis stands true which means that the two attributes are independent or not associated (i.e., the new medicine is not effective in controlling the fever). But if the calculated value of χ^2 is greater than its tabular value, our inference then would be that null hypothesis does not hold good which means the two attributes are associated and the association is not because of some chance factor but it exists in reality (i.e. the new medicine is effective in controlling the fever and as such may be prescribed). One important thing to be noted here is that χ^2 is not a measure of the degree of relationship or the form of relationship between two attributes, but is simply a technique of judging the significance of such association or relationship between two attributes.

15.5.2 Chi-Square as a Test of Goodness of Fit

Chi-square test is also used for finding out how far the observed frequency distribution conforms with the theoretical frequency distribution, i.e. goodness of fit is tested. It means that χ^2 as a test of goodness of fit is used to determine how well a theoretical distribution fits on the observed data. Many times, expected frequencies are found out with the help of mathematical techniques like laws of binomial, normal, and Poisson distributions, etc. Sometimes, it becomes very important for us to know how far actual frequencies resemble expected frequencies. This necessity mainly arises in the case of sample study. In such circumstances, we have to see how far actual frequencies obtained through sample study resemble the expected frequencies obtained through theoretical or mathematical distribution and whether the difference between them is significant or not. It is called goodness of fit.

In such a situation, we proceed on the null hypothesis of no difference between expected and actual frequencies. The criteria for acceptance and rejection are the same as for the test of independence. An acceptance of the null hypothesis states that the fit is good and a rejection conveys that the fit is poor. Another notable point in this regard is that in the case of goodness of fit, curves of expected

and actual frequencies are alike. On the other hand, if there is no goodness of fit then both curves are not alike.

As you know as a test of goodness of fit, chi-square is mainly used in sample studies. But one thing must be taken into consideration while using chi-square in different samples a too-good fit also raises question mark on its reliability. According to Chou, "It should be borne in mind that in repeated sampling, too good a fit is just as likely as too bad a fit. When the computed chi-square value is too close to zero, we should suspect the possibility that two sets of frequencies have been manipulated in order to force them to agree and therefore, the design of our experiment should be thoroughly checked.

15.5.3 Chi-Square as a Test of Homogeneity

The chi-square test is also used for testing homogeneity between different samples. For testing homogeneity, it is ascertained whether the two different samples are drawn from the same universe or not. In other words, the chi-square test is used for testing the significance of the difference between the two different sample values taken from the same difference. Hence, it is similar to the test of independence between two attributes. But at the same time, it is different from the test of independence on two points—firstly, a test of independence tries to find out if one attribute is independent of another whereas test of homogeneity tries to find out if the random samples have been drawn from the same population. Secondly, a test of independence uses a single sample whereas the test of homogeneity uses two or more samples.

15.6 STEPS IN APPLICATION OF THE CHI-SQUARE TEST

The following steps are followed to apply the χ^2 test:

- The process of the chi-square test begins with the assumption of the null hypothesis. It is assumed that there is no difference between expected and actual frequencies. Naturally, an alternative hypothesis of the difference between expected and actual frequencies is also formulated with the null hypothesis. It can be expressed as follows:

  H_o: $O_i = E_i$

  H_a: $O_i \neq E_i$

Where: O_i = Observed Frequency & E_i = Expected Frequency

Notations f_o and f_e or simply O and E are also used by some statisticians to denote observed and expected frequencies respectively.

- To calculate the value of χ^2, we must have actual and expected frequencies. Actual or observed frequencies are already available to us. In the second step, based on these actual frequencies, expected frequencies are calculated through the use of some appropriate rule depending upon the circumstances. The expected frequency is developed by assuming that a particular probability distribution applies to the concerned statistical population. Usually, in case of 2 × 2 or any contingency table, the expected frequency for any given cell is worked out as under:

$$\text{Expected frequency of any cell} = \left[\frac{(\text{Row total for the row of that cell}) \times (\text{Column total for the column of that cell})}{(\text{Grand total})}\right]$$

- In the next step, the difference between observed and expected frequencies are found out, i.e. (O–E)
- Thereafter, this difference is squared, i.e., $(O–E)^2$
- In the next step, this squared difference is divided by its expected frequency, i.e.$(O–E)^2/ E$. It should be repeated for all the cell frequencies.
- After that, the summation of $(O–E)^2 / E$ is found out. It is the required χ^2 value. It can be expressed as follows:

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

- After calculating the value of chi-square by the above-mentioned method, it is compared with the tabular value for making a decision regarding acceptance or rejection of the null hypothesis. To obtain the tabular value, we must decide on two aspects—level of significance and degrees of freedom.
- The level of significance means the maximum probable percentage of the numbers being wrong due to fluctuations of random sampling. For example, a 1% level of significance means that a maximum 1% number can be wrong due to fluctuations of random sampling. Similarly, a 5% level of significance means that a maximum 5% number can be wrong due to fluctuations of random sampling. To find out the tabular value of χ^2, the level of significance is decided by

the researcher as per his convenience and the purpose of the study. In practice 5% level of significance is more common.

- Another criterion to be decided for finding out the tabular value of χ^2 is the degree of freedom. The degree of freedom means the extent of freedom of selection. In any circumstances, when various class frequencies are given, then keeping the summation as before, the number of choices available to change class frequency is known as the desired degree of freedom. The desired degree of freedom is obtained by subtracting one from the total number of frequency. It can be expressed as follows:

  $$d.f. = n - 1$$

  If the class frequencies are arranged in rows and columns, then to find out the degree of freedom, one is subtracted from both the number of rows and number of columns because it is essential to keep the summation of rows and columns as before. In such cases, the formula for the degree of freedom becomes as follows:

  $$d.f. = (c - 1)(r - 1)$$

  Where: c = no. of columns

  r = no. of rows

- After deciding the level of significance and degree of freedom, the tabular value of χ^2 is seen at that specified level and degree. The tabular value of chi-square is the maximum limit up to which, if the calculated value is found, then it is considered to have arisen due to fluctuations of sampling.
- In the last step, a comparison is made between the calculated and tabular value of χ^2. If the calculated value is less than its tabular value, then the null hypothesis is accepted, which indicates that the difference between expected and actual frequencies is not considered as significant. On the contrary, if the calculated value is more than its tabular value, then the null hypothesis is rejected, and the difference between expected and actual frequencies is considered to be significant.

There is another notable thing in this regard is that when the degrees of freedom are greater than 30, the distribution of $\sqrt{2\chi^2}$ approximates a normal distribution, wherein the mean of $\sqrt{2\chi^2}$

distribution is $\sqrt{2\ \text{d. f.} - 1}$ and the standard deviation = 1. Accordingly, when the degree of freedom exceeds 30, the quantity $\left[\sqrt{2\chi^2} - \sqrt{2\text{d. f.} - 1}\right]$ maybe used as a normal variate with unit variance, i.e.

$$Z\alpha = \left[\sqrt{2\chi^2} - \sqrt{2\text{d. f.} - 1}\right]$$

Illustration 3: The table given below shows the data obtained during an epidemic of cholera:

	Attacked	Not Attacked	Total
Inoculated	31	469	500
Not Inoculated	185	1,315	1,500
Total	**216**	**1,784**	**2,000**

Test the effectiveness of inoculation in preventing the attack of cholera. The five percent value of χ^2 for one degree of freedom is 3.84.

Solution:

We will take the null hypothesis as follows:

Null Hypothesis: There is no association between inoculation and prevention of the attack of cholera, i.e. the two attributes are independent.

To calculate the value of χ^2, we must have actual and expected frequencies. Actual or observed frequencies are given in the question (attack of cholera is denoted by 'A' and inoculation is denoted by 'B'), which is as follows:

Observed Frequency

Attacked (A)	Not Attacked (α)	
31	469 (α B)	500

Inoculated (B)	(AB)		
Not Inoculated (β)	185 (Aβ)	1,315 (αβ)	1500
	216	1,784	2,000 (N)

On the basis of these actual frequencies, we can calculate the expected frequencies in the following manner.

(AB) = $\frac{500 \times 216}{2000}$ = 54 ∴ (αB) = (B) – (AB) or 500 – 54 = 446

(Aβ) = (A) – (AB) or 216 – 54 = 162 (αβ) = (β) – (Aβ) or 1500 – 162 = 1,338

It can also be shown in the following form:

Expected Frequency

Attacked (A)	Not Attacked (α)	
$\frac{500 \times 216}{2000}$ = 54 (AB)	500 – 54 = 446 (α B)	**500**
216 – 54 = 162 (Aβ)	1500 – 162 = 1,338 (αβ)	**1500**

Inoculated (B)	**216**	**1,784**	**2,000 (N)**

Not Inoculated (β)

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

Substituting the values:

$$\chi^2 = \frac{(31-54)^2}{54} + \frac{(469-446)^2}{446} + \frac{(185-162)^2}{162} + \frac{(1315-1338)^2}{1338}$$

$$= \frac{(-23)^2}{54} + \frac{(23)^2}{446} + \frac{(23)^2}{162} + \frac{(-23)^2}{1338}$$

$$= 9.8 + 1.19 + 3.27 + 0.40 = 14.66$$

Thus, the calculated value of χ^2 is **14.66**.

No. of degrees of freedom = (c – 1) (r – 1) = (2 – 1) (2 – 1) = 1

The tabular value of χ^2 at 5% level of significance with 1 d.f. = 3.84 whereas the calculated value of χ^2 is 14.66 which is much more than the tabular value. Hence, the null hypothesis is rejected which means inoculation and prevention from the attack of cholera are not independent. Thus, we can say that inoculation is effective in the prevention of cholera.

Alternative method of computing χ^2

We may use another method for the computation of the value of χ^2 in the case of a 2 × 2 contingency table. Let us assume that the observed frequencies are arranged in the following manner:

a	b	(a + b)

c	d	(c + d)
(a + c)	(b + d)	N

Then χ^2 can be computed by the following formula:

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

This formula can be applied to this question to compute the value of χ^2 directly without even finding out the expected frequencies. In such a situation, we need only actual frequencies which are already given in the question.

$$\chi^2 = \frac{[(31 \times 1315) - (185 \times 469)]^2 \times 2000}{(31+469)(185+1315)(31+185)(469+1315)}$$

$$= \frac{(40675 - 86765)^2 \times 2000}{(500)(1500)(216)(1784)}$$

$$= \frac{4232000000000}{28900800} = 14.6$$

Illustration 4: In a survey done on 1000 students who watch Discovery Channel and their IQ level, the following information is revealed.

	Watching Discovery	Not Watching Discovery	Total
High IQ	415	185	600
Low IQ	65	335	400
Total	480	520	1,000

Test at 5 % significance level if the students watching Discovery Channel have high IQ?

Solution:

Let us start with the null hypothesis that there is no relationship between watching Discovery Channel and IQ. Now, let's calculate the expected frequencies based on actual frequencies:

$\frac{480 \times 600}{1000}$ = 288	600 – 288 = 312	**600**
480 – 288 = 192	400 –192 = 208	**400**
480	**520**	**1000**

Observed Frequency (O)	Expected Frequency (E)	(O – E)	$(O – E)^2$	$(O – E)^2 / E$
415	288	127	16129	56
185	312	– 127	16129	51.7
65	192	– 127	16129	84
335	208	127	16129	77.54
			Total	269.24

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 269.24$$

No. of degrees of freedom = (c – 1) (r – 1) = (2 – 1) (2 – 1) = 1

The tabular value of χ^2 at the 5% level of significance with 1 d.f. is 3.841. Since the calculated value (269.24) is more than the tabular value (3.841), hence null hypothesis is rejected, and we conclude that children watching the Discovery Channel have high IQ.

Illustration 5: A die is thrown 150 times with the following results:

No. turned up:	1	2	3	4	5	6
Frequency:	19	23	28	17	32	31

Test the hypothesis that the die is unbiased.

Solution:

Let us take the null hypothesis that there is no significant difference between observed and expected frequencies in the throw of the die, i.e. die is unbiased. A die has 6 faces and the probability of turning up of each face is equal, hence, the expected frequencies for each face for 150 throws would be $\frac{150}{6} = 25$

O	E	(O – E)	$(O – E)^2$	$(O – E)^2/E$
19	25	– 6	36	1.44
23	25	– 2	4	0.16
28	25	3	9	0.36
17	25	– 8	64	2.56
32	25	7	49	1.96
31	25	6	36	1.44
			Total	7.92

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 7.92$$

No. of degrees of freedom = n – 1 = 6 – 1 = 5

The tabular value of χ^2 at 5% level of significance with 5 degree of freedom is 11.07. Since the calculated value of χ^2 (7.92) is less than its tabular value (11.07) hence, the null hypothesis is accepted. Thus, we can conclude that the die is unbiased.

Illustration 6: The following contingency table presents the analysis of eye colour and hair colour of 300 people. Use the χ^2 test to examine, is there any association between eye colour and hair colour?

Eye Colour	Hair Colour			Total
	Black	Fair	Brown	
Brown	30	10	40	80
Blue	40	20	40	100
Grey	50	30	40	120
Total	120	60	120	300

Solution:

Null Hypothesis: There is no association between eye colour and hair colour, i.e. both are independent.

Let's calculate the expected frequencies based on the actual frequencies given in the question.

Expected Frequency

Eye Colour	Hair Colour			Total
	Black	Fair	Brown	
Brown	$\frac{80 \times 120}{300}$ = 32	$\frac{80 \times 60}{300}$ = 16	80 – (32 +16) = 32	**80**
Blue	$\frac{100 \times 120}{300}$ = 40	$\frac{100 \times 60}{300}$ = 20	100 – (40 + 20) = 40	**100**

Grey	120 – (32 + 40) = 48	60 – (16 +20) = 24	120 – (32 + 40) = 48	**120**
Total	**120**	**60**	**120**	**300**

O	**E**	**(O – E)**	**$(O - E)^2$**	**$(O - E)^2/E$**
30	32	– 2	4	0.125
10	16	– 6	36	2.250
40	32	+ 8	64	2.000
40	40	0	0	0
20	20	0	0	0
40	40	0	0	0
50	48	+ 2	4	0.08
30	24	+ 6	36	1.50
40	48	– 8	64	1.33
			Total	7.285

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 7.285$$

No. of degrees of freedom = (c – 1) (r – 1) = (3 – 1) (3 – 1) = 4

The tabular value of χ^2 at 5% level of significance with 4 degrees of freedom is 9.488. The calculated value of χ^2 is 7.285 which is less than the tabular value. Hence, the null hypothesis is accepted. It means that there is no association between eye colour and hair colour.

Illustration 7: A survey of 320 families with 5 children each revealed the following distribution:

No. of boys: 5 4 3 2 1 0

No. of girls:	0	1	2	3	4	5
No. of families:	14	56	110	88	40	12

Is this result consistent with the hypothesis that male and female births are equally probable?

Solution:

Let us take the null hypothesis that male and female births are equally probable, i.e. there is no difference between the probability of male and female births.

In this question, the Binomial distribution can be applied to find out the expected frequencies because it fulfils all the conditions of the binomial distribution.

Probability of male birth = p = $\frac{1}{2}$

∴ Probability of female birth = q = $1 - \frac{1}{2} = \frac{1}{2}$

If a survey on 320 families with 5 children each is done, the binomial expansion will be as follows:

$320\ (p + q)^5$

or $320\left(\frac{1}{2} + \frac{1}{2}\right)^5$

$= 320\ [p^5 + 5p^4q + 10p^3q^2 + 10p^2q^3 + 5pq^4 + q^5]$

$=320[\left(\frac{1}{2}\right)^5 + 5\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right) + 10\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 + 10\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^3 + 5\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^5]$

$= 320\left(\frac{1}{32}\right) + 320\ (5)\left(\frac{1}{16}\right)\left(\frac{1}{2}\right) + 320\ (10)\ \left(\frac{1}{8}\right)\left(\frac{1}{4}\right) + 320\ (10)\left(\frac{1}{4}\right)\left(\frac{1}{8}\right) + 320\ (5)\left(\frac{1}{2}\right)\left(\frac{1}{16}\right) + 320\left(\frac{1}{32}\right)$

$= 10 + 50 + 100 + 100 + 50 + 10$

The various terms of the above binomial expansion show the required expected frequencies.

O	E	(O – E)	$(O - E)^2$	$(O - E)^2/E$

14	10	4	16	1.60
56	50	6	36	0.72
110	100	10	100	1.00
88	100	– 12	144	1.44
40	50	– 10	100	2.00
12	10	2	4	0.40
			Total	7.16

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 7.16$$

No. of degrees of freedom = n – 1 = 6 –1 = 5

The tabular value of χ^2 at a 5% level of significance with 5 degrees of freedom is 11.07. The calculated value of χ^2 is 7.16, which is less than the tabular value. Hence, the null hypothesis is accepted, and it can be concluded that male and female births are equally probable.

Illustration 8: A sample analysis of the examination results of 200 MBA was made. It was found that 46 students failed, 68 secured a third division, 62 secured a second division, and the rest were placed in the first division. Are these figures commensurate with the general examination result, which is in the ratio of 2:3:3:2, for various categories respectively?

Solution:

Let us take the null hypothesis that there is no difference in the observed and expected results.

The expected frequencies can be found based on the general examination results' ratio of 2:3:3:2. based on this ratio, the expected number of students failing, getting third division, second division, and first division, should be $\frac{200 \times 2}{10}$ = 40, $\frac{200 \times 3}{10}$ = 60, $\frac{200 \times 3}{10}$ = 60 and $\frac{200 \times 2}{10}$ = 40 respectively.

O	E	(O – E)	$(O – E)^2$	$(O – E)^2/E$
46	40	+ 6	36	0.900

68	60	+ 8	64	1.067
62	60	+ 2	4	0.067
24	40	– 16	256	6.400
			Total	8.434

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 8.434$$

No. of degrees of freedom = n – 1 = 4 –1 = 3

The tabular value of χ^2 at a 5% level of significance with 3 degrees of freedom is 7.81. The calculated value of χ^2 is 8.434 which is more than the tabular value. Hence, the null hypothesis is rejected and we conclude that the given results are not commensurate with the general examination results.

15.7 YATE'S CORRECTION

In 1934, F. Yate has suggested a correction for continuity in χ^2 value calculated in connection with a (2 × 2) table, particularly when cell frequencies are small and χ^2 is just on the significance level. When using χ^2 analysis, a minimum of 80 percent of the expected or theoretical frequencies in a cell must be at least five and no cell have an expected frequency of less than one. When any expected frequency is less than 5, then we use Yate's correction in a 2 × 2 contingency table consisting of 1 d.f. According to this correction, the observed frequency, which is less than 5, is increased by 0.5, and the other frequencies are so adjusted (by adding 0.5 and subtracting 0.5) that the row total and column total remain the same.

Grouping of Frequencies

If small theoretical frequencies occur, it is generally possible to overcome this difficulty by grouping two or more classes together. In other words, one or more classes with theoretical frequencies less than 5 may be combined into a single category before calculating the difference between observed and expected frequencies. The number of degrees of freedom would be determined by the number of classes after the regrouping. Yate's correction is used in (2 × 2)

contingency table. Grouping of frequencies is used in m × n (m > 2, n > 2) contingency tables where the expected frequencies less than 5 are added to adjacent frequency.

Illustration 9: The following information was obtained in a sample of 50 small general shops. Can it be said that there are relatively more women owners in villages than in town? Use the χ^2 test.

	Shops		
	In Towns	In Villages	Total
Run by men	17	18,	**35**
Run by women	3	12	**15**
Total	**20**	**30**	**50**

The value of χ^2 at 5% level of significance with 1 d.f. is 3.841.

Solution:

Null Hypothesis: The number of women owners is the same in the town and village.

Adding 0.5 to the cell frequency 3 and adjusting the other cell frequencies so that the row totals remain same we have the following contingency table:

Observed Frequencies Table

A	α	Total
16.5	18.5	**35**
3.5	11.5	**15**

B	**20**	**30**	**50**
			β

Expected Frequencies Table

	A	α	Total
B	$\frac{35 \times 20}{50}$ = 14	$\frac{35 \times 30}{50}$ = 21	**35**
β	$\frac{15 \times 20}{50}$ = 6	$\frac{15 \times 30}{50}$ = 9	**15**
	20	**30**	**50**

O	E	(O – E)	$(O - E)^2$	$(O - E)^2/E$
16.5	14	+ 2.5	6.25	0.45
18.5	21	– 2.5	6.25	0.30
3.5	6	– 2.5	6.25	1.04
11.5	9	+ 2.5	6.25	0.69
			Total	2.48

The tabular value of χ^2 at 5% level of significance with 1 d.f. is 3.841. The calculated value of χ^2 is 2.48 which is less than its tabular value. Hence, the null hypothesis is accepted. This means that the number of women owners is the same in the town and village.

15.8 CRITICAL OVERVIEW OF CHI-SQUARE TEST

You have seen in the above sections that the chi-square test is used in many areas for making comparison between sample variance and population variance, for finding out the independence or association between two attributes, or for finding out the homogeneity between different samples. It is also used to judge the goodness of fit regarding theoretical distribution. Thus, there is no doubt about the importance of χ^2 test.

The χ^2 test is very popular and frequently used because of its **merits**. The chief merit of the χ^2 test is that it does not assume the form of parent distribution or its parameters. Since the χ^2 test is based on observed and expected frequencies and not on parameters like mean and standard deviation, hence it does not need to make any assumption regarding parent distribution. Being a distribution-free test, it can be used in any type of population distribution which enhances the scope of the χ^2 test. Another reason for the popularity of this test is that in comparison to parametric tests, like t-test, Z-test, and F-test, the procedure of calculation and interpretation is easy in χ^2 test. Yet another advantage of the χ^2 test is its additive property which it possible to add the results of independent related samples. Due to all these merits, the chi-square test is frequently used in the area of business problems and social sciences.

Despite these advantages, the chi-square test has certain **limitations or demerits** also. As you know, certain conditions have to be fulfilled for the application of the χ^2 test. Fulfillment of these conditions is the biggest limitation of this test. Another notable thing in this regard is that the χ^2 test is not as reliable as parametric tests are. Thus, if in a situation χ^2 test and any parametric test both can be applied then preference should be given to the parametric test. Similarly, this test can be used only for testing of hypothesis. It is not useful for estimation. Another limitation of this test is that for using the chi-square test, it is necessary to have data regarding occurrence as well as non-occurrence of events. Yet another demerit of the chi-square test is that the χ^2 value cannot be calculated where repeated measurements on the same or matched groups are represented in one table. But all these limitations do not reduce the importance or popularity of the χ^2 test.

There is no doubt about its importance or popularity but its correct application is also an important and difficult task. One must be very **cautious** while using the χ^2 test. It should always be remembered that the test is to be applied only when the individual observations of the sample are independent. It means that the occurrence of one individual item event or observation does not affect the occurrence of any observation or event in the sample under consideration. A sample having small theoretical frequencies should be dealt with special care. The other possible reasons concerning the improper application or misuse of this test can be negligence regarding frequencies of non-occurrence, failure to equalize the sum of observed and the sum of expected frequencies, wrong determination of the degrees of freedom and wrong computation, etc. thus, the researcher should take into account all these precautions while using the χ^2 test and drawing inferences in respect of his hypothesis.

15.9 SUMMARY

The chi-square test was propounded by Karl Pearson in 1900. χ^2 test is a parametric as well as a non-parametric test. But it is mostly used as a non-parametric test. χ^2 test helps us to determine the extent of difference between the theory or expected value and the observed or the actual value. There are some conditions that must be fulfilled to apply the χ^2 test like a total number of items should be at least 50 and the frequency in any cell should not be less than five. The sample should be selected at random and data should be shown in absolute form. χ^2 test can be used to test if a random sample has been drawn from a normal population with mean μ and variance. σ_p^2. As a non-parametric test, χ^2 test can be used as a test of independence, goodness of fit, and homogeneity. χ^2 as a test of independence can establish if two or more attributes are associated or independent. χ^2 as a test of goodness of fit is used to determine how well a theoretical distribution fits on observed data. χ^2 as a test of homogeneity is used to find out if two or more randomly selected independent samples have been drawn from the same population or not. The formula $\sum \frac{(O-E)^2}{E}$ is used to calculate the value of χ^2. The calculated value of χ^2 is compared with the tabular value of χ^2 at certain degrees of freedom and level of significance. If the calculated value is less than the tabular value then the null hypothesis is accepted otherwise rejected. The degree of freedom refers to the number of classes to which a value can be assigned freely without exceeding the limitation placed. (n– 1) or (c –1)(r– 1) is used to ascertain the number of degrees of freedom. If the frequency in

any cell is less than 5, then Yate's correction is applied, whereby 0.5 is added to that frequency, and the other frequencies are adjusted so that the row and column totals remain the same.

15.10 GLOSSARY

Degree of freedom: No. of independent constraints in a set of data

Level of significance: Maximum probable percentage of the numbers being wrong due to fluctuations of random sampling.

Goodness of fit: Properness of theoretical distribution to the observed distribution.

15.11 CHECK YOUR PROGRESS

A. Fill in the blanks:

1. When observed and expected frequencies completely coincide, then the value of χ^2 will be……
2. Prof. ……propounded the χ^2 test.
3. The quantity χ^2 describes the magnitude of discrepancy between theory and …….
4. χ^2 distribution is a ……. probability distribution.
5. Yate's correction is applicable in a ……contingency table.

B. State whether each of the following statements is true or false: `

1. The formula for the calculation of χ^2 is $\sum \frac{(O-E)^2}{O}$. ( )
2. Calculated value of χ^2 is positive or negative. ( )
3. Degree of freedom is denoted by v. ( )
4. (c – 2) (r – 2) formula is used for determining the degree of freedom in a contingency table ( )
5. If the calculated value of χ^2 is less than its tabular value, then the null hypothesis is rejected. ( )

15.12 ANSWERS TO CHECK YOUR PROGRESS

A. (i) zero (ii) Karl Pearson (iii) observation (iv) continuous (v) (2 × 2)

B. (i) False (ii) False (iii) True (iv) False (v) False

15.13 TERMINAL QUESTIONS

1. Critically analyze the use of chi-square testing.

2. What is meant by Degrees of Freedom?

3. In which situation is Yate's correction applied?

4. What is the chi-square test? What are the steps in the application of the chi-square test?

5. Explain the critical overview of the applications of the chi-square test.

6. A sample of scores of seven students of a class is given as follows:

S. No.:	1	2	3	4	5	6	7
Scores (in percentage):	52	50	56	61	45	54	39

Use χ^2 test to determine if the above sample has been drawn from a student population whose variance is 25. Test at 5% significance level. The tabular value of χ^2 with 6 degrees of freedom is 14.1. [12.64, H_0 accepted]

7. A sample of 10 is randomly drawn from a certain population. The sum of the squared deviations from the mean of the given sample is 50. Test the hypothesis that the variance of the population is 5 at 5 percent level of significance. The tabular value of χ^2 with 9 degrees of freedom is 16.92. [10, H_0 accepted]

8. The data given below show the data obtained during an epidemic of malaria:

	Attacked	**Not Attacked**	**Total**
Inoculated	120	240	**360**
Not Inoculated	280	360	**640**

Total	**400**	**600**	**1,000**

Test the effectiveness of inoculation in preventing the attack of malaria. The tabular value of χ^2 with 1 degree of freedom is 3.841. [10.41, H_0 rejected]

9. The police records in a certain city show the following data relating to the number of accidents that occurred during the first week of January 2012. You are required to find whether the accidents are uniformly distributed over the week.

Day:	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Total
No. of accidents:	20	12	13	17	19	20	18	119

χ^2 value for 6 degrees of freedom at 5% level of significance is 12.59. [3.77, H_0 accepted]

10.200 digits were chosen at random from a set of tables. The frequencies of the digits were:

Digits:	0	1	2	3	4	5	6	7	8	9
Frequency:	18	19	23	21	16	25	22	20	21	15

Use χ^2 test to assess the correctness of the hypothesis that the digits were distributed in equal numbers in the tables from which these were chosen. Value of χ^2 for 9 d.f. at 5% level of significance is 16.919. [4.3,H_0 accepted]

11. Two research studies were conducted to classify people into income groups based on sampling studies. Their result was as follows:

Research Studies	Income Groups			**Total**
	Poor	Middle	Rich	
A	160	30	10	**200**
B	140	120	40	**300**
Total	**300**	**150**	**50**	**500**

Value of χ^2 for 9 d.f. at 5% level of significance is 16.919. [55.54, H_0 rejected]

12. Five coins are tossed 3,200 times and the number of heads appearing each time is noted. In the end, the following results were obtained:

No. of heads:	0	1	2	3	4	5
Frequency:	80	570	1100	900	500	50

Use the chi-square test of goodness of fit to determine whether the coins are unbiased. Value of χ^2 for 5 d.f. at 5% level of significance is 11.07. [58.8, H_0 rejected]

13. In an experiment on the immunization of goats from anthrax the following results were obtained:

	Died	Survived	Total
Inoculated	2	10	12
Non inoculated	6	6	12
Total	8	16	24

Calculate χ^2 by Yate's correction and derive your inference on the efficacy of the vaccine. [1.6876, H_0 accepted]

15.14 SUGGESTED READINGS

1. Roy Ramendu, 'Principles of Statistics' Prayag Pustak Bhawan, Allahabad
2. Gupta S. P. & Gupta M. P., 'Business Statistics' Sultan Chand & Sons, New Delhi
3. Shukla S. M. & Sahai S. P., 'Advanced Statistics' Sahitya Bhawan Publications, Agra

UNIT 16: SIGN TEST AND MEDIAN TEST

Structure

OBJECTIVES

After studying this unit, you shall be able to understand:

- Different types of non-parametric tests
- Applicability of non-parametric tests
- Advantages & limitations of non-parametric tests

16.1 INTRODUCTION

In the previous unit, you have studied about chi-square test, which is the most popular non-parametric test. Apart from the chi-square test, a number of other useful non-parametric tests have been developed during recent years. As you know that non-parametric tests are distribution-free tests which do not assume that a particular distribution is applicable or that a certain value is

attached to a parameter of the population; thus, they are easy to apply. This is the reason that even where these tests have corresponding parametric testing methods, the non-parametric versions are preferred for being more realistic in not requiring the hypothesized population to be normal or near normal.

Among a variety of non-parametric tests that have become available to us, we shall dwell upon only those that are more frequently used in a number of real-life situations. In this unit, you will study such frequently used and popular non-parametric tests like Sign Test, the Median Test, Fisher Irwin Test, the Mann-Whitney U Test, the McNemer Test, the Wilcoxon Test, One sample runs test, etc.

16.2 SIGN TEST

The sign test is one of the earliest and simplest non-parametric tests. Its name comes from the fact that it is based on the direction (plus or minus signs) of a pair of observations and not their numerical magnitude. Thus, in this test, the magnitude of differences between the predicted values and observed values is not important; rather, only the direction of difference, i.e., + or – sign, is relevant. It is useful for evaluating the effectiveness of two types of methods whose effects cannot be quantified but can only be judged as superior/inferior or good/bad, or preferable / not preferable. For example, a group of students are asked to evaluate two different types of teaching methods. The evaluation of the two methods is converted into signs; a plus means preference for the first method, a minus means preference for the second method, and a zero represents a tie, i.e., no preference. We count (+) signs and (–) signs and exclude tie evaluations. The only requirement necessary for using this test is that the population distribution is approximately symmetric about the mean μ_0. The sign test is of two types:

- One-sample sign test
- Two-sample sign test

16.2.1 One-Sample Sign Test

The one-sample sign test is a very simple non-parametric test applicable when we sample a continuous symmetrical population, in which case the probability of getting a sample value less

than the mean is ½, and the probability of getting a sample value greater than the mean is also ½. The procedure of conducting a sign test is as follows:

- First of all, a random sample of size n is selected from the population, and we take the null hypothesis that the population mean is equal to the hypothesized mean, i.e., H_0: $\mu = \mu_0$.
- Each of the n sample values is observed to find out whether it is greater than μ_0 or less than it. Sample values greater than μ_0 are assigned a plus (+) sign and referred to as a success, and those less than μ_0 are assigned a minus (–) sign and referred to as a failure.
- If there is any item whose value is the same as that of the mean, then zero (o) is assigned to it, and these values are simply discarded. It reduces the sample size.
- The total number of signs is denoted by 'n', and the number of less frequent signs is denoted by S.
- Thereafter, the critical value for a two-sided alternative at 5% significance level is calculated and is denoted by 'K'. The formula for calculating the value of 'K' is as follows:

$$K = \frac{n-1}{2} - (0.98)\sqrt{n}$$

- In the last step, a comparison is made between the value of S and K. If the value of S is greater than the value of K, then the null hypothesis is accepted. If however, the value of S is less than or equal to K, then the null hypothesis is rejected.

For performing a one-sample sign test when the sample is small, we can use tables of binomial probabilities. Using the Z critical value of the normal distribution, the null hypothesis is tested. The following formula is used:

$$Z = \frac{(p \pm 0.50) - n/2}{\sqrt{n/4}}$$

Where: p = the number of plus signs

n = the total of plus and minus signs (zero signs excluded)

When $p < n/2$, (+ 0.5) is used and when $p > n/2$, (– 0.5) is used in the formula.

If the calculated Z value is higher than the table value, the null hypothesis is rejected; otherwise, null hypothesis is accepted.

When the sample happens to be large, we use the normal approximation to the binomial distribution. The following formula is used for the calculation of the Z value:

$$Z = \frac{S - np}{\sqrt{np(1-p)}} \quad \text{or} \quad Z = \frac{S - np}{\sqrt{n.p.q}}$$

Where: p = proportion of successes & q = 1–p S = No. of positive signs

If the calculated 'Z' value is less than its critical value, then the null hypothesis is accepted, and if the calculated 'Z' value is higher than the tabular value, then it is rejected.

An important point to be noted here is that if it is expected that the population is not a continuous symmetrical binomial population, then the null hypothesis can be expressed in terms of median instead of mean.

Illustration 1: A salesman made 12 visits to his area sales manager and noted that he had to wait for 10,15, 20, 17, 11, 25, 30, 27, 36, 40, 5, and 26 minutes respectively, before being called into his office. The area sales manager claims that the salesmen wishing to meet him do not have to wait for more than 20 minutes before being called in. Using the sign test, verify at a 0.05 level of significance the claim made by the area sales manager.

Solution:

H_0: $\mu = 20$ minutes

H_1: $\mu > 20$ minutes

Now we will provide plus and minus signs to sample values on the basis of whether it is greater than 20 or less than 20.

Time (in minutes):	10	15	20	17	11	25	30	27	36	40	5	26
Sign:	–	–	0	–	–	+	+	+	+	+	–	+

Total no. of signs or n = 11

Total no. of less frequent signs, i.e. (–) or S = 5

$$K = \frac{n-1}{2} - (0.98)\sqrt{n}$$

$$= \frac{11-1}{2} - (0.98)\sqrt{11}$$

$$= 5 - 3.25 = 1.75$$

Since the value of S (5) is greater than the value of K (1.75), the null hypothesis is accepted. It means the claim made by the area sales manager is valid.

Illustration 2: Suppose playing four rounds of golf at the City Club, 11 professionals totaled 280, 282, 290, 273, 283, 283, 275, 284, 282, 279, and 281. Use the sign test at a 5% level of significance to test the null hypothesis that professional golfers' average $\mu_0 = 284$ for four rounds against the alternative hypothesis $\mu_0 < 284$.

Solution:

H_0: $\mu_0 = 284$

H_1: $\mu_0 < 284$

Now we will provide plus and minus signs to sample values on the basis of whether it is greater than 284 or less than 284.

Score:	280	282	290	273	283	283	275	284	282	279	281
Sign:	–	–	+	–	–	–	–	0	–	–	–

Total no. of signs excluding zero, i.e. ‘n’ = 10

Total no. of less frequent sign, (+) i.e. ‘S’ = 1

This question can be solved by different alternative methods which are illustrated below:

(i) Under the Experimental Method

$$K = \frac{n-1}{2} - (0.98)\sqrt{n}$$

$$= \frac{10-1}{2} - (0.98)\sqrt{10}$$

$$= 4.5 - 3.099 = 1.4$$

Since the value of S (1) is less than the value of K (1.4), the null hypothesis is rejected. It means that the golfers' average is less than 284 for four rounds of golf.

(ii) Under the Binomial Probability Method

When the sample size is small, we can use the binomial probability distribution to perform a sign test.

n = 10, p = ½ q = ½

No. of less frequent sign (+) is 1. Hence, the probability of one or fewer successes with n=10 and p = ½ can be worked out as under:

$$P(1) = {}^{10}C_1\, p^1 q^9 + {}^{10}C_0\, p^0 q^{10}$$

$$= 10\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^9 + 1\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^{10}$$

$$= 0.010 + 0.001 = 0.011$$

Since the value of P(S), i.e., 0.011, is less than 0.05 (i.e., desired significance level) hence, the null hypothesis must be rejected. It means that the golfers' average is less than 284 for four rounds of golf.

You should remember that the null hypothesis is accepted if P(S) > α, and the null hypothesis is rejected if P(S) < α.

(iii) Under the Normal Curve Method

n = 10, p = ½ q = ½

Based on signs, the observed proportion of success = $\hat{p} = \frac{1}{10} = 0.1$

The standard error of proportion assuming the null hypothesis p = ½ is as under:

$$\text{S.E.}_{\text{prop}} = \sqrt{\frac{p.q}{n}} = \sqrt{\frac{\frac{1}{2}\times\frac{1}{2}}{10}} = 0.1581$$

For testing the null hypothesis, i.e., p =1/2, against the alternative hypothesis p < 1/2, a one-tailed test (left-tailed) is appropriate.

Since the significance level is 5%, hence, the acceptance region is 0.5 – 0.05 = 0.45 area. By using the table of area under a normal curve, we find that the corresponding z value for 0.45 of the area is 1.64. Now, we will find out the limit of the acceptance region by deducting the standard error of proportion from p (because it is a left-tailed test, thus, $\text{S.E.}_{\text{prop}}$ should be deducted, not added).

Limit of acceptance region = $P – z.\ \text{S.E.}_{\text{prop}}$

$$= \frac{1}{2} – (1.64)\ (0.1581)$$

$$= 0.5 – 0.2593 = 0.2407$$

As the observed proportion of success is only 0.1, which comes in the rejection region (because it is under the area of 0.2407), hence the null hypothesis is rejected, and consequently, the alternative hypothesis is accepted, which means that professional golfers' average is less than 284 for four rounds of golf.

16.2.2 Two-Sample Sign Test

The sign test has important applications in problems where we deal with paired data. It can be applied to n paired observations belonging to two symmetric populations. Thus, it is also known as a sign test for paired data. The test statistic and the decision rule are the same as those for a sample sign test. The difference lies only in the manner in how plus and minus signs are assigned. For each subject or item, 1st score is compared with 2nd score. If the difference is positive, i.e., 1st score is greater than 2nd score, we assign '+' sign; if the difference is negative, i.e., 1st score is lower than 2nd score, then '– ' sign is assigned. For subjects, where the two values or scores are equal, '0' is assigned to them, and these pairs are discarded. In case the two samples are not of equal size, then some values of the larger sample that do not have a pair are discarded. Two sample

sign test is mainly used for two repeated measures to test if there is any significant difference in the mean of two samples.

Illustration 3: The pulse rate of 22 patients measured before and after administering a drug is as follows:

Patient	Pulse rate before taking drug	Pulse rate after taking drug	Patient	Pulse rate before taking drug	Pulse rate after taking drug
1	73	75	12	70	72
2	71	73	13	70	69
3	69	70	14	67	70
4	68	69	15	74	75
5	74	73	16	72	74
6	72	73	17	71	71
7	73	73	18	73	75
8	71	72	19	71	69
9	70	68	20	70	72
10	69	74	21	73	75
11	73	70	22	74	75

Use the sign test at a 5% level of significance to test the hypothesis that “the drug does not affect pulse rate”.

Solution:

Patient	Pulse rate before taking drug	Pulse rate after taking drug	Sign
1	73	75	–
2	71	73	–
3	69	70	–
4	68	69	–
5	74	73	+
6	72	73	–
7	73	73	0
8	71	72	–
9	70	68	+
10	69	74	–
11	73	70	+
12	70	72	–
13	70	69	+
14	67	70	–
15	74	75	–
16	72	74	–
17	71	71	0

18	73	75	–
19	71	69	+
20	70	72	–
21	73	75	–
22	74	75	–

Total no. of plus sign = 5, Total no. of minus sign = 15 ∴ Sample size or n = 20

We will discard 2 pairs which are assigned with '0' sign, hence the sample size would become 22 – 2 = 20

We will solve this question by using normal approximation to binomial distribution.

$$Z = \frac{S - np}{\sqrt{np(1-p)}}$$

$$= \frac{5 - \left(20 \times \frac{1}{2}\right)}{\sqrt{20 \times \frac{1}{2}\left(1 - \frac{1}{2}\right)}}$$

$$= \frac{5 - 10}{\sqrt{20 \times \frac{1}{2} \times \frac{1}{2}}} = \frac{-5}{\sqrt{5}}$$

$$= -2.23$$

Since the calculated value of Z (–2.23) is less than tabular value of Z (–1.64) at 5% level of significance, hence, the null hypothesis is rejected. It means that the drug has an effect on the pulse rate.

16.3 THE MEDIAN TEST

In the previous unit, you have studied about chi-square test, which applies to two or more independent samples measured at the nominal level. But if the samples are measured at an ordinal level, then we can apply the median test. Similarly, in the previous section, you have studied about sign test, which is mainly used in those cases where we deal with n sets of paired observations. But one condition necessary for the application of sign test is that two samples of the same size should be drawn, provided the resultant sample data are paired outcomes. In practical situations, we have to deal with such problems that require the selection of two independent samples, not necessarily of the same size, from different populations. In such situations, it becomes important to verify whether the samples are drawn from populations that differ in their central values. By comparing the medians, the median test is applied to determine whether the samples have been drawn from populations with the same median. It determines the significance of the difference between the medians of two or more random samples. The procedure of conducting the median test is as follows:

- In the first step, the median of the combined samples is computed. In this step, after selecting the samples, observations contained in each are combined, arranged in order of magnitude, and the median is found. Median is the middle observation where n is an odd number and it is the mean of the two middle observations where n is an even number.
- In the next step, all the observations in the first sample (n_1) are compared with the median value and classified into two groups: (i) above the median (a_1) and (ii) below the median (b_1). Similarly, after comparing with the median, all the observations in the second sample (n_2) are classified into two categories—above the median (a_2) and below the median (b_2).
- Thereafter, the resultant data are presented in the form of a 2×2 contingency table.

Classification of Sample Data for the Median Test

	Above M_d	Below M_d	Sample Size
Sample I	a_1	b_1	$n_1 = a_1 + b_1$
Sample II	a_2	b_2	$n_2 = a_2 + b_2$
Total	$a_1 + a_2$	$b_1 + b_2$	$n = n_1 + n_2$

- While classifying the sample data in the 2×2 contingency table, if any observation is found to be equal to the value of the median, then it can either be deleted from the sample or it may be included in the above median class. If the size of the sample is sufficiently large, then the observations equal to the median value can be deleted from the sample, but if the size of the sample is small, then it should be included in the above median class.
- If the sample size is large (at least 30 or more than 30) then we will apply the chi-square test for testing of null hypothesis. In that case, after finding out the expected frequencies on the basis of 2 × 2 contingency table, we will compute the χ^2 value and compare it with its tabular value at a pre-determined level of significance with 1 d.f. If the computed value is less than the tabular value, then the null hypothesis is accepted, and if the computed value is more than the tabular value, then the null hypothesis is rejected.
- If the sample size is small, then we will apply Fisher's exact probability test for testing of null hypothesis. In that case, we will find out Fisher's exact probability for the set of frequencies arranged in the form of a 2×2 contingency table by using the hypergeometric distribution in the following manner:

$$P_{(a_1\, a_2)} = \frac{\left(n_1 C_{a_1}\right)\left(n_2 C_{a_2}\right)}{\left(n_1 + n_2\ C_{\,a_1 + a_2}\right)}$$

The decision to accept or reject the null hypothesis is taken by comparing the computed value of P with the level of significance (α). If the computed value of P is less than the level of significance, then the null hypothesis is rejected, and if the value of P is more than α then the null hypothesis is accepted.

Illustration 4: Two large shipments marked S_1 and S_2 of TV tuners are received by an importer. He selects two samples consisting of 25 tuners each from the two shipments. On getting them checked for the number of defective pieces, he is provided with the following data on the number of defective tuners:

Sample I (S_1):	2	1	0	4	2	3	6	5	3	1
Sample II (S_2):	3	2	0	6	3	4	8	6	5	2

Verify at a 0.05 level of significance by using the median test if the number of defective items in the two shipments is the same as the median number.

Solution:

Null Hypothesis (H_0): The number of defective items in the two shipments is the same as the median number.

Alternative Hypothesis (H_1): The number of defective items in the two shipments is not the same as the median number.

Now, we will arrange the observations of the 1st and 2nd sample in ascending order to find out the value of the median.

S.No:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Values:	0	0	1	1	2	2	2	2	3	3	3	3	4	4	5	5	6	6	6	8

Median = the value of $\frac{n+1}{2}$ the item

= the value of $\frac{20+1}{2}$ th item = the value of 10.5^{th} item

= average of 10^{th} and 11^{th} item

= $\frac{3+3}{2} = 3$

Now, we will compare the values of the 1st and 2nd sample with the median value (3) and grouped them into above median and below median classes. It can be expressed in the form of the following 2×2 contingency table:

	Above M_d	Below M_d	Sample Size
S_1	5	5	10
S_2	7	3	10
Total	12	8	20

Since, the size of sample (20) is small (less than 30) thus we will apply Fisher's exact probability test.

$$P_{(a_1\, a_2)} = \frac{\left(n_1 C_{a_1}\right)\left(n_2 C_{a_2}\right)}{\left(n_1 + n_2\ C\ a_1 + a_2\right)}$$

$$P_{(5,7)} = \frac{(10_{C_5})(10_{C_7})}{(20\ C_{12})} = \frac{252 \times 120}{125970} = 0.24$$

Since $P_{(5,7)}$ is 0.24 which is more than the level of significance (0.05), hence, null hypothesis is accepted. It means the two shipments have the same number of defective pieces as the median number.

16.4 WILCOXON MATCHED PAIRS TEST

This test is appropriate for matched pairs data, i.e., for testing the significance of the relationship between a dichotomous variable and a continuous variable with related samples. This test is also known as the Wilcoxon signed rank test because in sign test, we deal only with the sign of the difference between values, whereas this test not only tests the direction, but also the magnitude of

differences between matched pairs. This test is particularly suitable for matched pairs of before and after experiment type. It is mainly used in case of matched pairs, such as a study where husband and wife are matched, or when we compare the output of two similar machines or the results of a before-after experiment. Thus, it is an important non-parametric test that takes into consideration the direction and magnitude of differences between matched pairs. The procedure of conducting the Wilcoxon signed rank test is as follows:

- First of all, depending upon the nature of the problem, a null hypothesis of no difference between the two series under consideration is taken.
- Differences between each pair of scores or values are worked out.
- Ranks are assigned to the differences from the smallest to the largest without regard to sign. It means rank 1 is given to the smallest difference, rank 2 to the next in order, and so on. While assigning the ranks, a sign of difference is not taken into consideration.
- While using this test, we may come across two types of tie situations:
    - One situation arises when the two values of some matched pairs are equal, i.e. the difference between values is zero. Such pairs are dropped from the calculation. Thus, pairs carrying equal scores or a zero difference is excluded from further calculations.
    - Another tie situation arises when two or more pairs have the same difference value. In this case, we work out the average rank of the concerned positions and assign them this average rank. For example, suppose after assigning rank 1, 2 and 3, we have two same different values. If their difference had not have been equal, then they would have been awarded with 4^{th} and 5^{th} rank. Hence, we will assign both of them the average of 4^{th} and 5^{th} rank, i.e., [(4+5)/2] = 4.5. The next difference value will be assigned the 6^{th} rank.
- Once the differences have been ranked, each is then assigned the sign of the actual difference.
- In the next step, the test statistic (T) is calculated which happens to be the smaller of the two sums, viz. the sum of negative ranks and sum of positive ranks.
- If the total number of matched pairs, after considering the dropped pairs, is equal to or less than 25, then the table of critical values of T is used for accepting or rejecting the null hypothesis. The null hypothesis is accepted if the calculated value of the T statistic is larger than the table value and if T is equal to or less than the table value, the null hypothesis is rejected.

- When the number of matched pairs is greater than 25 then z statistic is worked out to decide on acceptance or rejection of the null hypothesis. The value of z is calculated in the following manner:

$$Z = \frac{T - U_T}{\sigma_T}$$

Where: U_T = mean σ_T = Standard Deviation T = Sum of ranks of the smaller sign group

Mean or $U_T = \frac{n(n+1)}{4}$ Standard Deviation or $\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$

Where: n = number of matched pairs, excluding dropped pairs

Illustration 5: Students in the 5th standard were given two sets of practice workbooks, A and B, to practice mathematics. The data given below shows the marks obtained by students practicing from workbooks A and B, respectively. The researcher is interested in knowing if there is a perceptible difference in the scores that can be attributed to the type of workbook children use.

Student No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A:	73	43	47	53	58	47	52	58	38	61	56	56	34	55	65	75
B:	51	41	43	41	47	32	24	58	43	53	52	57	44	57	40	68

Solution:

The null and alternative hypothesis for this problem can be taken as follows:

H_0 : There is no difference between the scores of the two groups of students.

H_a: There is a difference between the scores of the two groups.

Using the Wilcoxon matched-pairs test, we work out the value of the test statistic T as under:

Pair	**Work-book A**	**Work-book B**	**Difference (di)**	**Rank of Diff. |di|**	**Signed Ranks** +	–
1	73	51	+22	13	+13	…..
2	43	41	+2	2.5	+2.5..…	
3	47	43	+4	4.5	+4.5	…..
4	53	41	+12	11	+11	…..
5	58	47	+11	10	+10	….
6	47	32	+15	12	+12	….
7	52	24	+28	15	+15	….
8	58	58	0	….	…..	…..
9	38	43	– 5	6	….	– 6
10	61	53	+8	8	+8	….
11	56	52	+4	4.5	+4.5	….
12	56	57	–1	1	…...	–1
13	34	44	–10	9	….	–9
14	55	57	–2	2.5	…...	–2.5
15	65	40	+25	14	+14	….
16	75	68	+7	7	+7	….
				Total	**+101.5**	**–18.5**

We will discard the pair no.8 because its difference value is zero.

∴ n = 16 – 1 = 15

& T = Sum of smaller sign group = 18.5

The table value of T at 5% level of significance when n = 15 is 25 (using a two-tailed test because our alternative hypothesis is that there is difference between the two groups). The calculated value of T is 18.5 which is less than the table value of 25. As such, we reject the null hypothesis and conclude that there is difference between the two groups.

16.5 WILCOXON-MANN-WHITNEY TEST (U TEST)

The U test is a very popular test amongst the rank sum tests. It is also popularly known as the Wilcoxon-Mann-Whitney test. Like the earlier tests, the Mann-Whitney U test is also based on data derived from two independent samples. It is used to determine whether two independent samples have been drawn from the same population or two different populations having the same distribution. It is considered as an improvement over the sign test and the Fisher-Irwin test because it uses ranking information rather than plus and minus sign. Mann-Whitney U and Wilcoxon Matched pairs are the same in that they compare between two medians to suggest whether both samples come from the same population or not. If both of your samples are not entirely independent from each other and have some factor in common, i.e., geographical location or before/after treatment, the Wilcoxon Matched Pairs test can be applied. If you have two independent samples, you should use the Mann-Whitney U test. As an extremely versatile non-parametric test, it is used in situations where samples are of small and unequal size. This test applies under very general conditions and requires only that the populations sampled are continuous. The procedure of conducting a U test is as follows:

- Two independent samples normally of different sizes are drawn from a single population or two different populations. A sample of the smaller size is taken as consisting of n_1 observations, and that of the larger size as comprising n_2 observations.
- Null and alternate hypothesis are taken. The null hypothesis states that the two sets of score do not have differences whereas the alternative hypothesis states that the two sets of scores do differ systematically. It may be one-tailed or two-tailed.

- The two samples are combined, and all the n = (n_1 + n_2) observations are arranged in ascending order, starting from the lowest and finally moving to the highest. Thereafter, ranks are assigned. The values of the combined samples n_1 and n_2 are ranked from the lowest to the highest rank, irrespective of the samples, rank 1 to the smallest score, rank 2 to the next in order, and so on, indicating beside each the identification of the sample. Repeated values are ranked with the average of their initial ranks.
- Then the sum of ranks of 1st sample is obtained and it is denoted as R_1 and then the sum of ranks of 2nd sample is obtained and it is denoted as R_2.
- In the next step, we work out the value of the test statistic, i.e., U, which is a measurement of the difference between the ranked observations of the two samples as under:

$$U = n_1 \times n_2 + \frac{n_1 (n_1+1)}{2} - R_1$$

- Then the critical value of U is taken from the U table for n_1 and n_2. If the U table is not available and the sample size is large (n_1 and n_2 >8) then U statistic can be converted into a Z-statistic. If the null hypothesis that the n_1+ n_2 observations came from identical populations is true, then the U statistic has a sampling distribution with:

$$\text{Mean or } \overline{U} = \frac{n_1 \times n_2}{2} \quad \& \quad \text{Standard Deviation or } \sigma_u = \sqrt{\frac{n_1 \,.\, n_2(n_1+n_2+1)}{12}}$$

Hence, the Z-statistic can be calculated by the following formula:

$$Z = \frac{U - (n_1\, n_2)/2}{\sqrt{\frac{n_1 \,.\, n_2(n_1+n_2+1)}{12}}}$$

- If the calculated value of Z happens to be less than or equal to a critical value, then the null hypothesis is accepted. On the other hand, if the calculated value of Z is larger than the critical value, then it is rejected.
- In case the sample size is small, i.e. n_1 or n_2 < 8, then we can apply an alternative method. U is found out by deducting the minimum Ws from Ws, where Ws is the smaller of R_1 or R_2 and S is the number of items in the sample with the smaller sum. Then we will compare this calculated value with the critical value of U from Wilcoxon's table and decide on acceptance or rejection of the null hypothesis.

Illustration 6: Data related to scores obtained by boys and girls in a test are given below:

Boys (B):	44	56	32	36	52	48	40	44	56	52	36	32
Girls (G):	40	48	44	36	44	24	32	16	36	44	28	30

Apply U test at 10% significance level to test that both boys and girls have come from a population with same mean.

Solution:

H_0: Sample of boys and girls have come from a population with the same mean.

H_a: Sample of boys and girls have come from a population with different mean.

Now, we will arrange all observations in ascending order and assign rank to them.

Sample Values	Ranks	Ranks of Boys (B)	Ranks of Girls (G)
16 (G)	1	-	1
24 (G)	2	-	2
28 (G)	3	-	3
30 (G)	4	-	4
32 (B)	6	6	-
32 (B)	6	6	-
32 (G)	6	-	6
36 (B)	9.5	9.5	-
36 (B)	9.5	9.5	-
36 (G)	9.5	-	9.5
36 (G)	9.5	-	9.5

40 (B)	12.5	12.5	-
40 (G)	12.5	-	12.5
44 (B)	16	16	-
44 (B)	16	16	-
44 (G)	16	-	16
44 (G)	16	-	16
44 (G)	16	-	16
48 (B)	19.5	19.5	-
48 (G)	19.5	-	19.5
52 (B)	21.5	21.5	-
52 (B)	21.5	21.5	-
56 (B)	23.5	23.5	-
56 (B)	23.5	23.5	-
Total		**R_1 = 185**	**R_2 = 115**

Now, we will find out the value of U statistic as under:

$$U = n_1 \times n_2 + \frac{n_1 (n_1+1)}{2} - R_1$$

$$= (12 \times 12) + \frac{12 (12+1)}{2} - 185$$

$$= 144 + 78 - 185 = 37$$

Or $$U = n_1 \times n_2 + \frac{n_2 (n_2+1)}{2} - R_2$$

$$= (12 \times 12) + \frac{12\,(12+1)}{2} - 115$$

$$= 144 + 78 - 115 = 107$$

Since $n_1 = 12$ and $n_2 = 12$ (both are greater than 8), so, the sampling distribution of U approximates closely with normal curve. U is transformed into a Z-statistic by the following formula:

$$Z = \frac{U - (n_1\, n_2)/2}{\sqrt{\frac{n_1 \,.\, n_2(n_1+n_2+1)}{12}}}$$

$$= \frac{37-(12\times12)/2}{\sqrt{\frac{12\times12(12+12+1)}{12}}} \qquad \text{[when U = 37]}$$

$$= \frac{37-72}{17.32} = -2.02$$

or $$Z = \frac{107-(12\times12)/2}{\sqrt{\frac{12\times12(12+12+1)}{12}}} \qquad \text{[when U = 107]}$$

$$= \frac{107-72}{17.32} = +2.02$$

Since, it is a two tailed test, the critical value of Z at 10% significance level is ± 1.64. The calculated Z value ± 2.02 is more than the critical value hence null hypothesis is rejected and we conclude that the sample of boys and girls has been drawn from population with different means.

Illustration 7: Two samples with values 90, 94, 36 and 44 in one case and the other with values 53, 39, 6, 24 and 33 are given. Test using U U-test that the samples have been drawn from populations with the same mean at a 10% significance level.

Solution:

H_0: The two samples come from populations with the same mean.

H_1: The two samples come from populations with different means.

Now, we will arrange all observations in ascending order and assign ranks to them.

Sample Values	**Ranks**	**Ranks of I Sample**	**Ranks of II Sample**
6 (II)	1	……	1
24 (II)	2	……	2
33 (II)	3	……	3
36 (I)	4	4	……
39 (II)	5	……	5
44 (I)	6	6	…….
53 (II)	7	…….	7
90 (I)	8	8	……
94 (I)	9	9	..…...
	Total	**R_1 = 27**	**R_2 = 18**

As the number of items in the two samples is less than 8 (n_1 = 4 and n_2 = 5), we cannot use the normal curve approximation technique. We shall use the table giving values of Wilcoxon's (unpaired) distribution.

Ws = Smaller of two sums = 18

S = No. of items in sample with smaller sum = 5

W_l = Larger of two sums = 27

L = No. of items in sample with larger sum = 4

Minimum Value of Ws = 1 + 2 + 3 + 4 + 5 = 15 (when S = 5)

Maximum Value of W_l = 6 + 7 + 8 + 9 = 30 (when L = 4)

U = Ws – Minimum Ws = 18 – 15 = 3

or U = Maximum W_l – W_l = 30 – 27 = 3

The probability value of U as per Wilcoxon's table from cell in column 3, row stubbed by s = 5 and L = 4 is 0.056. It is the required probability of getting a value as small as or smaller than 3 and now we should compare it with the significance level of 10%. Since the alternative hypothesis is that the two samples come from populations with different means, a two-tailed test is appropriate and accordingly 10% significance level will mean 5% in the left tail and 5% in the right tail. In other words, we should compare the calculated probability with the probability of 0.05. Since, the calculated probability (0.056) is greater than 0.05, hence, the null hypothesis is accepted and we conclude that the two samples come from populations with the same mean.

16.6 McNEMER TEST

McNemar's test is one of the important non-parametric tests which are often used when the data happens to be nominal and relates to two related samples. It is used to determine whether there is evidence of a difference between the proportions of two related samples. Based on a single sample count of data, it compares the pre-decision and post-decision response outcomes. In other words, the McNemer test is used for two related samples in situations where the attitudes of people are noted before and after treatment to test the significance of change in opinion, if any. The experiment is designed for the use of this test in such a way that the subjects initially are divided into equal groups as to their favourable and unfavourable views about a system. After some treatment, the same number of subjects is asked to express their views about the given system,

whether they favour it or do not favour it. The responses of the same subjects before and after administering the treatment can be presented in the following 2 × 2 contingency table form:

Table of Sample Responses for Testing the Significance of Change

Before Treatment	After Treatment	
	Favour	Do Not Favour
Favour	A	B
Do Not Favour	C	D

In this table, A indicates no change as the respondents' view remains positive both before and after the treatment. Similarly, D also indicates no change in respondents' views as it remains negative both before and after treatment. But B and C exhibit a change in respondents' view due to the effect of treatment. B represents the number of respondents who felt positive before the treatment and negative thereafter. Similarly, C indicates those respondents who felt negative before the treatment and positive thereafter. Thus, B and C are the only two cell frequencies relevant in deciding whether the change is significant or not. Since (B + C) indicates total change in respondents' view, the expectation under the null hypothesis is that (B + C)/2 cases change in one direction and the same proportion in the other direction. McNemer test statistic uses a transformed χ^2 test model as follows:

$$\chi^2 = \frac{(|B - C| - 1)^2}{(B+C)} \qquad \text{(with 1 d.f.)}$$

The correction of (– 1) is done to the above-mentioned χ^2 formula to make it a continuous distribution from a discrete distribution. In the last step, this calculated value is compared with the tabular value of χ^2 with 1 degree of freedom at a pre-determined significance level. If the calculated value is less than the tabular value, then the null hypothesis of no significant change is accepted, and if the calculated value is more than the tabular value, then null hypothesis is rejected.

McNemer test has an advantage over χ^2 test as that matched pairs are taken into consideration in this test whereas they are not considered in χ^2 test.

Illustration 8: A company has been working on a new branding strategy, which it thinks is more effective. Having adopted it, the manager wants to know if the new strategy is more effective than expected. A sample of 50 respondents is selected to determine their response both before and after the adoption of the new strategy. The analysis of sample response data produces the following results:

Before Adoption	After Adoption	
	Positive	Negative
Positive	10	16
Negative	12	11

Verify at a 5% level of significance whether the new branding strategy is more effective.

Solution:

H_0: The new branding strategy is not more effective.

H_1: The new branding strategy is significantly more effective.

$$\chi^2 = \frac{(|B - C| - 1)^2}{(B+C)} = \frac{(|16 - 12| - 1)^2}{(16+12)} = \frac{9}{28} = 0.32$$

The tabular value of χ^2 at 5% level of significance with 1 d.f. is 3.84. Since, the calculated value (0.32) is less than table value (3.84), hence, null hypothesis is accepted. It means that the new branding strategy is not more effective.

16.7 ONE SAMPLE RUNS TEST

One sample runs test is a test used to judge the randomness of a sample on the basis of the order in which the observations are taken. Many times, we have to deal with such situations where we do not have any control over the selection of data items. In such cases, it becomes very difficult to judge whether the selected sample is random or not. In these circumstances, we should use a runs test to test randomness in a sample. One important point to be noted here is that it is a necessary but not sufficient test for random sampling. If a supposedly random sample fails the runs test, this indicates that there are unusual, non-random periodicities in the order of the sample inconsistent with random sampling.

A run is a succession of identical letters or any other kind of symbols that is followed and preceded by different letters or no letters at all. The number of events, items, or symbols in a run is known as the length of a run. In a random data set, the probability that $(I+1)^{th}$ value is larger or smaller than the I^{th} value follows a binomial distribution, which forms the basis of the runs test. For example, XX YYYYY XXX ZZZZ XX represents a run. We can group the continued identical letters through underlines to subdivide a run into sub runs in the following manner:

XX YYYYY XXX ZZZZ XX

In this case, we have 5 sub runs (r), 7 occurrences of 'X' (n_1), 5 occurrences of 'Y' (n_2), and 3 occurrences of 'Z' (n_3). Thus, the length of run or total number of observations (N) is (7+5+3) =15.

If the size of any one kind of observations is less than 10 (i.e., n_1 or n_2 or $n_3 < 10$), then the calculated value of r is compared with the tabular value of r obtainable from Run Table. But when the size of all kinds of observations is more than or equal to 10, then we calculate the Z-statistics based on r in the following way:

$$Z = \frac{r - \mu_r}{\sigma_r}$$

Where: $\mu_r = \frac{2n_1n_2}{n_1 + n_2} + 1$ & $\sigma_r = \sqrt{\frac{2n_1n_2\,(2n_1\ n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1 + n_2 - 1)}}$

One sample runs test is not limited only to test the randomness of a series of attributes. It can even be applied to a sample consisting of numerical values by segregating the values into above median

and below median classes or runs. It is mainly helpful in testing for trends or cyclical patterns concerning economic data.

Illustration 9: A sample of 26 individuals is interviewed, with 16 women (W) and 10 men (M). These were interviewed in the following order:

M WWWW MMM WW M WWW MM WWW MMM WWWW

Use a runs test to test the randomness of this sample at a 5% significance level.

Solution:

H_0: The samples are random.

H_1: The samples are not random.

No. of runs = r = 10

No. of occurrence of women = $n_1 = 4 + 2 + 3 + 3 + 4 = 16$

No. of occurrence of men = $n_2 = 1 + 3 + 1 + 2 + 3 = 10$

Total no. of observations = N = 16 + 10 = 26

Since both n_1 and $n_2 \geq 10$, hence the Z-statistic will be calculated as follows:

$$\mu_r = \frac{2n_1n_2}{n_1 + n_2} + 1 = \frac{2(16)(10)}{16+ 10} + 1 = 13.3$$

$$\sigma_r = \sqrt{\frac{2n_1n_2\,(2n_1\ n_2 - n_1 - n_2\,)}{(n_1+n_2)^2(n_1 + n_2 - 1)}} = \sqrt{\frac{2(16)(10)(2\times16\times10-16-10)}{(16+10)^2(16+10-1)}}$$

$$= \sqrt{\frac{94080}{16900}} = 2.359$$

$$Z = \frac{r - \mu_r}{\sigma_r} = \frac{10 - 13.3}{2.359} = -1.398$$

At 5% significance level, the critical value of Z for two-tailed test is ± 1.96. Since calculated Z value is less the table value, hence the null hypothesis is accepted, and we may conclude that the sample is random.

CRITICAL APPRAISAL OF NON-PARAMETRIC TESTS

For testing hypotheses, we have two types of tests—parametric tests and non-parametric tests. We should choose a parametric test if we are sure that our data are sampled from a population that follows a normal distribution. But many times, we have to deal with such cases where various assumptions required for standard tests of significance, such as that the population is normal, samples are independent, standard deviation is known, etc. cannot be met, and then we can use non-parametric methods. We should select a non-parametric test in the following three situations:

- The outcome is a rank or a score and the population is clearly not normal.
- Some values are "off the scale", i.e. too high or too low to measure. Even if the population is normal, it is impossible to analyze such data with a parametric test since we don't know all of the values. A non-parametric test is easy to use with these data.
- The data is measured on an ordinal scale, and the population is not distributed in a Gaussian manner.

Non-parametric tests have many **advantages** over parametric tests. The biggest advantage of non-parametric tests is their versatility. These tests can be used for all kinds of data, whether the population is normal or non-normal, quantitative or qualitative. It is most suitable for ranked data. When we deal with such data that can be ranked according to respondents' preference, but their exact quantification is not possible, then we have only the option of non-parametric tests. Similarly, it is also the best option to deal with categorical or nominal data. Sometimes, we work with such data that is obtained through samples belonging to different populations. In these circumstances, we have to make some unrealistic assumptions to apply parametric tests. But there is no such problem in applying non-parametric tests. When the sample size is small or only a few observations are available then also only non-parametric tests should be applied. The main reason for popularity of non-parametric tests is their easy calculations in comparison to parametric tests. Since, non-parametric tests are easy to understand, simple to calculate, applicable to all kinds of data and less time consuming, hence, they are liked by researchers.

Although non-parametric tests have so many advantages but it also have some **disadvantages,** because of which first preference is always given to parametric tests. Non-parametric tests are less powerful than parametric tests because they are not based on many assumptions. Lack of assumptions limits the scope of making inferences. Thus, if the sampled data fulfills all desired assumptions, or the data is measured at an interval or ratio scale, then it is always considered better to use parametric tests than non-parametric tests. Similarly, if the size of the sample is large, then the calculations involved in non-parametric tests become too lengthy. Hence, in case of large samples, non-parametric tests should be avoided. Another problem with the implementation of non-parametric tests is the availability of critical value tables. To arrive at significant decisions, critical values are required. However, some of these values have not yet been compiled in relevant tables, and existing tables are always, also not easily available.

16.9 SUMMARY

Non-parametric tests are those tests that are not based on the parameters of the population; they are distribution-free tests. In this unit, you have studied some popular and commonly used non-parametric tests other than the chi-square test, which you have already studied in the previous unit.

One of the important non-parametric tests is Sign test which uses direction of differences to test if population mean is equal to hypothesized mean. There are two types of sign test—one sample sign test and two sample sign tests.

The median test is applied to determine whether the samples have been drawn from populations with the same median. It determines the significance of the difference between the medians of two or more random samples.

Another important non-parametric test is Wilcoxon Matched Pairs test which is suitable for matched pairs of before and after experiment type. In this test, the direction as well as the magnitude of difference is considered.

Yet another non-parametric test is Wilcoxon Mann Whitney test which is also known as U test. It measures the degree of separation between two independent samples. It is used to determine whether two independent samples have been drawn from the same population or from two different populations having the same distribution.

McNemer test is used for two related samples in situations where the attitudes of people are noted before and after treatment to test the significance of change in opinion if any; whereas one sample run test is used to test the randomness of the sample.

Non-parametric tests have the advantage of being versatile, usable on ranked data, nominal data and small sized sample. It has also some disadvantages, like these tests are less powerful and not suitable for large sample. Thus, parametric tests are preferred over non-parametric tests.

16.10 GLOSSARY

Sign Test-It is based on the direction (plus or minus signs) of a pair of observations and not their numerical magnitude.

16.11 CHECK YOUR PROGRESS

1. (A) Fill in the blanks:

(i) Wilcoxon Matched Pairs Test is also known as Wilcoxon……Test.

(ii) The number of events in a run is known as……of run.

(iii) McNemer test statistic uses a transformed ……test model.

(iv)……test is a very popular test amongst the rank sum test.

(v) In the Sign test, sample values greater than μ_0 are assigned ……...sign.

(B) State whether each of the following statements is true or false:

(i) We should select non-parametric tests when distribution of the population is clearly normal. ( )

(ii) If the samples are measured at an ordinal level, then we can apply the median test.( )

(iv) In the U test, when n_1 and $n_2 < 8$, then U should be converted into Z. ( )

(v) In the Median Test, we compare the computed value with a critical value of Z for testing of null hypothesis. ( )

(vi) Non-parametric tests are more suitable for small samples. ( )

16.12 ANSWERS TO CHECK YOUR PROGRESS

A. (i) Signed Rank (ii) length (iii) Chi-Square (iv) U (v) Plus

B. (i) False (ii) True (iii) False (iv) False (v) True

16.13 TERMINAL QUESTIONS

1. What is the objective of using the Sign Test?

2. State the rules regarding testing of null hypothesis in the Median Test.

3. Describe three advantages of non-parametric tests.

4. Briefly describe two non-parametric tests, explaining the significance of each such test.

5. Discuss the advantages and disadvantages of non-parametric tests.

6. Using the Sign Test at a 5% significance level, determine if the claim of the school that all its students get 80% marks on an average is true or not:

S.N.	1	2	3	4	5	6	7	8	9	10	11	12
Marks (%)	81	70	93	94	82	80	76	78	83	95	75	89

[H_0 accepted]

7. The following are the numbers of artifacts dug up by two archaeologists at an ancient cliff dwelling on 30 days

By X: 1 0 2 3 1 0 2 2 3 0 1 1 4 1 2 1 3 5 2 1 3 2 4 1 3 2 0 2 4 2

By Y: 0 0 1 0 2 0 0 1 1 2 0 1 2 1 1 0 2 2 6 0 2 3 0 2 1 0 1 0 1 0

Use the sign test at a 1% level of significance to test the null hypothesis that the two archaeologists, X and Y, are equally good at finding artifacts against the alternative hypothesis that X is better. [H_0 rejected]

8. A physical instructor claims that a particular exercise, when done continuously for 7 days, reduces body weight at least by 3.5 kg. Five overweight girls did the exercise for 7 days, and their body weights were found as under:

Girls	1	2	3	4	5
Weight before exercise	70	72	75	71	78
Weight after exercise	66	70	72	66	72

Making use of the sign test, verify the claim at $\alpha = 0.05$ that the exercise reduces weight by at least 3.5 kg. [H_0 accepted]

9. Two sections of MBA students were taught the same course on cost accounting by two different methods of teaching, known as T_1 and T_2. A sample of 6 students from each of the two sections was given a class test carrying 20 marks. The marks obtained were found as under:

Sample I (T_1)	15	10	11	12	18	15
Sample II (T_2)	12	17	14	11	09	15

Test at 0.05 level of significance using the median test if the marks obtained by MBA students in the two sections are the same as the median number. [H_0 accepted]

10. Given below are the matched pairs of data relating to the production capacity of two machines, A and B

Machine A	142	140	144	144	142	146	149	150	142	148
Machine B	138	136	147	139	143	141	143	145	136	136

Using Wilcoxon's signed rank test, verify the null hypothesis at a 1% level that there is no difference in the production capacity between the two machines. [H_0 rejected]

11. The security department of a night club wishes to select one out of two brands of hand torch batteries, B1 and B2, for normal day-to-day use. A sample of 5 batteries of brand B1 and B2 were tested for the length of useful life measured in hours. The test resulted in the following data:

B1 ($n_1 = 5$)	25	31	26	33	35	
B2 ($n_2 = 6$)	24	30	28	32	29	34

Use the U test at 5% level of significance to test the hypothesis that the two brands of batteries have the same length of useful life. [H_0 accepted]

12. In a before-and-after experiment, the responses obtained from 300 respondents were classified as follows:

Before Treatment	**After Treatment**	
	Unfavourable	**Favourable**

Favourable	60	90
Unfavourable	120	30

Test at 5% significance level, using McNemer test if there is any significant difference in the opinion of people after the treatment. [H_0 rejected]

13. Many years ago, 30 mango trees were planted alongside a road. A researcher found healthy (H) and diseased (D) trees in the following order:

HH DD HHHHH DDD HHHH DDDDD HHHHHHHHH

Use a runs test to test the randomness of this sample at a 1% significance level. [H_0 rejected]

16.14 SUGGESTED READINGS

(i) Hooda R P, 'Statistics for Business and Economics', MacMillan Business Books, New Delhi

(ii) Roy Ramendu & Banerjee S, 'Fundamentals of Research Methodology', Kitab Mahal, Allahabad

(iii) Shukla S. M. & Sahai S. P., 'Advanced Statistics', Sahitya Bhawan Publications, Agra

UNIT 17: F-TEST & ANALYSIS OF VARIANCE

Structure

OBJECTIVES

After studying this unit, you shall be able to understand:

- Concept and application of the F-test
- Concept of ANOVA
- Technique of Analysis of Variance

17.1 INTRODUCTION

In the previous two units, you have studied about chi-square test and other non-parametric tests. You must know that parametric tests are more powerful than non-parametric tests. Thus, you should rely more upon parametric tests for inferring conclusions or testing hypotheses.

In earlier units, you have already studied about some important parametric tests like the t-test, z-test, etc. The significance of the difference between the means of two samples can be judged by either z-test or the t-test. But when we are studying the significance of difference amongst more than two sample means at the same time, both these tests are not useful, and we have to use the analysis of variance. Another important parametric test is F-test which is used to test the significance of population variance for independent estimates.

In this unit, you will study about F-test and analysis of variance which will help you to judge the significance of variance between more than two samples.

17.2 F-TEST

Variance ratio test or F-test is an important test in the field of parametric tests. It is popularly known as F-test because great statistician R. A. Fisher has first used the term 'variance' and evolved this test. F-test is particularly useful when multiple sample cases are involved and the data has been measured on interval or ratio scale. The object of the F-test is to find out whether the two independent estimates of population variance differ significantly, or whether the two samples may be regarded as drawn from the normal populations having the same variance. The F-test is a very useful test which can be used to test the equality of variance of two normal populations. It can analyze variance for more than two independent samples. It can also be used for analysis of covariance. Thus, it is an important, popular and useful parametric test which can be applied in all fields like economics, business, education, agriculture, etc.

17.2.1 Assumptions of the F-Test

The F-test is based on certain assumptions that must be fulfilled for its application. These assumptions are as follows:

- The first assumption is the normality of the population. It means that values in each group are normally distributed.
- The second assumption is homogeneity of groups. It means that the variance within each group should be equal for all groups. This assumption is needed in order to combine or pool the variances within the groups into a single 'within-group' source of variation.
- The third assumption is independence of error. It means that the variation of each value around its own group mean should be independent for each value.
- The last assumption is randomness. It means that the sample items should be drawn randomly from the population.

17.2.2 Technique of the F-Test

F-test is based on the ratio of two variances. That is why it is called the variance ratio test. The ratio of two variances follows the F distribution, which is based on the above-mentioned assumptions. In this test, first of all, a null hypothesis is taken, which states that there is no

difference between the variance of the two populations. For testing this hypothesis, we have to work out the value of F (ratio of variances). F is calculated as follows:

$$F = \frac{S_1^2}{S_2^2}$$

Where, $S_1^2 = \frac{\sum(X_1 - \bar{X}_1)^2}{n_1 - 1}$ & $S_2^2 = \frac{\sum(X_2 - \bar{X}_2)^2}{n_2 - 1}$

An important point to be noted here is that the numerator is always of greater variance. It means that S_1^2 is always the larger estimate of variance (i.e. $S_1^2 > S_2^2$) . It can be expressed in the shape of the following formula:

$$F = \frac{\text{Larger Estimate of Variance}}{\text{Smaller Estimate of Varaince}}$$

For $v_1 = n_1 - 1$ = Degrees of freedom for a sample having a larger variance
$v_2 = n_2 - 1$ = Degrees of freedom for a sample having smaller variance

After computing the value of F, it is compared with the tabular value of F for v_1 and v_2 (degrees of freedom for larger and smaller variances) at the desired level of significance (5% or 1%). If the calculated value of the F ratio is less than the tabular value of F, then F-ratio is not significant and the null hypothesis is accepted. Then it can be inferred that both the samples have come from the population having the same variance. On the other hand, if the calculated value of the F-ratio is more than the tabular value of F, then the F-ratio is considered significant and the null hypothesis is rejected.

Illustration 1: Two random samples were drawn from two normal populations, and their values are as follows:

A:	66	67	75	76	82	84	88	90	92		
B:	64	66	74	78	82	85	87	92	93	95	97

Test whether the two populations have the same variance at the 5% level of significance.
(Hint: F = 3.36 at 5% level for $v_1 = 10$ and $v_2 = 8$)

Solution:

Let us take the null hypothesis that the two populations have the same variance.

A (X_1)	$(X_1 - \bar{X}_1)$	$(X_1 - \bar{X}_1)^2$	B(X_2)	$(X_2 - \bar{X}_2)$	$(X_2 - \bar{X}_2)^2$
66	–14	196	64	–19	361
67	–13	169	66	–17	289
75	–5	25	74	–9	81

76	–4	16	78	–5	25
82	+2	4	82	–1	1
84	+4	16	85	+2	4
88	+8	64	87	+4	16
90	+10	100	92	+9	81
92	+12	144	93	+10	100
			95	+12	144
			97	+14	196
720	**0**	**734**	**913**	**0**	**1298**

$\bar{X}_1 = \frac{\sum X_1}{n_1} = \frac{720}{9} = 80$ $\bar{X}_2 = \frac{\sum X_2}{n_2} = \frac{913}{11} = 83$

$S_1^2 = \frac{\sum(X_1 - \bar{X}_1)^2}{n_1 - 1} = \frac{734}{9-1} = 91.75$

$S_2^2 = \frac{\sum(X_2 - \bar{X}_2)^2}{n_2 - 1} = \frac{1298}{11-1} = 129.8$

$F = \frac{S_1^2}{S_2^2} = \frac{129.8}{91.75} = 1.4$ (The variance of the second sample has been made the numerator because the variance of the second sample is larger than first sample.

The tabular value of F for $v_1 = 10$ and $v_2 = 8$ at 5% level of significance is 3.36. Since the calculated value of F (1.4) is less than the tabular value (3.36), hence the null hypothesis is accepted. Thus, it may be concluded that the two populations have the same variance.

Illustration 2: In a study of wheat productivity in a sample of common 10 subdivisions of equal area of two agricultural plots it was seen that the sum of squared deviations of items from the mean was 0.92 and 0.26 respectively. Test at 5% significance level whether samples taken from two random populations have the same variance.

Solution:

Let us take the null hypothesis that there is no difference between the variance of the two populations.

Given that: $n_1 = 10$, $n_2 = 10$, $\sum(X_1 - \bar{X}_1)^2 = 0.92$, $\sum(X_2 - \bar{X}_2)^2 = 0.26$

$S_1^2 = \frac{\sum(X_1 - \bar{X}_1)^2}{n_1 - 1} = \frac{0.92}{10-1} = 0.102$

$S_2^2 = \frac{\sum(X_2 - \bar{X}_2)^2}{n_2 - 1} = \frac{0.26}{10-1} = 0.028$

$$F = \frac{S_1^2}{S_2^2} = \frac{0.102}{0.028} = 3.64$$

$v_1 = n_1 - 1 = 10 - 1 = 9$, $v_2 = n_2 - 1 = 10 - 1 = 9$

The tabular value of F for $v_1 = 9$ and $v_2 = 9$ at 5% significance level is 3.18. Since the calculated value of F (3.64) is more than the tabular value (3.18); hence, the null hypothesis is rejected. It means that the samples have been drawn from populations having different variances.

17.3 ANALYSIS OF VARIANCE

Analysis of variance is frequently referred to as ANOVA. It has been defined as the statistical technique for the separation of variation due to a group of causes from the variation due to other groups. It is an extremely useful technique concerning research in the fields of economics, biology, education, sociology, psychology, business or industry, and in research of several other disciplines. The ANOVA technique was initially used in agrarian research and is now actively used in research based on experimental design, whether in natural science or social science. This technique is used when multiple sample cases are involved. ANOVA is specially designed to test whether the means of more than two quantitative populations are equal. It involves classifying and cross-classifying data and then testing if the mean of a specified classification differs significantly.

According to Donald L. Harnett & James L. Murphy, "The essence of ANOVA is that the total amount of variation in a set of data is broken down into two types, that amount which can be attributed to chance and that amount which can be attributed to specified causes." There may be variation between samples and also within sample items. ANOVA consists of splitting the variance for analytical purposes. You know that the t-test is used for testing whether two population means are equal, whereas ANOVA is used for testing equality between means of multiple populations. Thus, ANOVA may be considered as an extension of the t-test.

ANOVA is a technique that can be applied in various fields. For example, this technique helps us in explaining whether various varieties of seeds or chemical fertilizers, or soils differ significantly so that the policy decision can be taken accordingly in the field of agricultural research. Similarly, through the application of this technique, the differences in various types of feed prepared for a particular class of animal or various types of drugs manufactured for curing a specific disease may be studied or judged to be significant or not. It can also be used in the field of policy decisions relating to business. For example, a manager of a big concern can analyze the performance of various salesmen working under his supervision and control to know whether their performance differs significantly. Similarly, it could be determined whether the mean qualities of the outputs of the various machines differed significantly. Similarly, it could independently be determined whether the mean qualities of outputs of the various machines differed significantly. Such a study

would determine whether uniformity in quality of outputs could be increased by standardizing the procedures of the operations or it could be increased by standardizing the machines. In this way, ANOVA could be proven a very important statistical technique for taking business-related policy decisions.

You should always remember that the analysis of variance test is not intended to serve the ultimate purpose of testing for the significance of the difference between two sample variances; rather, its purpose is to test for the significance of the differences among sample means. It is done through the mechanism of the F-test for testing the significance of the difference between two variances, but the test is so designed that the variances being compared are different only if the means under consideration are not homogeneous. In this way, significant values of F indicate that the means are significantly different from each other.

17.3.1 Sources of Variation

The analysis of variance is an important technique in the context of all those situations when the researcher wants to compare more than two populations. For analyzing the difference between various populations, we have to decide whether the differences among sample means are only due to chance or whether the differences occur because the means of various populations from which the samples have been drawn are different. There can be two types of variation, and the ANOVA technique helps us in studying these two types of variations in the data—one 'between the various samples' and another 'within sample'.

If the variations within the sample and between the samples are not significantly different from each other, then the sample is just like variations within the sample. If the variation between samples is much greater than the variation within the samples, it means that the samples come from different types of universes; otherwise, there would not have been a significant difference in the variations between the samples and within the samples. Therefore, in the analysis of variance, we find out the relationship of the variation between the samples and the variation within the samples. If the means of all the populations are equal, then the variability between samples would result only from chance and hence would be the same as the variability arising from within samples. On the other hand, if the population means are not equal, the variability between samples would be more than the variability within samples.

The measure of variability used in the analysis of variance is called a 'Mean Square', which is calculated by the following formula:

$$\text{Mean Square} = \frac{\text{Sum of squared deviation from mean}}{\text{Degree of freedom}}$$

For measuring the variability within the samples, deviations are taken from the respective sample means, and the sum of squared deviations is divided by the degree of freedom (total sample size minus number of samples), which is known as 'Mean Square Within Samples'. This mean square represents a measure of variability due to chance or experimental error. For measuring the variability between samples, deviations of sample means are taken from the grand mean of all observations, and the sum of squared deviations is divided by degree of freedom (number of samples minus one), which is known as 'Mean Square Between Samples'. This mean square represents the group effect or possible difference between samples.

If the means of all populations are equal, there is no group effect and the mean square samples will also represent variability due to chance alone. Consequently, when the sample means in the population are equal, the mean square within samples and the mean square between samples should not be much different and their ratio should be close to one. Unusually large ratios would indicate that the sample means are not equal in the population.

17.3.2 Rationale of ANOVA

The conceptual rational behind ANOVA is that the amount of variation in a set of data can be attributed to two things viz., chance and specified causes, and using ANOVA, we can split this variance for analytical purposes. ANOVA allows for investigating any number of factors that are hypothesized to influence the dependent variable. The basic principle of ANOVA is to test for differences among the populations by examining the amount of variation within the samples and the relative amount of variation between the samples. While using ANOVA, we assume that each of the samples is drawn from a normal population and that each of these populations has the same variance. It is also assumed that factors other than the one or more being tested are effectively controlled.

After selecting independent random samples for each data category, a ratio of the amount of variation between samples and the amount of variation within samples is worked out; this is known as F ratio. It can be expressed in the shape of the following formula:

$$F = \frac{\text{Estimate of population variance based on between samples variance}}{\text{Estimate of population variance based on within samples variance}}$$

Ordinarily, variance between samples would be greater than variance within samples. If the case is reverse, i.e., the variance between the samples is less than the variance within the samples, the position of the numerator and the denominator should be interchanged and conclusions drawn accordingly, but this will happen very rarely. After computing the F value, it is compared with the tabular value of F for the given degrees of freedom. If the calculated value of F is more than or equal to the tabular value, then the null hypothesis of no significant difference between sample means is rejected. It should be remembered that the ANOVA test is always one-tailed, since a low

calculated value of F from the sample data would mean that the fit of the sample means to the null hypothesis is a very good fit.

The application of the ANOVA test is based on some **assumptions,** which are as follows:

- Normality of population
- Homogeneity
- Randomness
- Independence of error

You can see that the assumptions of ANOVA test and F test are same. The fulfillment of these assumptions will obviously enhance the reliability of this test but if the populations are unimodal and sample sizes are approximately equal then the violation of assumption of normality of population would not affect the applicability of the test.

17.3.3 ANOVA Technique

Through the analysis of variance, the researcher can investigate any number of factors that are hypothesized. If the researcher takes only one factor and investigates the differences amongst its various categories, having numerous possible values, then the researcher uses one-way ANOVA and in case he investigates two factors simultaneously, then two-way ANOVA is used by him. For better decision making, two independent variables affecting a dependent variable can be studied. Based on classification of data or involvement of factors, the ANOVA technique can be divided into various classes like One-way ANOVA, Two-way ANOVA, ANOVA in Latin Square Design, etc. Different methods can be applied in various situations, which are summarized in the following form:

I. One-Way ANOVA
- (a) Direct Method
- (b) Shortcut Method
- (c) Coding Method

II. Two-way ANOVA
- (a) With no repeated values
- (b) With repeated values
- (c) Graphic Method

I - One Way ANOVA

In case of one-way or single factor ANOVA, only one factor is considered, and it is observed that this single factor is important in studying the variation within the samples and the variation between the samples. We have to examine if there are differences within that factor. In one-way

classification, the data are classified according to only one criterion and null hypothesis is taken which states that the arithmetic means of populations from which the k samples were randomly drawn are equal to one another. It can be expressed as follows:

H_0: $\mu_1 = \mu_2 = \mu_3$……..= μ_k

We can apply different alternative methods for testing of this null hypothesis which are discussed below:

(a) Direct Method

One-way ANOVA test under direct method involves following steps:

- First of all, mean of each sample is calculated:
  $\bar{X}_1, \bar{X}_2, \bar{X}_3$,…….$\bar{X}_k$ (when there are k samples)
- Thereafter, mean of sample means is calculated in the following manner:
  $$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \cdots\ldots\ldots\ldots\ldots\ldots + \bar{X}_k}{\text{No.of Samples (k)}}$$
- In the next step, deviation of the sample means from the mean of sample means is calculated.
- Thereafter, these deviations are squared and multiplied by number of items in the corresponding sample and their summation is obtained. This is known as sum of squares for variance between the samples or 'SS Between'. It can be expressed in the following form:
  SS between = $n_1(\bar{X}_1 - \bar{\bar{X}})^2 + n_2(\bar{X}_2 - \bar{\bar{X}})^2 + \ldots\ldots\ldots\ldots\ldots\ldots + n_k(\bar{X}_k - \bar{\bar{X}})^2$
- Then the sum of squares for variance between the samples is divided by degrees of freedom between samples which provide 'Mean Square Between'. Symbolically,
  $$\text{MS between} = \frac{\text{SS Between}}{(k-1)}$$
  Where (k –1) = degrees of freedom between samples
- In the next step, 'SS Within' is calculated. For this, deviation of the value of sample items for all samples from corresponding sample mean is calculated, such deviations are squared and their summation is obtained. It is known as sum of squares for variance within samples or 'SS Within'. It can be expressed in the following form:
  SS Within = $\sum(X_{1i} - \bar{X}_1)^2 + \sum(X_{2i} - \bar{X}_2)^2 + \ldots\ldots\ldots\ldots\ldots\ldots + \sum(X_{ki} - \bar{X}_k)^2$
  with i =1,2,3…….
- Thereafter, 'Mean Square Within Sample' is calculated by dividing the sum of squares for variance within sample with degrees of freedom within sample. Symbolically,
  $$\text{MS Within} = \frac{\text{SS Within}}{(n-K)}$$
  Where, n = total no. of items in all the samples, i.e. $n_1 + n_2 + \ldots\ldots + n_k$
  k = total number of samples

Thus, (n – k) represents degrees of freedom within samples.

- In the last step, F-ratio is calculated by the following formula:

$$\text{F-ratio} = \frac{\text{MS Between}}{\text{MS Within}}$$

Thereafter, this calculated value of F is compared with tabular value of F for given degrees of freedom at specified significance level. If the calculated value of F-ratio is less than the tabular value then null hypothesis is accepted and if the calculated value is more than the tabular value then null hypothesis is rejected.

This ratio is used to judge whether the difference among several sample means is significant or is just a matter of sampling fluctuations.

Additive Property of ANOVA Technique

The sum of square of deviation for total variance can be found out by adding the sum of square for variance within samples and sum of square for variance between samples. Symbolically,

SS for Total Variance = SS Between + SS Within

We can find this sum of square for total variance by an alternative method also. Its procedure involves adding the squares of deviations when the deviations for the individual items in all the samples have been taken from the mean of the sample means. Symbolically,

$$\text{SS for Total Variance} = \sum \left(X_{ij} - \bar{\bar{X}}\right)^2$$

i = 1,2,3,……..

j = 1,2,3,……

Degrees of freedom for total variance = (n – 1) = (k – 1) + (n – k)

It means that degrees of freedom for total variance will be equal to the number of items in all the samples minus one. It can also be found out by adding the degrees of freedom for between samples and degrees of freedom for within samples. This is the reason for additive property of ANOVA technique.

On the basis of various steps involved in one-way or single factor ANOVA technique, their calculations can be summarized in the form of following analysis of variance table:

Analysis of Variance Table for One-way ANOVA

Sources of Variation	**Sum of Squares (SS)**	**Degree of Freedom (d.f.)**	**Mean Square (MS)**	**F-ratio**
		(k – 1)		

(i)Between Samples	$n_1(\bar{X}_1 - \bar{\bar{X}})^2 + n_2 (\bar{X}_2 - \bar{\bar{X}})^2 + \ldots\ldots\ldots\ldots\ldots\ldots + n_k (\bar{X}_k - \bar{\bar{X}})^2$	k = No. of samples	$\frac{\text{SS Between}}{k-1}$	$\frac{\text{MS Between}}{\text{MS Within}}$
(ii) Within Samples	$\sum(X_{1i} - \bar{X}_1)^2 + \sum(X_{2i} - \bar{X}_2)^2 + \ldots\ldots\ldots\ldots\ldots\ldots + \sum(X_{ki} - \bar{X}_k)^2$ with i =1,2,3…….	(n – k) n = total no. items	$\frac{\text{SS Within}}{(n-k)}$	
(iii) Total	$\sum (X_{ij} - \bar{\bar{X}})^2$ i = 1,2,3,…….. j = 1,2,3,……	(n – 1)		

Illustration 3: To assess the significance of possible variation in performance in a certain test between the convent schools of a city, a common test was given to a number of students taken at random from the senior fifth class of each of the four schools concerned. The results are given below. Make an analysis of variance of data.

Schools			
A	B	C	D
8	12	18	13
10	11	12	9
12	9	16	12
8	14	6	16
7	4	8	15

Solution:

Let us take the null hypothesis—$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

Mean of Each Sample

$$\bar{X}_1 = \frac{8+10+12+8+7}{5} = \frac{45}{5} = 9$$

$$\bar{X}_2 = \frac{12+11+9+14+4}{5} = \frac{50}{5} = 10$$

$$\bar{X}_3 = \frac{18+12+16+6+8}{5} = \frac{60}{5} = 12$$

$$\bar{X}_4 = \frac{13+9+12+16+15}{5} = \frac{65}{5} = 13$$

Mean of Sample Means

$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4}{k}$$

$$= \frac{9+10+12+13}{4} = \frac{44}{4} = 11$$

SS Between

$$\text{SS Between} = n_1(\bar{X}_1 - \bar{\bar{X}})^2 + n_2(\bar{X}_2 - \bar{\bar{X}})^2 + n_3(\bar{X}_3 - \bar{\bar{X}})^2 + n_4(\bar{X}_4 - \bar{\bar{X}})^2$$

$$= 5(9-11)^2 + 5(10-11)^2 + 5(12-11)^2 + 5(13-11)^2$$

$$= 20 + 5 + 5 + 20$$

$$= 50$$

MS Between

$$\text{MS between} = \frac{\text{SS Between}}{(k-1)}$$

$$= \frac{50}{4-1} = 16.7 \qquad \text{(there are 4 samples)}$$

SS Within

$$\text{SS Within} = \sum(X_{1i} - \bar{X}_1)^2 + \sum(X_{2i} - \bar{X}_2)^2 + \sum(X_{3i} - \bar{X}_3)^2 + \sum(X_{4i} - \bar{X}_4)^2$$

$$= \{(8-9)^2 + (10-9)^2 + (12-9)^2 + (8-9)^2 + (7-9)^2\}$$
$$+ \{(12-10)^2 + (11-10)^2 + (9-10)^2 + (14-10)^2 + (4-10)^2\}$$
$$+ \{(18-12)^2 + (12-12)^2 + (16-12)^2 + (6-12)^2 + (8-12)^2\}$$
$$+ \{(13-13)^2 + (9-13)^2 + (12-13)^2 + (16-13)^2 + (15-13)^2\}$$

$$= \{1+1+9+1+4\} + \{4+1+1+16+36\} + \{36+0+16+36+16\} + \{0+16+1+9+4\}$$

$$= 16 + 58 + 104 + 30 = 208$$

MS Within

$$\text{MS Within} = \frac{\text{SS Within}}{(n-k)}$$

$$= \frac{208}{20-4} = \frac{208}{16} = 13$$

F-ratio

$$\text{F-ratio} = \frac{\text{MS Between}}{\text{MS Within}}$$

$$= \frac{16.7}{13} = 1.285$$

The above-mentioned calculations can be summarized in the form of following table:

Source of Variation		**SS**	**d.f.**	**MS**	**F-ratio**	**5% F-Limit**
(i)	Between sample	50	4 – 1= 3	16.7	1.285	F(3,16)
(ii)	Within sample	208	20 – 4 = 16	13		3.24
Total		258				

The calculated value of F (1.285) is less than tabular value (3.24), hence, null hypothesis is accepted and we may conclude that the samples could have come from the same universe.

(b) Short cut Method

The above-mentioned direct method of one-way ANOVA technique is very lengthy and time consuming. Instead of direct method, a short-cut method can be employed for the problems relating

to one-way ANOVA and we will obtain the same results. In fact, short-cut method is more popular than direct method and usually it is used in practice for one-way ANOVA because it is less time consuming, convenient and it reduces considerably the computational work. The short-cut method involves following steps:

- First of all, total of the values of individual items in all the samples is found out and it is known as 'T'. Symbolically,

  $T = \sum X_{ij}$ Where, i = 1,2,3……. j = 1,2,3,…..
- Thereafter, 'correction factor' is worked out as under:

  $$\text{Correction factor} = \frac{(T)^2}{n}$$
- In the next step, we find the sum of squares for total variance by squaring all the item values and taking its total and subtracting the correction factor form it.

  $$\text{Total SS} = \sum X^2_{ij} - \frac{(T)^2}{n}$$
- Then, we find out the sum of squares for variance between samples. To obtain this figure, the square of each sample total $(T_j)^2$ is divided by number of items in the concerning sample, their summation is found out and the correction factor is subtracted from this summation.

  $$\text{SS Between} = \sum \frac{(T_j)^2}{n_j} - \frac{(T)^2}{n} \qquad \text{where j = 1,2,3,…….}$$
- In the next step, the sum of squares for variance between samples is subtracted from sum of squares for total variance and the resultant figure indicates sum of squares for within samples. Symbolically,

$$\text{SS Within} = \left\{\sum X^2_{ij} - \frac{(T)^2}{n}\right\} - \left\{\sum \frac{(T_j)^2}{n_j} - \frac{(T)^2}{n}\right\}$$

$$= \sum X^2_{ij} - \sum \frac{(T_j)^2}{n_j}$$

Thereafter, the ANOVA table is constructed in the manner similar to one used for direct method.

Illustration 4: Set up an analysis of variance table for the following per acre production data for three varieties of wheat, each grown on 4 plots and state if the variety differences are significant.

Per acre production data

Plot of land	Variety of Wheat		
	A	B	C
1	6	5	5
2	7	5	4
3	3	3	3

4 8 7 4

Solution:

We will solve this problem by short-cut method

H_0: $\mu_1 = \mu_2 = \mu_3$

Null hypothesis assumes no significant difference in three varieties of wheat.

$T = \sum X_{ij}$

$= 6+7+3+8+5+5+3+7+5+4+3+4 = 60$

Correction factor = $\frac{(T)^2}{n} = \frac{(60)^2}{12} = 300$

Total SS = $\sum X^2_{ij} - \frac{(T)^2}{n}$

$= (6)^2 + (7)^2 + (3)^2 + (8)^2 + (5)^2 + (5)^2 + (3)^2 + (7)^2 + (5)^2 + (4)^2 + (3)^2 + (4)^2 - \frac{(60)^2}{12}$

$= 332 - 300 = 32$

SS Between = $\sum \frac{(T_j)^2}{n_j} - \frac{(T)^2}{n}$

$= \frac{(24)^2}{4} + \frac{(20)^2}{4} + \frac{(16)^2}{4} - \frac{(60)^2}{12}$

$= 144 + 100 + 64 - 300 = 308 - 300 = 8$

SS Within = $\sum X^2_{ij} - \sum \frac{(T_j)^2}{n_j}$

$= 332 - 308 = 24$

The ANOVA table is as follows:

Source of Variation	SS	d.f.	MS	F-ratio	5% F-Limit
(i) Between sample	8	3 – 1= 2	$\frac{8}{2} = 4.00$	$\frac{4.00}{2.67} = 1.5$	F (2,9) = 4.26
(ii) Within sample	24	12 – 3 = 9	$\frac{24}{9} = 2.67$		
Total	32				

The calculated value of F (1.5) is less than tabular value (4.26), hence, null hypothesis is accepted and we may conclude that the difference in wheat output due to varieties is insignificant and is just a matter of chance.

(c) Coding Method

Sometimes, we have to deal with big figures. It makes the calculation procedure very cumbersome. In such situation, we may take help of the coding method. It is an extension of short-cut method. Thus, coding method is used to simplify problems which involve big figures. Coding refers to the addition, subtraction, multiplication or division of data by a constant. If all the n items are either multiplied or divided by a common factor called constant or if a constant is added to or

subtracted from each of the n items, then the value of F-ratio is not affected. This means that the original measurement can be coded to simplify calculations without the need for any subsequent adjustments of the results. Once the given figures are converted with the help of some common figure, then all the steps of the short-cut method can be adopted for obtaining and interpreting F-ratio.

Illustration 5: A random sample of five motor-car tyres is taken from each of 3 brands manufactured by three companies. The lifetime of these tyres (as measured by the mileage run) is shown below. On the basis of the data, test whether the average lifetime of the 3 brands of tyres are equal or not.

Lifetime of Tyres ('000 miles)

	Brand	
A	**B**	**C**
35	32	34
34	32	33
34	31	32
33	28	32
34	29	33

Solution:

$H_0 : \mu_1 = \mu_2 = \mu_3$

The null hypothesis assumes that there is no difference between the three brands of tyres.

In order to simplify the calculations, each observation is reduced by 30. The coded data is:

A	**B**	**C**
5	2	4
4	2	3
4	1	2
3	–2	2
4	–1	3

$T = \sum X_{ij}$

$= 5 + 4 + 4 + 3 + 4 + 2 + 2 + 1 - 2 - 1 + 4 + 3 + 2 + 2 + 3 = 36$

Correction factor $= \frac{(T)^2}{n} = \frac{(36)^2}{15} = 86.4$

Total SS $= \sum X^2_{ij} - \frac{(T)^2}{n}$

$=(5)^2+(4)^2+(4)^2+(3)^2+(4)^2+(2)^2+(2)^2+(1)^2+(-2)^2+(-1)^2+(4)^2+(3)^2+(2)^2+(2)^2+(3)^2-86.4$

$= 138 - 86.4 = 51.6$

SS Between $= \sum \frac{(T_j)^2}{n_j} - \frac{(T)^2}{n}$

$= \frac{(20)^2}{5} + \frac{(2)^2}{5} + \frac{(14)^2}{5} - 86.4$

$= 80 + 0.8 + 39.2 - 86.4$

$= 120 - 86.4 = 33.6$

SS Within $= \sum X_{ij}^2 - \sum \frac{(T_j)^2}{n_j}$

$= 138 - 120 = 18$

The ANOVA table is as follows:

Source of Variation	SS	d.f.	MS	F-ratio	5% F-Limit
(i) Between sample	33.6	3 – 1= 2	$\frac{33.6}{2} = 16.8$	$\frac{16.8}{1.5} = 11.2$	F (2,12) = 3.89
(ii) Within sample	18.0	15– 3 = 12	$\frac{18}{12} = 1.5$		
Total	51.6				

The calculated value of F (11.2) is more than tabular value (3.89), hence, null hypothesis is rejected and we may conclude that the average lifetime of the 3 brands of tyres are not equal.

I- Two Way ANOVA

You must have noticed that in one-way ANOVA, the treatments constitute different levels of a single factor which is controlled in the experiment. But in real life situations, a researcher may be interested in knowing about the effect of more than one factor simultaneously or we may have to face many such situations in which response variable of interest may be affected by more than one factor. For example, agricultural output can be affected by type of fertilizer and variety of seed, product sale can be affected by advertising levels and price levels, production can be affected by different varieties of machines and different categories of labour. In these cases, we will apply two-way ANOVA. Thus, two-way ANOVA technique is used when the data are classified on the basis of two factors. We can design the test in such a way that analysis of variance can be used to test for the effects of the two factors simultaneously. With the two-way ANOVA, we can test two sets of hypothesis with the same data at the same time. The biggest advantage of this technique is that it enables the researcher to examine the interactions between the factors. A two-way design may have repeated measurements of each factor or may not have repeated values.

(a) With no repeated values

The procedure for analysis of variance in two-way classification is slightly different from the procedure which is followed in case of one-way classification. When we do not have repeated values, the sum of squares within samples cannot be computed directly. This residual or error variation is calculated by subtracting the sum of squares for variance between varieties of one treatment and sum of squares for variance between varieties of the other treatment from the sum of squares for total variance. The calculation procedure involves following steps:

- While calculating two-way ANOVA, if the figures are cumbersome then coding can be done initially and thereafter next steps are followed.

- Values or coded values of individual items in all the samples are totaled and this summation is known as 'T'. Symbolically,

$$T = \sum X_{ij}$$

- Thereafter, correction factor is found out in the following way:

$$\text{Correction Factor} = \frac{(T)^2}{n}$$

- In the next step, the correction factor is subtracted from the sum of squared individual items and the resultant figure is the sum of squares of deviation for total variance. Symbolically,

$$\text{Total SS} = \sum X^2_{ij} - \frac{(T)^2}{n}$$

- Now, we have to find the sum of squares of deviations for variance between columns. To calculate this figure, different columns are totaled and square of each column total is divided by the number of items in the concerning column and these figures are summed up and after that the correction factor is subtracted from this summation. Symbolically,

$$\text{SS Between Column} = \sum \frac{(T_j)^2}{n_j} - \frac{(T)^2}{n}$$

- After calculating 'SS Between Column' we have to calculate 'SS Between Row'. The sum of squares of deviations for variance between rows is calculated by subtracting the correction factor from sum of square of row totals divided by number of items in the concerning row. Symbolically,

$$\text{SS Between Rows} = \sum \frac{(T_i)^2}{n_i} - \frac{(T)^2}{n}$$

- In the next step, the sum of squares of deviations for variance between columns and sum of squares of deviations for variance between rows is subtracted from sum of squares of deviations for total variance and the resultant figure shows the sum of squares of deviations for residual or error variance. It can be expressed as follows:

  SS Residual = Total SS – (SS between columns + SS between rows)

- For obtaining the value of F-ratio, we must know the degrees of freedom for different sum of squares which can be worked out as under:

  d.f. for total variance = (c . r – 1)
  d.f. for variance between columns = (c – 1)
  d.f. for variance between rows = (r – 1)
  d.f. for residual variance = (c – 1) (r–1)

Where: c = No. of columns & r = No. of rows

- Then, a two-way ANOVA table is constructed in the following manner:

Analysis of Variance Table for Two-way ANOVA

Source of Variation	**Sum of Squares (SS)**	**Degrees of Freedom (d.f.)**	**Mean Square (MS)**	**F-Ratio**

(i)Between Columns (ii)Between Rows (iii)Residual or error	$\sum \frac{(T_j)^2}{n_j} - \frac{(T)^2}{n}$ $\sum \frac{(T_i)^2}{n_i} - \frac{(T)^2}{n}$ Total SS – (SS between columns + SS between rows)	$(c-1)$ $(r-1)$ $(c-1)(r-1)$	$\frac{\text{SS between Columns}}{(c-1)}$ $\frac{\text{SS between rows}}{(r-1)}$ $\frac{\text{SS residual}}{(c-1)(r-1)}$	$\frac{\text{MS between columns}}{(\text{MS residual})}$ $\frac{\text{MS between rows}}{(\text{MS residual})}$
Total	$\sum X^2_{ij} - \frac{(T)^2}{n}$	$(c.r-1)$		

It must be clear to you that residual variance is the basis of F-ratio in two-way ANOVA with no repeated value. The reason for occurrence of residual variance is fluctuations of sampling. Both the F-ratios are compared with their corresponding tabular values for given degrees of freedom at a specified level of significance. The acceptance and rejection criteria for null hypotheses on the basis of F value remain the same.

Illustration 6: The following data represent the number of units of production per day turned out by 3 different workers using 4 different types of machine. Perform a two-way ANOVA on the data given below:

Workers	**Machine Type**			
	A	**B**	**C**	**D**
I	38	40	41	39
II	45	42	49	36
III	40	38	42	42

(Use coding method subtracting 40 from the given numbers).

Solution:

Let us take the hypothesis that there is no significant difference in mean productivity with respect to machine type and different workers.

On subtracting 40 from each value, we get

Workers	**Machine Type**				**Total**
	A	B	C	D	
I	–2	0	+1	–1	–2
II	+5	+2	+9	–4	+12
III	0	–2	+2	+2	+2
Total	**+3**	**0**	**+12**	**–3**	**+12**

It is clear from the above table that 'T' or $\sum X_{ij} = 12$

Correction Factor = $\frac{(T)^2}{n} = \frac{(12)^2}{12} = 12$

Total SS = $\sum X^2_{ij} - \frac{(T)^2}{n}$

$= (-2)^2 + (5)^2 + (0)^2 + (0)^2 + (2)^2 + (-2)^2 + (1)^2 + (9)^2 + (2)^2 + (-1)^2 + (-4)^2 + (2)^2 - 12$

$= 4 + 25 + 0 + 0 + 4 + 4 + 1 + 81 + 4 + 1 + 16 + 4 - 12 = 144 - 12 = 132$

Sum of squares between machines $= \sum \frac{(T_j)^2}{n_j} - \frac{(T)^2}{n}$

$= \frac{(3)^2}{3} + \frac{(0)^2}{3} + \frac{(12)^2}{3} + \frac{(-3)^2}{3} - 12$

$= 3 + 0 + 48 + 3 - 12 = 42$

Sum of squares between workers $= \sum \frac{(T_i)^2}{n_i} - \frac{(T)^2}{n}$

$= \frac{(-2)^2}{4} + \frac{(12)^2}{4} + \frac{(2)^2}{4} - 12$

$= 1 + 36 + 1 - 12 = 26$

SS Residual = Total SS – (SS between machines + SS between workers)

= 132 – (42 + 26) = 132 – 68 = 64

The ANOVA table is as follows:

Source of Variation	**SS**	**d.f.**	**MS**	**F-ratio**	**5% F-Limit**
(i) Between machines	42	4– 1= 3	$\frac{42}{3} = 14$	$\frac{14}{10.67} = 1.31$	F (3,6) = 4.76
(ii) Between workers	26	3– 1 =2	$\frac{26}{2} = 13$	$\frac{13}{10.67} = 1.22$	F(2,6) = 5.14
(iii) Residual	64	(4 –1)(3–1) =6	$\frac{64}{6} = 10.67$		
Total	132	(4×3 – 1) =11			

Since the calculated value of both the F-ratios (1.31, 1.22) are less than their tabular value (4.76, 5.14) hence, both the null hypothesis are accepted and we may conclude that there is no significant difference in the mean productivity with respect to machine type as well as workers.

(b) With repeated values

Sometimes, we may have to face such cases in a two-way design where there are repeated measurements for all of the categories. There is only one difference in calculation procedure of two-way design with no repeated values and two-way design with repeated values. Total SS, SS between columns and SS between rows are computed in the same way. In case of repeated values, we have to calculate interaction variation. Interaction in a two-way analysis implies that the two treatments are not independent and the effect of a particular treatment in one factor depends on the level of the other factor and vice-versa. Sum of squares for variance within samples is calculated in the same manner as in case of one-way ANOVA. Interaction variation is computed on the basis of left-over sums of squares and left-over degrees of freedom. A significant interaction effect

indicates that the effect of treatment for one factor is strongly influenced by the other factor. The ANOVA table is prepared in the usual manner.

Illustration 7: Is the interaction variation significant in case of the following information concerning mileage based on different brands of gasoline and cars.

	Brands of Gasoline		
Cars	**X**	**Y**	**Z**
A	12	10	9
	12	9	11
B	12	7	10
	11	8	11
C	10	11	8
	11	11	7

Solution:

$H_{0:}$ There is no significant interaction between cars and brands of gasoline.

$T = \sum X_{ij}$

$= (12 + 12 + 10 + 9 + 9 + 11) + (12 + 11 + 7 + 8 + 10 + 11) + (10 + 11 + 11 + 11 + 8 + 7)$

$= 63 + 59 + 58 = 180$

Correction factor $= \frac{(T)^2}{n} = \frac{(180)^2}{18} = 1800$

Total SS $= \sum X^2_{ij} - \frac{(T)^2}{n}$

$= (12)^2 + (12)^2 + (10)^2 + (9)^2 + (9)^2 + (11)^2 + (12)^2 + (11)^2 + (7)^2 + (8)^2 + (10)^2 + (11)^2 + (10)^2 + (11)^2 + (11)^2 + (11)^2 + (8)^2 + (7)^2 - 1800$

$= 1846 - 1800 = 46$

SS Between Columns $= \sum \frac{(T_j)^2}{n_j} - \frac{(T)^2}{n}$

$= \left\{\left(\frac{68 \times 68}{6}\right) + \left(\frac{56 \times 56}{6}\right) + \left(\frac{56 \times 56}{6}\right)\right\} - 1800$

$= 770.67 + 522.67 + 522.67 - 1800 = 1816.01 - 1800 = 16.01$

SS Between Rows $= \sum \frac{(T_i)^2}{n_i} - \frac{(T)^2}{n}$

$= \left\{\left(\frac{63 \times 63}{6}\right) + \left(\frac{59 \times 59}{6}\right) + \left(\frac{58 \times 58}{6}\right)\right\} - 1800$

$= 661.5 + 580.17 + 560.67 - 1800 = 1802.34 - 1800 = 2.34$

SS within is calculated by subtracting each item within a group with its mean.

[e.g. A & X → (12 +12)/2 = 12; A & Y→(10 + 9)/2 = 9.5 & so on]

SS Within =

$(12-12)^2 + (12-12)^2 + (12-11.5)^2 + (11-11.5)^2 + (10-10.5)^2 + (11-10.5)^2 + (10-9.5)^2 + (9-9.5)^2 + (7-7.5)^2 + (8-7.5)^2 + (11-11)^2 + (11-11)^2 + (9-10)^2 + (11-10)^2 + (10-10.5)^2 + (11-10.5)^2 + (8-7.5)^2 + (7-7.5)^2$

$= 0 + 0 + 0.25 + 0.25 + 0.25 + 0.25 + 0.25 + 0.25 + 0.25 + 0.25 + 0 + 0 + 1 + 1 + 0.25 + 0.25 + 0.25 + 0.25 = 5$

SS Interaction = Total SS – (SS Column + SS Rows + SS Within)

= 46 – (16.01 + 2.34 + 5) = 46 – 23.35 = 22.65

The ANOVA table is as follows:

Source of Variation	**SS**	**d.f.**	**MS**	**F-ratio**	**5% F-Limit**
(i) Between columns	16.01	3– 1= 2	$\frac{16.01}{2} = 8$	$\frac{8}{0.56} = 14.28$	F (2,9) = 4.26
(ii) Between rows	2.34	3– 1 =2	$\frac{2.34}{2} = 1.17$	$\frac{1.17}{0.56} = 2.09$	F (2,9) = 4.26
(iii) Within Sample	5.00	18 – 9 = 9	$\frac{5}{9} = 0.56$		
(iv) Interaction	22.65	17–(2+2+9) =4	$\frac{22.65}{4} = 5.66$	$\frac{5.66}{0.56} = 10.1$	F(4,9) = 3.63
Total	46	18 – 1 = 17			

The value of calculated F-ratio for interaction (10.1) is more than its tabular value (3.63), hence the null hypothesis is rejected. It means that there is significant interaction between cars and brand of gasoline; hence column effect and row effect results are of no use.

(c) Graphic Method

If you have to deal with problems requiring two-way ANOVA with repeated values, then you have an option of graphic method also. Thus, in a two-way design, graphic method can also be employed to study the interaction among different factors. In graphic method, one factor is plotted on X-axis and another factor is plotted on Y-axis. Averages for all the samples are plotted on the graph and connected by distinct lines. If the lines connecting each sample items do not cross each other, then it is indicative of no interaction, whereas if the lines cross each other it implies an interaction between the factors. The graphic representation of lines concerning each sample items indicates about the type of interaction. For example, the interaction can be of an ordinal type where the rank order of effects related to one factor remains the same. If the rank order of the effects in relation to other factor changes then interaction will be of disordinal type. This disordinal interaction can be of non-cross over or cross over type.

17.4 SUMMARY

In this unit, you have studied about F-test and ANOVA. F-test is based on the ratio of two variances. It is used to find whether two samples can be regarded as drawn from normal populations having same variance. F test is based on assumptions of normality, homogeneity, randomness and independence of error. ANOVA is the statistical technique for the separation of variation due to a group of causes from the variation due to other groups. It is specially designed to test whether the means of more than two quantitative populations are equal. If the data is classified according to only one factor, then one-way ANOVA is applied whereas if the data is classified according to two factors then two-way ANOVA is applied. In one-way ANOVA, F-ratio is calculated as the ratio of mean square between and mean square within. The decision regarding acceptance and rejection of null hypothesis is taken on the basis of comparison between calculated and tabular value of F. Two-way ANOVA studies the effect of more than one factor simultaneously. In case of no repeated values, two F-ratios are calculated—one between columns and another between rows. In case of repeated values, besides the above-mentioned two F-ratios, another F-ratio for interaction variation is also computed. Acceptance and rejection criterion remains same for one-way and two-way ANOVA. In case of cumbersome data, coding method can be applied to simplify the data.

17.5 GLOSSARY

- **ANOVA**—Analysis of variance
- **SS**—Sum of squares of deviations for variance
- **MS**—Mean Square
- **Interaction effect**—effect of treatment for one factor on another factor

17.6 CHECK YOUR PROGRESS

A. Fill in the blanks:

(i) The term ‘variance’ was first used by the statistician…….

(ii) The ANOVA technique was initially used in ……..research.

(iii) Degree of freedom for computing mean square within sample consists of

…….minus……..

(iv) ……method is used for simplifying computational work in ANOVA technique.

(v) Under the graphic method,…….between factors is indicated by crossing of lines.

B. State whether each of the following statements are true or false:

(i) Homogeneity of groups is an essential assumption for the application of the F-test. ( )

(ii) F-test is based on the ratio of two standard deviations. ( )

(iii) The objective of the analysis of variance test is to test the significance of the difference between the variances of two samples. ( )

(iv) Interaction variation is computed in case of two-way ANOVA with no repeated values. ( )

(v) (n – k) indicates degree of freedom for residual variance. ( )

17.7 ANSWERS TO CHECK YOUR PROGRESS

A. (i) R. A.Fisher (ii) agrarian (iii) total sample size, no. of samples

(iv) coding (v) interaction

B. (i) True (ii) False (iii) False (iv) False (v) False

17.8 TERMINAL QUESTIONS

1. What is the objective of F-test?

2. Define the term 'Mean Square'

3. State the formula of the correction factor.

4. Describe the assumptions and technique of F-test.

5. Explain the meaning of analysis of variance. Describe briefly the technique of ANOVA for two-way classification.

6. Two random samples have been drawn from two normal populations:

Sample 1:	75	68	65	70	84	66	55
Sample 2:	42	44	56	52	46		

Test using variance ratio at 5% significance level whether the two populations have same variance. [$F = 2.37$, H_0 accepted]

7. In a sample of 8 observations, the sum of squared deviations of items from the mean was 84.4. In another sample of 10 observations, the value was found to be 102.6. Test whether the difference is significant at 5% level.

You are given that at 5% level, critical value of F for v_1 =7 and v_2 = 9 degrees of freedom is 3.29 and for v_1 = 8 and v_2 = 10 degrees of freedom, its value is 3.07. [F= 1.06, H_0 accepted]

8. The three samples below have been obtained from normal populations with equal variances. Test the hypothesis that the sample means are equal:

8	7	12
10	5	9
7	10	13
14	9	12
11	9	14

The table value of F at 5% level of significance for v_1 =2 and v_2= 12 is 3.88.

[F=4, H_0 rejected]

9. A company is interested in knowing if the three salesmen are performing equally well. The weekly sales record of the three salesmen are:

A (rupees)	300	400	300	500	0
B (rupees)	600	300	300	400	---
C (rupees)	700	300	400	600	500

[F = 3.98, H_0 accepted]

10. Three experiments determine the moisture content of samples of power, each man taking a sample from each of 4 consignments. The results are given below:

Experiment	**Consignment**			
	I	**II**	**III**	**IV**
A	9	10	9	10
B	12	11	9	11
C	11	12	10	12

Perform an analysis of variance on these data and discuss whether there is any significant difference between consignments or between experiments.

[Between consignments: F= 4.02, H_0 accepted; Between experiments: F=6.91, H_0 rejected]

11. Set up ANOVA table for the following information relating to three drugs testing to judge the effectiveness in reducing blood pressure for three different groups of people:

Amount of Blood Pressure Reduction in Millimeters of Mercury

	Drug		
Group of People	**X**	**Y**	**Z**
A	14	10	11
	15	9	11
B	12	7	10
	11	8	11
C	10	11	8

11 11 7

(i) Do the drugs act differently?
(ii) Are the different groups of people affected differently?
(iii) Is the interaction term significant?

[(i) F= 36.9, 19.0, 18.78; all three H_0 rejected]

17.9 SUGGESTED READINGS

1. Gupta S. P., 'Statistical Methods' Sultan Chand & Sons, New Delhi
2. Das N. G. 'Statistical Methods' Tata McGraw Hill, New Delhi
3. Bajpai Naval, 'Business Statistics' Pearson

UNIT 18: MULTIVARIATE ANALYSIS TECHNIQUES

Structure

OBJECTIVES

After studying this unit, you shall be able to understand:

- Basic concepts of multivariate analysis
- Classification of multivariate techniques
- Application of multivariate techniques

18.1 INTRODUCTION

In the previous units, you have studied about different types of parametric and non-parametric tests. These tests deal with such data which are based on a single factor or at most two factors. But in real life, we have to deal with many such phenomenons which are simultaneously affected by more than one factor. In case of such complex phenomenon, univariate and bivariate techniques are not helpful for analyzing them. We have to take help of multivariate analysis techniques.

Multivariate analysis refers to all statistical methods that simultaneously analyze multiple measurements on each individual or object under study. In other words, any simultaneous analysis of more than two variables can be considered as multivariate analysis. Thus, multivariate analysis techniques help to understand and interpret various relationships among variables.

In this unit, you will study about some commonly used multivariate analysis techniques like factor analysis, multiple regression, cluster analysis, MDS, conjoint analysis, etc.

18.2 CONCEPT & IMPORTANCE OF MULTIVARIATE ANALYSIS

In order to overcome the limitations of univariate and bivariate techniques, multivariate analysis techniques have emerged. Univariate analysis is helpful only in providing information about the level and distribution of the variable under study. Bivariate analysis helps only in establishing degree of relationship between the variables. If a series of univariate analysis is separately carried out for each variable, it may lead to incorrect interpretation of results. Multivariate analysis techniques help us in properly analyzing the complex simultaneous relationship between the variables.

According to **Paul E. Green**, "Multivariate techniques are a collection of procedures for analyzing the association between two or more sets of measurements that were made on each object in one or more samples of objects. If only two sets of measurements are involved, the data typically are referred to as bivariate." In simple words, when three or more variables are to be analyzed simultaneously then multivariate techniques are used.

Since, it is possible to convert complex real-life data into composite meaningful scores with the help of multivariate techniques; hence, it proves to be a powerful tool in the field of research relating to various areas like economics, sociology, psychology, anthropology, biology, agriculture, medicine, etc. It also helps in decision-making relating to business problems. In short, whenever a phenomenon is affected by many factors or variables, then multivariate techniques should be applied for its proper analysis. According to **K. Takeuchi**, "if the researcher is interested in making probability statements on the basis of sampled multiple measurements, then the best strategy of data analysis is to use some suitable multivariate statistical technique."

Multivariate techniques are empirical in nature which analyzes complex data. It transforms a large volume of data into a smaller number of composite meaningful scores. In case of a true multivariate situation, all the variables must be random and they should be interrelated in such ways that their different effects cannot be interpreted separately in a meaningful way. You should always remember that multivariate techniques are called by this name because it involves multiple variates or multiple combinations of variables and not because of multiple observations.

Another noteworthy thing in this regard is that due to its complicated lengthy calculation procedure, it is very difficult to apply multivariate techniques manually. To solve a problem through these techniques, knowledge of fundamental concepts of univariate analysis, linear algebra, vector spaces, orthogonal and oblique projections is necessary which makes the whole thing very complicated. We must use computer software programmes to solve problems through multivariate techniques. The increasing use of specialized computer software programmes in the field of research in recent years is the main reason of popularity of multivariate techniques.

18.3 TERMINOLOGY

You must be familiar with most of the terms which are used in the context of multivariate analysis because these same terms are used in univariate and bivariate analysis. But there are some additional terms also which are frequently used in this context and you should be aware of their meanings to properly understand the concept of multivariate analysis and these terms are explained below:

- **Variate—**A variate is a linear combination of variables with empirically determined weights. The variables are specified by the researcher, whereas the weights are determined by the multivariate technique to meet a specific objective. Mathematically, it can be expressed as follows:

Variate value = $w_1X_1 + w_2X_2 + w_3X_3 + \ldots\ldots\ldots\ldots\ldots\ldots + w_nX_n$

Where X_n is the observed variable and w_n is the weight determined by the multivariate technique. Thus, the entire combination of the set is represented by a single value known as variate value.

- **Metric and Non-metric variables—**Metric variables refer to those data or variables which are measured on an interval or ratio scale like age, weight, etc. Non-metric variables refer to those data or variables which are measured on a nominal or ordinal scale like name, sex, religion, etc.
- **Dependence and Interdependence techniques—**When there are one or more variables that are dependent on independent variables then dependence techniques are used. If the classification of variables into dependent and independent categories is not possible and there is mutual dependence between variables then interdependence techniques are used.

- **Explanatory and Criterion variables—**The variable which causes change in the value of another variable is known as independent or explanatory or exogenous variable. On the contrary, the variable whose value is changed due to the effect of another variable is known as dependent or criterion or endogenous variable.
- **Observable and Latent variables—**Explanatory variable can be segregated into two groups—observable and latent variable. If the explanatory variable can be directly observed then it is known as observable variable and if it cannot be directly observed then it is known as latent variable.
- **Dummy variable—**It is also known as pseudo variable. This term is used in a technical sense and is useful in algebraic manipulations in the context of multivariate analysis. X_i (i = 1, 2,…..m) is called a dummy variable if only one of X_i is 1 and the others are all zero.
- **Residuals—**The residual represents the unexplained portion of the dependent variable. It is that part of the dependent or endogenous variable which is not explained by a multivariate technique. It can be used to identify unspecified relations or problems in estimation technique.
- **Co-linearity**—It expresses the relationship between two or more independent variable. If any single independent variable is highly correlated with a set of other independent variables then it is known as multi co-linearity.

18.4 CLASSIFICATION OF MULTIVARIATE TECHNIQUES

Multivariate techniques can be classified on different bases. Its classification depends upon the answers to certain questions. The first question is:

- Can the variables be grouped into dependent and independent variables?

It means that is it possible to categorize the variables into two groups—one dependent variables and another independent variables. If we get a positive answer, i.e. some of the involved variables are dependent on others; then **dependence methods** are used. On the contrary, if we get a negative answer, i.e. none of the variable involved is a dependent variable and there is interdependence among variables; then **interdependence methods** are used. Thus, on the basis of presence of dependent variables, multivariate techniques can be classified into two groups—dependence methods and interdependence methods. They are also known as functional methods and structural methods respectively.

If we find some dependent variables then it is an indication to apply dependence methods. In the context of dependence methods, the next relevant question for its further classification is:

- How many variables are treated as dependent in a single analysis?

It means that we have to ascertain that, is there only one dependent variable or several dependent variables. After finding out the number of dependent variables, the next relevant question is:

- How are the variables measured?

It means that we have to find out which scale is used to measure the value of variables. As you know that if the data is measured on interval or ratio scale then it is known as metric and if it is measured on nominal or ordinal scale then it is known as non-metric. Thus on the basis of nature of variables, multivariate techniques can be classified into **metric** and **non-metric** methods.

In the context of dependence methods, if there is only one dependent variable and it is metric then the most popular multivariate technique is **Multiple Regression** and if this single dependent variable is non-metric then **Multiple Discriminant Analysis** is used. In case of several dependent variables if they are metric then **Multivariate Analysis of Variance** is popular and if these several dependent variables are non-metric then **Canonical Analysis** and **Conjoint Analysis** are the most desirable techniques.

In the context of interdependence methods, if the variables are metric then **Factor Analysis, Cluster Analysis** and **Metric Multidimensional Scaling**, etc. can be applied. But if the variables are non-metric then **Non-metric Multidimensional Scaling** and **Latent Structure Analysis** would be appropriate.

18.5 MULTIVARIATE TECHNIQUES

As you know that one multivariate technique cannot be used in all types of situations. In various circumstances, different types of multivariate techniques are applied. Some of the important multivariate techniques are discussed below:

18.5.1 Multiple Regression

Multiple regression is the most commonly used multivariate technique. It is an extension of simple regression. When we want to study the combined effect of two or more independent variables on one dependent variable, then multiple regression is an appropriate method. Thus, it examines the relationship between a single metric dependent variable and two or more metric independent variables. In multiple regression, a linear composite of explanatory variables is formed in such a way that it has maximum correlation with a criterion variable. The main objective of this technique is to predict the variability of the dependent variable based on its covariance with all the independent variables. In other words, the basic objective of multiple regression is to produce a model in the form of a linear equation that identifies the best weighted combination of independent variables in the study to optimally predict the criterion variable. It means that if the levels of independent variables are given, then through the multiple regression analysis model, the level of the dependent phenomenon can be predicted. For example, on the basis of a given amount of

rainfall and fertilizer, the yield of a crop can be estimated with the help of multiple regression or similarly on the basis of advertising and personal selling, amount of sales can be estimated. One important thing in this regard is that in simple regression, the estimate of dependent variable is always linear but when there are two independent variables then the estimates are in a plane. If these independent variables are three or more then they are in hyper plane.

Multiple regression is often used as a forecasting tool. It can be applied for different purposes like to predict the value of one variable from a combined knowledge of several other variables or to examine the relationship of one variable with a set of other variables or to statistically explain the variance of one variable using a set of other variables or to find out how much better a variable can be predicted if one or more predictor variables are added to the mix, etc.

Like simple regression analysis model, multiple regression is also based on the concept of regression equations. Multiple regression equation expresses the average relationship of various variables and on the basis of this average relationship the most appropriate estimate for the dependent variable is made. This equation expresses the combined effect of many independent variables on the dependent variable at the same time. The general form of a multiple regression model is as follows:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots\ldots\ldots\ldots + b_kX_k + \varepsilon$$

Where: Y = computed value of dependent variable

$X_1, X_2, \ldots\ldots X_k$ = known independent variables

$b_1, b_2, \ldots\ldots\ldots b_k$ = regression coefficients

a = constant

ε = error or residual

A noteworthy thing in this regard is that when deviations are taken from actual mean then regression equation becomes small because in such case the value of constant **a** becomes zero. Actually, the constant **a** shows the intercept made by the regression plane, thus, when the regression line passes through the origin then the value of **a** is zero. In such situation, equation in case of 3 variables will be:

$$Y = b_1X_1 + b_2X_2$$

In multiple regression, it is assumed that the conditional distributions of dependent variable for given independent variables are normal and these conditional distributions are assumed to have equal standard deviations. Multiple regression equation can also be determined by least square method. At the time of determining regression equations by least square method, there is only one difference between simple regression and multiple regression. Since, in simple regression, there are only two unknown values (a and b); thus, it can be solved by only two equations. But in case of multiple regression, more equations are required because there are more unknown values. After finding out these unknown values, they are substituted in the basic equation and thereby multiple regression equation is determined. Multiple regression equations for one dependent and two independent variables are as follows:

$$\sum Y = na + b_1 \sum X_1 + b_2 \sum X_2$$
$$\sum X_1Y = a \sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1X_2$$
$$\sum X_2Y = a \sum X_2 + b_1 \sum X_1X_2 + b_2 \sum X_2^2$$

In practice, Y and the several X variables are converted to standard scores; z_y, z_1, z_2,……z_k; each z has a mean of 0 and standard deviation of 1.

The main drawback of multiple regression is multi collinearity. In case a variable is omitted even then its effect may still be included if the excluded variable is correlated with one of the included variables. Thus the estimated coefficient of the included variable reflects both the included and excluded variable. This situation of multi collinearity is the biggest challenge for multiple regression.

18.5.2 Multiple Discriminant Analysis

Multiple Discriminant analysis is such a method in the category of dependence methods which is used in that case when the dependent or criterion variable is measured at nominal level and independent or predictor variable is measured at interval or ratio scale. It is particularly useful in that situation when individuals or objects can be classified into one of two or more mutually exclusive and exhaustive groups on the basis of a set of independent variables. The main purpose of this technique is to predict an object's likelihood of belonging to a particular group based on several independent variables. It is used for various purposes like classification of objects into different categories or to examine any significant difference between groups or to develop such discriminant function that discriminates between the different groups or evaluation of accuracy of classification, etc.

Multiple Discriminant analysis is very useful in real life situations involving one non-metric dependent variable and several metric independent variables. For example, suppose we have to find out brand preference for two different brands, viz. **A** and **B** in relation to individual's income, age and education. In this case, regression analysis cannot be applied because the dependent variable (brand preference) cannot be measured at interval or ratio scale. Here, Discriminant analysis would be appropriate.

If the dependent variable can be classified into only two groups then the term 'two group Discriminant analysis' or simply 'Discriminant analysis' is used. On the other hand, when more than two groups can be formed out of dependent variable, then the term 'Multiple Discriminant Analysis' is used.

Assumptions

There are some assumptions which must be fulfilled for the application of Discriminant analysis which are as follows:

- There should be mutually exclusive groups. Thus, each item or object should belong to only one group and there should not be any confusion about its classification.
- All cases must be independent.
- There should not be much difference in the sizes of groups of dependent variable.
- Independent variables should be measured at interval scale.
- There should not be multi collinearity.

Discriminant Function Equation

The Discriminant function is represented by the following linear equation:

$D_i = b_0 + b_1X_1 + b_2X_2 + \ldots\ldots\ldots\ldots\ldots + b_kX_k$

Where:

D_i = score on Discriminant function i

$b_1, b_2, \ldots\ldots b_k$ = discriminant coefficients

b_0 = constant

$X_1, X_2, \ldots\ldots X_k$ = independent variables

The numerical values and signs of the b's indicate the importance of the independent variables in their ability to discriminate among the different classes of individuals. Hence, through this technique it can also be determined that which independent variables are most useful in predicting whether the respondent is to be put into one group or the other.

Steps

The Discriminant analysis procedure involves the following steps:

- First of all, dependent and independent variables are identified.
- In the next step, with the help of above-mentioned linear equation, Discriminant function coefficients are determined.

  When the number of independent variables (n) is equal to 2, we have a straight line classification boundary. Every individual on one side of the line is classified as Group I and on the other side; everyone is classified as belonging to Group II. When n is equal to 3, the classification boundary is a two-dimensional plane in 3 space and in general the classification is an n–1 dimensional hyper-plane in n space. In n-group Discriminant analysis, a Discriminant function is formed for each pair of groups. If there are 6 groups to be formed, we would have $6(6 – 1) /2 = 15$ pairs of groups and hence there would be 15 discriminant functions.
- After estimating the Discriminant functions, their significance is judged through following procedure:

  F-test is applied to test if the Discriminant model as a whole is significant.

If the F-test shows significance, then the individual independent variables are assessed to observe which variables differ significantly in mean by group and these are used to classify the dependent variable.

- ❖ In the last step, results are interpreted. When the Discriminant coefficient is positive, higher the values of the independent variable, the greater are the chances that individual belonging to that category.
- ❖ This technique can also be used for judging the statistical significance between two groups. For this, Mahalanobis statistic D^2 is computed which is a generalized distance between two groups, where each group is characterized by the same set of n variables and where it is assumed that variance-covariance structure is identical for both groups. Its formula is as follows:

  $D^2 = (U_1 – U_2)\, v^{-1}\, (U_1 – U_2)'$

  Where:

  U_1 = the mean vector for group I

  U_2 = the mean vector for group II

  v = the common variance matrix

  This D^2 statistic can be transformed into F statistic which can be used to observe if the two groups are statistically different from each other.

According to D. G. Morrison, "the utility of linear Discriminant analysis lies in its strength in segregating two groups to the maximum extent." There is no doubt about the utility of Discriminant analysis. It provides a predictive equation, measures the relative importance of each variable and is also a measure of the ability of the equation to predict actual class-groups concerning the dependent variable.

18.5.3 Multivariate Analysis of Variance (MANOVA)

Multivariate or Multiple Analysis of Variance (MANOVA) is used to observe the main and interaction effects of categorical variables on multiple dependent interval variables. MANOVA uses one or more categorical independents as predictors, like ANOVA, but unlike ANOVA, there is more than one dependent variable. Where ANOVA tests the differences in means of the interval dependent for various categories of the independent(s), MANOVA tests the differences in the vector of means of the multiple interval dependents, for various categories of the independent(s). In other words, it can be said that multivariate analysis of variance is an extension of bivariate analysis of variance in which the ratio of between-groups variance to within-groups variance is calculated on a set of variables instead of a single variable. Thus, if there is only one metric dependent variable and several non-metric explanatory variables then ANOVA would be appropriate but if several metric dependent variables are involved along with many non-metric explanatory variables then MANOVA would be the most appropriate choice.

This example will help you to understand the nature of appropriate situations for application of MANOVA. A study is conducted and we try two different text books, and we are interested in the

students' improvement in Mathematics and Physics. In that case, we have two dependent variables, and our hypothesis is that both together are affected by the difference in textbooks. We could now perform MANOVA to test this hypothesis. Instead of a univariate F value, we would obtain a multivariate F value (Wilks' lambda) based on a comparison of the error variance / covariance matrix and the effect variance / covariance matrix. The covariance is included because the two measures are probably correlated and we must take this correlation into account when performing the significance test. Obviously, if we were to take the same measure twice, then we would really not learn anything new. If we take a correlated measure, we gain some new information, but the new variable will also contain redundant information that is expressed in the covariance between the variables. If the overall multivariate test is significant, we conclude that the respective effect of textbook is significant. Now, the question arises that whether only Math skills improved, only physics skills improved, or both. In fact, after obtaining a significant multivariate test for a particular main effect or interaction, customarily one would examine the univariate F tests for each variable to interpret the respective effect. In other words, one would identify the specific dependent variables that contributed to the significant overall effect.

Objectives

MANOVA can be used for following purposes:

- To compare groups formed by categorical independent variables on group differences in a set of interval dependent variables
- To identify the independent variables which differentiate a set of dependent variables the most.
- To test hypotheses concerning multivariate differences in group responses to experimental manipulations.

Variations of MANOVA

In different circumstances, different variations of MANOVA can be applied. Three basic variations of MANOVA are as follows:

- Hotelling's T—When there is one dichotomous independent variable and multiple dependent variables then this variation of MANOVA is applied which is similar to two-group T-test.
- One-Way MANOVA—When there is one multi-level nominal independent variable and multiple dependent variables then this variation of MANOVA is applied which is similar to one-way F situation.
- Factorial MANOVA—When there are multiple nominal independent variables and multiple dependent variables then this variation of MANOVA is applied which is similar to factorial ANOVA design.

One common feature among all these variations of MANOVA is that they form linear combinations of the dependent variables which best discriminate among the groups in the particular experimental design.

18.5.4 Canonical Correlation Analysis

It is a technique which is applied in case of several non-metric dependent variables. Actually, both metric and non-metric data are used in this multivariate technique. Canonical correlation analysis was propounded by Hotelling. In this technique an attempt is made to simultaneously predict a set of criterion variables from their joint co-variance with a set of explanatory variables. Thus, it may be considered as an extension of multiple regression analysis. In canonical correlation analysis, a set of weights are obtained for the dependent and independent variables in such a way that linear composite of the criterion variables has a maximum correlation with the linear composite of the explanatory variables. It measures the extent of association between the discriminant scores and the groups. Its main objective is to discover factors separately in the two sets of variables such that there would be maximum possible multiple correlation between sets of factors.

For example, if we want to analyze relationship of a set of predictor variables (x_1, x_2) to a set of criterion variables (y_1, y_2) then it can be shown as follows:

	x_1 x_2	y_1 y_2
x_1 x_2	R_{xx}	R_{xy}
y_1 y_2	R_{yx}	R_{yy}

Here, Rxx shows inter correlation between predictor or independent variables, Ryy shows inter correlation between criterion or dependent variables and Rxy shows the cross-correlation between predictor and criterion variable.

Mathematically, in canonical correlation analysis, the weights of the two sets viz., a_1, a_2, a_3,…….a_k and b_1, b_2, b_3, ……..b_j are so determined that the variables $X = a_1X_1 + a_2X_2 +……+a_kX_k + a$ and $Y = b_1Y_1 + b_2Y_2 + ….+b_jY_j + b$ have a maximum common variance. The process of finding weights requires factor analyses with two matrices. The fundamental equation for canonical analysis is as follows:

$$M = R_{yy}^{-1} R_{yx} \, . \, R_{xx}^{-1} R_{xy}$$

Here M is the asymmetric canonical matrix M.

18.5.5 Conjoint Analysis

Conjoint analysis is a decompositional method used to evaluate objects presented as a combination of attributes. It is related to traditional experimentation in which the effects of levels of independent variables are determined on a dependent variable. In this multivariate technique, overall evaluation

of the object or product is made by the respondents and that product is presented in the form of a combination of different levels of certain attributes. When as a bunch of attributes, a product is evaluated; the contribution of each attribute to overall evaluation can be ascertained. This technique is a very useful technique which is frequently used in the field of business and market research, like purchase decision or adoption decision or understanding consumers' preferences for products or services, etc. It is a powerful tool for new product development.

Conjoint analysis is most useful in behavioural studies and in marketing studies where the predictor variables are often called attributes and the dependent variable is often an overall evaluation of a product. The basic philosophy behind conjoint analysis is that subjects evaluate the value or utility of a product or service or idea whether it is real or hypothetical, by combining the separate amounts of utility provided by each attribute. It is considered as a decompositional technique because a subject's overall evaluation is decomposed to give utilities for each predictor variable and each level of a predictor variable.

Objectives

Conjoint analysis is done basically for following two objectives:

- To determine the contributions of various predictor variables and their respective values or levels to the dependent variable
- To establish a predictive model for new combinations of values taken from the predictor variables

The main advantage of conjoint analysis is that it is able to accommodate metric or non-metric dependent variables and it is also able to use non-metric variables as predictors.

Decision Process for Conjoint Analysis

For a conjoint analysis, a product class is considered along with a set of subjects who would evaluate products in that class. A set of attributes or predictor variables is selected to describe the product class. The possible levels of each attribute are selected. A product in the product class is then simply a combination of attribute levels, i.e. one level per attribute. A typical conjoint analysis involves the following steps:

- **Attribute and level selection—**The first step is to decide which attributes and which levels of each attribute are to be included in the study. For example, for a purchase decision of laptop, three attributes brand name, memory size and price are selected. Usually, more attributes are generated than are possible to analyze, so their number may need to be reduced. Generally, 5 to 15 attributes are considered for a study. Then specific attribute levels must be chosen. If an attribute is continuous, both the range of attribute levels and the number of levels need to be determined. Enough intermediate levels should be used to adequately cover the range. Generally, 2 to 5 attribute levels are considered. For example, 3 different brands, 3 different memory sizes and 2 different price ranges are selected.

- **Concept generation—**After determining attributes and levels, the concepts to be shown to respondents need to be generated. Once attributes and levels have been chosen, it may be possible to ask the respondents to evaluate every possible combination of attribute levels. This is called a full factorial design but this design is very lengthy and complicated to apply. In practice, only a subset of total concepts is presented to the respondent and it is known as fractional factorial design.
- **Parameter estimation—**In this stage, respondents are asked to rate each of the combinations. A regression is run to estimate the utility for each level of each attribute. Several different kinds of models can be used to analyze full profile data. Two different models part-worth and vector can be applied for this purpose. The conjoint analysis model can be represented by the following formula:

  $U(X) = \sum_{i=1}^{m} \sum_{j=1}^{k_i} \alpha_{ij} X_{ij}$

  Where:

  U(X) = Overall utility of a product/object

  α_{ij} = Part-worth contribution associated with j^{th} level (j =1,2.3……k) of i^{th} attribute (i=1,2, ……..m)

  k_i = No. of levels of attribute

  m = No. of attributes

  X_{ij} = 1, if the j^{th} level of i^{th} attribute is present or equal to zero
- **Analysis for decision-making—**With the help of results obtained from conjoint analysis, different kinds of decisions can be taken. For example, on the basis of conjoint analysis of different combinations of brand, memory size and price range it could be decided that which combination is most preferred by the respondents.

18.5.6 Factor Analysis

Factor analysis is the most popular multivariate technique in the category of interdependence methods. It is mainly used in those circumstances when there is clear interdependence among variables and we want to find out latent or hidden factors which create this commonality. It tries to represent a set of observed variables in terms of a number of common factors plus a factor which is unique to each variable. In other words, factor analysis is a set of procedures which is applied to resolve a large set of measured variables in terms of relatively few categories, known as factors.

For example, we are interested in observing the set of attributes that car buyers consider while purchasing a car. After consulting experts, we have identified a set of 11 features in a car on which consumers base their choice of car. These features are: Price of the car (X_1), Interior of the car (X_2), Air-conditioning (X_3), Fuel Economy (X_4), Engine Power (X_5), Seating capacity (X_6), Exterior (X_7), Availability of loan (X_9), Resale value of the car (X_{10}) and Safety (X_{11}). If you carefully observe these variables then you can expect that many of the attributes may be correlated. For example, a person who is sensitive towards the price of a car would also be sensitive towards

the fuel economy of the car and also about the availability of convenient loan facility. This means that we can reduce the above set of variables into factors which are highly correlated to one another. If we can summarize a large number of measurements with a smaller number of factors without losing too much information, we have achieved some economy of description, which is one of the goals of scientific investigation.

In factor analysis, on the basis of correlation between variables, a large number of variables are grouped into a smaller number of factors and the value of these factors or new variables or latent variables is found out by adding up the values of the original variables. These factors are linear combinations of data and factor loading representing the correlation between the particular variable and factor is usually shown in the form of a data matrix. The factor analysis model can be shown in the form of following data matrix or score matrix of n objects with k measures or variables:

Variables

Objects	a	b	c	k
1	a_1	b_1	c_1	k_1
2	a_2	b_2	c_2	k_2
3	a_3	b_3	c_3	k_3
-	-	-	-	-
-	-	-	-	-
-	-	-	-	-
n	a_n	b_n	c_n	k_n

In this matrix, a_1 shows the score of 1st object on variable a and so on. Scores on each measure are standardized using the following formula:

$$x_i = \frac{(X-\bar{X})^2}{\sigma_1}$$

Hence, the sum of scores in any column of matrix is 0 and the variance of scores in any column is 1.0

Objectives

Factor analysis can be carried out for fulfilling different purposes, main of them are as follows:

- It can be used as an explanatory technique to summarize variables and identify commonalities through the use of correlation between variables.
- Through factor analysis, large volume of data can be maximally represented through simplification. Thus, data reduction is a main objective of factor analysis.

- Factor analysis allows us to test theories involving variables which are hard to measure directly.
- It also helps us in finding out the relative weight of each factor.

Assumptions

Factor analysis is based on following three basic assumptions:

- ❖ The data should be metric. In case of non-metric data, dummy variables need to be used.
- ❖ The sample size must be sufficient so that true structures can be identified. Generally, it should vary from 5 to 20.
- ❖ The variables should be correlated to each other.

Terminology

To properly comprehend the concept of factor analysis, you should know the true meaning of the following terms which are frequently used in the context of factor analysis:

Factor—A factor is an underlying dimension that account for several observed variables. According to the nature of study and number of involved variables, there could be one or more factors.

Factor Loadings—Factor-loadings are those values which explain how closely the variables are related to each one of the factors discovered. It is a matrix representing the correlation between different combinations of variables and factors. Li(j) expression shows the factor loading of the variable j on factor i, where i = 1,2,3,……..n and j =1,2,3,……n.

Correlation Coefficient Matrix—It is the matrix of correlation coefficients of the original observations between different pairs of input variables.

Communality—Communality shows how much of each variable is accounted for by underlying factor taken together. It is shown by the expression h_i^2 . It is the sum of squares of the factor loadings of the variable i on all factors: $h_i^2 =$

$$\sum_{i=1}^{n} L_{ij}^2$$

Eigen Value—It is the sum of squares of the factor loadings of all the variables on a factor. Eigen value is also known as latent root. It indicates the relative importance of each factor for the particular set of variables being analyzed.

Rotation—If the factor loading matrix possesses a simple structure, it is easy to make interpretations. A simple structure can be observed by looking at the factor loading of each factor. If the factor loading is high on one factor and very low on the other, it is said to possess a simple structure. If there is no simple structure, then the n-dimensional space of the factors should be rotated by an angle such that the factor loadings are revised to have a simple structure which will ease the process of interpreting the factors. If the factors are independent, orthogonal rotation is

done and if the factors are correlated, an oblique rotation is made. Communality for each variable will remain undisturbed regardless of rotation but the eigen values will change as result of rotation.

Types of Factor Analysis

Factor analysis can be of following types:

R-type factor analysis—This is the most common form of factor analysis. In R-mode, rows are cases, columns are variables, and cell entries are scores of the cases on the variables. The objective is to identify groups of variables forming latent dimensions or factors. It examines variables and groups them according to underlying structures or factors. Thus, in R-mode, the factors are clusters of variables on a set of people or other entities, at a given point of time.

Q-type Factor Analysis—It is also known as inverse factor analysis where correlations are sought between respondents instead of cases. In Q-mode, the rows are variables and the columns are cases and the cell entries are scores of the cases on the variables. It forms group of respondents based on their similarity on a set of characteristics.

R-type and Q-type are the most common form of factor analysis. Beside these two forms, O-mode, T-mode and S-mode factor analysis are also used. These three forms are applied in case of time series analysis.

Methods of Factor Analysis

Factor analysis is done in a number of ways. The main methods of factor analysis are discussed below:

Centroid Method—It is a very popular method of factor analysis which was developed by L.L. Thurstone. It is easy to understand and it involves relatively simpler calculations in comparison to other techniques. It paves the way to comprehend other methods of factor analysis. The centroid method tends to maximize the sum of loadings ignoring signs. This method extracts the largest sum of absolute loadings for each factor in turn. It is defined by linear combinations in which all weights are either +1.0 or –1.0. The factors are estimated based only on common variance and excludes specific and error variance.

Principal Component Analysis Method—This method finds new variables which are linear combinations of the observed variables so that they have maximum variation and are uncorrelated. It was developed by H. Hotelling. It is suitable in a situation where the objective is to summarize data while retaining maximum amount of variation. This method provides a unique solution, so that the original data can be reconstructed from the results. In this method, the total variance in the data is considered and the factors based on total variance are called principal components. These

factors account for the greatest portion of total variance. Usually, it can be seen that principal component analysis method provides an improved solution over the centroid method.

Factor Rotation

The mathematical process used to obtain a factor solution from a correlation matrix is such that each successive factor, each of which is uncorrelated with the other factors, accounts for as much of the variance of the observed variables as possible. Rotation serves to make the output more understandable and is usually necessary to facilitate the interpretation of factors. In other words, factor rotation is the process of manipulating or adjusting the factor axis to obtain a more relevant factor solution.

There are two types of rotations used in factor analysis. One of them is **Orthogonal Rotation**. If the factor axes are maintained at a 90^o angle, then it is known as orthogonal rotation. They are so rotated that the factor axis best fits the variables. There are some variations of orthogonal rotation. **Varimax rotation** is an orthogonal rotation of the factor axes to maximize the variance of the squared loadings of a factor (column) on all the variables (rows) in a factor matrix. The variance of a factor is largest when its smallest loadings tend towards zero and its largest loadings tend towards unity. This is the most common rotation option. Another orthogonal rotation option is **Quartimax rotation** which minimizes the number of factors needed to explain each variable. This type of rotation often generates a general factor on which most variables are loaded to a high or medium degree.

The other type of factor rotation is non-orthogonal or **Oblique Rotation.** In this rotation, factors are allowed to be correlated. This will result in higher eigen values but diminished interpretability of the factors.

18.5.7 Cluster Analysis

Cluster analysis is an exploratory data analysis tool for solving classification problems. It is used for combining observations into groups such that each group is homogeneous with respect to certain characteristics and each group is different from other groups with respect to same characteristics. Thus, its object is to sort cases into groups or clusters. Technically, a cluster consists of variables that correlate highly with one another and have comparatively low correlations with variables in other clusters. Hence, the basic objective of cluster analysis is to determine how many mutually exclusive and exhaustive clusters exist in the population and thereafter it aims to describe its composition. In case of cluster analysis, data may be thought of as points in a space where the axes correspond to the variables. For example, if income is shown on X axis and savings on Y axis then data collected on different individuals' income and savings can

be plotted in the form of points on the graph. On the basis of these points, different clusters can be formed. Suppose, two clusters 'A' and 'B' are formed where A shows people having low income and high savings while B shows group of people having high income and low savings. In this situation, people within a cluster have similar characteristics but they differ significantly with other clusters against the same characteristics. There would be some people who have not been grouped into any cluster as their classification is not obvious or they do not belong to these specific groups or clusters.

So, cluster analysis is a tool of discovery which reveals associations and structure in data and its results may contribute to the definition of a formal classification scheme. It is applied mainly for following two objectives:

- To find out clusters (i.e. mutually exclusive and exhaustive groups) in the population
- Data reduction by grouping individuals into clusters

Clustering Procedure

Clustering can be done in a number of ways. The main methods of clustering can be classified into two broad categories:

Hierarchical Method—Hierarchical clustering develops a tree like cluster using bottom up approach (agglomerative clustering) or top down approach (divisive clustering). Agglomerative methods start with each observation as a cluster and with each step combine observations to form clusters until there is only one large cluster. On the other hand, divisive methods begin with one large cluster and proceed to split into smaller cluster items that are most dissimilar.

Non-Hierarchical Method—Non-hierarchical clustering takes a fixed number of clusters and then tries to find the optimal solution by ensuring maximum cluster difference. It is the partitioning of the sample. Each cluster has a seed point and all objects within a prescribed distance are included in that cluster. Another way of non-hierarchical clustering is to loop through the sample, assigning each case to the seed point to which it is closest. Sequential threshold, Parallel threshold and Optimizing partitioning are three different approaches of non-hierarchical clustering.

When respondents are to be grouped rather than variables then cluster analysis is a better technique than factor analysis. But cluster analysis is a subjective method of segmentation, thus, there is a chance of biasness in the analysis. Another limitation of cluster analysis is that it is not possible to test the statistical significance of results obtained from this analysis.

18.5.8 Multidimensional Scaling

Multidimensional scaling is a data reduction technique whose main objective is to uncover hidden structure of a data set. It is used to measure an item in more than one dimension at a time. It is an important tool in measuring the perception or preferences of people across a large number of attributes. The basic assumption behind multidimensional scaling is that people perceive a set of objects as being more or less similar to one another on a number of dimensions instead of only one. It means that similarity or dissimilarity of objects is determined by respondents on more than one dimension simultaneously. In practice, perceptions are represented spatially in a multidimensional plot and are called perceptual maps.

Types of MDS

Multidimensional scaling is an important technique in the category of interdependence methods. MDS techniques are also known as techniques for dimensional reduction. It can be applied in both metric and non-metric data. Thus, multidimensional scaling can be divided into two broad categories--Metric MDS and Non-Metric MDS.

In case of metric MDS, both input and output data are metric in nature. Thus, there is a stronger relationship between input and output data and most of the qualities of input data are observed.

In case of non-metric MDS, input data is non-metric and output data is metric in nature because generally, input data is measured at ordinal scale and it is plotted on spatial map which is interval scaled. When data happens to be non-metric, rank ordering of each pair is done in terms of similarity and then judged similarities are transformed into distances through statistical manipulations and are consequently shown in n-dimensional space in a way that the inter-point distances best preserve the original inter-point proximities.

Significance

MDS is a very useful technique which is frequently used by the researchers for different purposes. It is a good scaling technique to measure the respondents' attitudes against multiple attributes. With the help of MDS, objects, individuals or both can be scaled from a minimum of information. It also helps in identifying the most salient attributes which happen to be primary determinants for making a specific decision. Besides that, spatial maps are particularly useful in identifying the gaps in people perceptions and thus locate newer opportunities.
Though it is a useful technique but there are some limitations associated with it. For example, the idea of concepts like similarity and preferences is not clear and each respondent's perception is different from one another. Similarly, it becomes difficult to interpret results when the researcher is trying to relate changes in stimuli to changes in perceptual maps.

18.5.9 Latent Structure Analysis

Latent structure analysis is another technique in the category of interdependency methods. It is carried out mainly for two purposes—to extract latent factors and express relationship of observed variables with these factors as their indicators and to classify a population of respondents into pure types. Thus you can see that its objectives are similar to the objectives of factor analysis. This type of analysis is appropriate when the variables involved in a study do not possess dependency relationship and happen to be non-metric. These techniques are particularly useful when the amount of data is huge and intractable. Latent structure analysis involves large number of models like latent class analysis, latent trait analysis, latent profile models, etc.

18.6 SUMMARY

Multivariate analysis refers to all statistical methods that simultaneously analyze multiple measurements on each individual or object under investigation. Multivariate analysis originated from univariate and bivariate statistics. Multivariate techniques can be classified on the basis of three factors: (i) grouping of variables into dependent and independent variables, (ii) number of dependent variables treated in a single analysis, (iii) the scale of measurement of the variables. Multivariate analysis has primarily been classified into two groups, dependence and interdependence technique, based on the relationship they examine.

In this unit, you have studied about the concept of some popular multivariate techniques. Multiple regression analysis is a measure of relationship involving a single dependent variable and two or more independent variables. Multiple Discriminant analysis is used when dependent variable is nominally scaled and independent variable is interval or ratio scaled. Multivariate analysis of variance or MANOVA analyses significant differences between means for more than one metric dependent variable at a time. Canonical correlation analysis is used to predict a set of dependent variables from their joint covariance with a set of explanatory variables. Conjoint analysis is a decompositional method used to evaluate objects presented as a combination of attributes. All these techniques belong to the category of dependence methods.

Factor analysis, Cluster analysis, Multidimensional scaling and Latent structure analysis are included in the category of interdependence methods. Factor analysis is set of procedures used for data reduction and summarizing data. Two important methods of factor analysis are centroid method and principal component analysis method. Cluster analysis is used for combining observations into homogeneous groups. Multidimensional scaling is a data reduction technique whose main objective is to uncover hidden structure of a data set.

18.7 GLOSSARY

Metric data—Variables measured on interval or ratio scale

Non-Metric data—Variables measured on nominal or ordinal scale

Explanatory variable—Variable which causes change in another variable

Criterion variable—Variable whose value is changed due to effect of another variable

Eigen value—Sum of squared values of factor loadings relating to a factor

Factor rotation—Process of manipulating or adjusting factor axis

Multi Collinearity—High correlation of a single independent variable with a set of other independent variables

18.8 CHECK YOUR PROGRESS

A. Fill in the blanks:

(i). Dummy variable is also known as ……. Variable.

(ii) In case of single metric dependent variable, ……is the most desirable technique.

(iii). In Multiple Discriminant Analysis, there should be ……groups.

(iv).....analysis is a decompositional method.

(v). Agglomerative and ……..method are two approaches of hierarchical clustering.

(vi). Multidimensional scaling is an important technique in the category of …… methods.

B. State whether each of the following statements is true or false:

(i). Metric variables are measured on nominal scale. ( )

(ii) Hotelling' T is a variation of MANOVA. ( )

(iii). Varimax and Quartimax rotation are types of oblique rotation. ( )

(iv). Canonical Correlation Analysis was propounded by Spearman. ( )

(v). Rows are cases and columns are variables in R-type factor analysis. ( )

(vi). Co-linearity expresses the relationship between two dependent variables. ( )

18.9 ANSWERS TO CHECK YOUR PROGRESS

A. (i) Pseudo (ii) Multiple Regression (iii) Mutually Exclusive

(iv) Conjoint (v) Divisive (vi) interdependence

B. (i) False (ii) True (iii) False (iv) False (v) True (vi) False

18.10TERMINAL QUESTIONS

1 What is orthogonal factor rotation?

2. State the cases where interdependency methods should be used.

3. Define Metric Variables.

4. Differentiate between explanatory and criterion variables

5. What do you mean by Multivariate techniques?

6. Explain the classification of multivariate techniques with suitable examples.

7. Discuss the assumptions and method of Multiple Discriminant Analysis.

8. Write a note on Factor Analysis Method.

9. Explain the different types of clustering methods. Discuss its importance and limitations.

10. Briefly describe any two multivariate techniques belonging to the category of dependency methods.

18.11 SUGGESTED READINGS

1. Roy Ramendu & Banerjee Subhojit, 'Fundamentals of Research Methodology' Kitab Mahal, Allahabad
2. Kothari C. R., 'Research Methodology: Methods & Techniques', New Age International Publishers, New Delhi
3. Gupta Shashi K. & Rangi Praneet, 'Research Methodology (Methods, Tools and Techniques)', Kalyani Publishers, New Delhi

BLOCK:6 RESEARCH METHODOLOGY

UNIT 19: CONCEPT, APPROACHES AND METHODS

OBJECTIVES

The objectives of this unit are as follows:

- To understand the concept and meaning of research.
- To understand the need of research.
- To understand the process of formulating of research problem.

- To understand the importance of scientific thinking in research.

19.1: INTRODUCTION

Various discipline of study requires an update in its existing knowledge with the latest facts and theories. To achieve this, it is suggested to research the existing knowledge related to the specific area of study. The word research is in use since immemorial and is considered as an important tool for updating the existing knowledge with the latest facts. For the student of commerce and business management, research methodology is an important subject of study. Understanding of this subject helps the students in developing problem-solving skills and logical thinking to look into required aspects of the phenomena or situation to come out with optimum and logical solution. This unit of research methodology explained the fundamental concepts and approaches of research methods. After reading this unit students would be able to develop understanding about the processes and characteristics of the research methodology.

19.2: DEFINING RESEARCH AND RESEARCH METHODOLOGY

There are various ways by which the term "research" is defined in the literatures. Literally research is the systematic investigation and study of area of problem or an area of interest. It is a process that starts with the analysis of the situation itself to find out the causes of characteristics of the problem situation. The situation may be related to anything. For an organization, the situation may be related to the problem of absenteeism of employee or assessing the customer' satisfaction to take the suitable decision. The manager needs an impartial assessment of the nature of the problem and on the basis of the suggested assessment, the manager has to take the decision.

The research is conducted by the researcher. **The researcher** is a person who is doing research work by investigating the situation and is supposed to draw a valid conclusion related to the situation. **Research work** involves a scientific exploration and investigation of the defined situation. The scientific study and investigation is carried out by researchers by using standardized tools and techniques as provided in the subject matter of "**Research Methodology".**

Research in common parlance refers to a search for knowledge. One can also define research as a scientific and systematic search for pertinent information on a specific topic. In fact, research is an art of scientific investigation. The Advanced Learner's Dictionary of Current English lays down the meaning of research as "**a careful investigation or inquiry especially through search for new facts in any branch of knowledge**." Redman and Mory define research as a "**systematized effort to gain new knowledge**."

Now you have understood the meaning of research. As the word "research" has been defined from various perspectives it would be useful to interpret various definitions of research to understand its characteristics from different perspectives. Some definitions of research are as follows-

- Research is a diligent and systematic inquiry or investigation into a subject in order to discover facts or principles.
- Research could also be defined as the scientific investigation of phenomena, which includes collection, analysis, interpretation and presentation of data according to defined objectives to draw conclusions.
- Research is a structured enquiry that utilizes acceptable scientific methodology to solve problems and create new knowledge that is generally applicable

The word **research methodology** is used to describe set of tools and techniques used in research method(s). Methodology is a complete science of research method to be followed. Some perspectives which are associated with research methodology are discussed in next section.

19.3: PERSPECTIVES OF RESEARCH

The first step to research is the exploration and understanding of the problem or situation for study and investigation. After that the researcher tries to find out the reasons for that problem. Hence, research is a two stage processes involving following stages.

Stage 1: Analysis and understanding of characteristics of phenomena under study

Stage 2: Analysis and understanding of the reasons for those characteristics

Research is also defined from different perspectives. These perspectives are as follows

Paradigm: It refers to the school of thought which are followed to carry out the research

Approach: It is all about the way a researcher can search for the answer of the question and communicate the answer

Instruments: It refers to standardized tools and methods used in research

19.4: CHARACTERISTICS OF RESEARCH

After defining research from different perspectives and understanding the stages, you have to understand the characteristics of research and research methodology. Remember, while we are describing characteristics of the research, we have considered **research as a process** that involves various interconnected sequential activities. In a day-to-day environment, various other activities compel the individual (manager or researcher) to take a decision. Various activities would be considered as research only when it fulfills certain requirements. We have identified these requirements of research by examining the definitions of research presented in previous section. Now we can list the basic requirement of the investigation and study as follow.

- Research is a **systematic** process of investigation and study.
- Research is **universal** in nature and it is used in all discipline of study.
- Research involves **collection and analysis** of data.
- Research involves **scientific logics and reasoning** in drawing inferences and conclusion.
- Research is a **cyclic process** as end of a research leaves scope for further research.
- Research uses **proven and standardised tools** for data collection and analysis.
- Research is normally **conclusive** in nature.

As said earlier, research is a process and a process involves set of interconnected and sequential activities. Every activity could not be considered as research. There are certain characteristics of the activities associated with the research

- **Activities must be systematic**: It implies that a set of logical sequences for activities are followed in research.
- **Activities must be controlled**: It indicates that the factors and variables under study are properly manipulated to infer the cause-effect relations between two variables. It also explains that the researcher set up his research study in such a way that minimizes the effects of other factors affecting the relationship between variables.
- **Activities must be valid:** It implies that the purpose for which the activities and processes related to a specific research is carried out; those are able to do so.
- **Activities must be verifiable:** This concept implies that whatever a researcher conclude on the basis of his/her findings can be verified by you and others
- **Activities must be rigorous:** It means researcher must be scrupulous in ensuring that the procedures followed to find answers to questions are relevant, appropriate and justified. Although degree of rigor varies markedly for different types of research.

Practice Exercise: ***Interact with a "research scholar" of your university and "marketing manager" of a company to understand the characteristics of various steps and activities they follow in "research process" and "decision making". Identify the activities to be considered as research on the basis of characteristics of the research process as described in this section?***

19.5: OBJECTIVES OF RESEARCH

The purpose of research is to discover answers to questions through the application of scientific procedures. The objectives of research could be listed as follow

- To gain familiarity with a phenomenon or to achieve new insights into it
- To portray accurately the characteristics of a particular individual, situation or a group
- To determine the frequency with which something occurs or with which it is associated with something else
- To test a hypothesis of a causal relationship between variables

19.6: STEPS IN RESEARCH

There are many models available and taught concerning how to conduct a research process and the process presented here is only one of many. In general research process consists of following steps

Step 1: Decide on a topic

Step 2: Develop an overview of the topic

Step 3: Determine the information requirements

Step 4: Collecting the information

Step 5: Organizing the information

Step 6: Analyzing the information

Step 6: Drawing inferences and conclusion

Step 7: Communicate the conclusion and findings

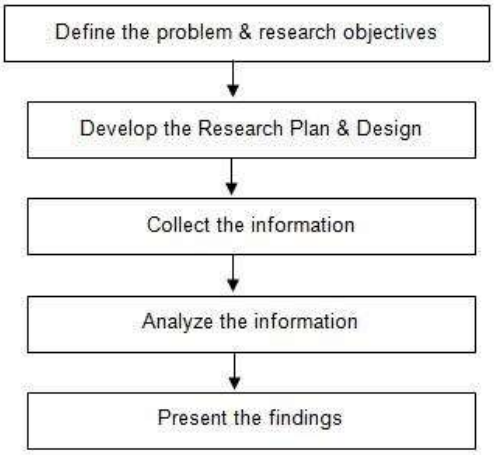


Fig: 1: Steps in Research Process

Practice Exercise: ***You have to conduct an opinion poll for the post of "President of Student Union" under proposed election in your university. Write down the steps required to conduct this opinion poll. Compare the steps selected by you with the steps suggested above in this section for the research process. List down the sequence of activities required to be performed for each steps to complete this opinion poll.***

19.7: NEEDS OF THE RESEARCH

In this section we shall try to explain the practical uses of research in business and commerce. Business environment needs careful study of the situation for taking decision. Research can help in the effort to investigate a specific problem encountered in the business and commerce to take **rational decision** for solving the problems. In other area of study research could be used to update the existing knowledge of the subject matter. Research is important in each and every field.

Research helps us to know that how our business is running and it also helps us to know that what all changes we have to do in our business to earn more profit

We can summarize the need of research as follow.

- To improve quality of life by updating the existing knowledge of the area of study
- To help manager to take rational and good decision
- Effective business plan for product launching
- Getting feedback from customers for marketing planning

Research may be used in **basic** area of knowledge or in **applied** area of, depending upon the purpose of the research. In basic area of study a research is oriented towards search and validation of new concepts and theories to update the related knowledge. In case of applied area research is used to solve specific business problem

Practice Exercise:: ***Do you feel that the research is required for all managerial situations? Try to explain your view with suitable examples.***

Hints: Research is its conceptual form may not be used for all managerial decisions. Some managerial decisions are based on intuition and experiences also. It depends on the extent of investigation and requirement of the decision maker to consider a carry out the steps in research process according to the defined characteristics of research or not.

19.8: FORMULATION OF RESEARCH QUESTION

Lets now understand the steps of research process. The first step is formulating and defining the research question. Research Question is a specific question about a part of the research problem that will be investigated and answered by the researcher. A researcher might have several questions about the problem. Among various questions, one specific question from the interest is to be selected as research problem and can be answered through systematic investigation, based on the fundamental principles of research.

The following points should be kept in mind while defining a research question:

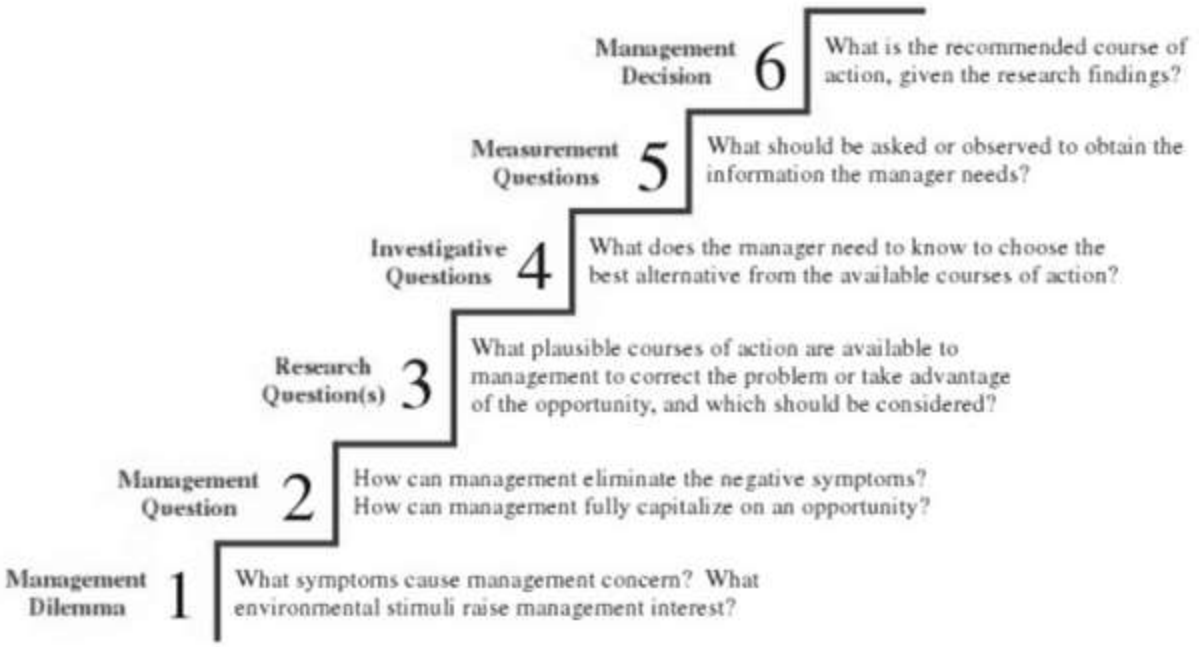
1) The right question must be addressed from research problem.

2) The research question must be completely defined.
3) The research question must be validated from the expected answers required by the decision maker.
4) The researcher must ensure that sufficient data would be available for investigation related to different areas of research question.
5) Proper scientific and statistical tools must be available to analyse the data related to the research question.

In order to develop a research question, a research interacts extensively with the decision maker or users of the outcome of the research. The purposes of interactions are just to validate the definitions of research problem and exploring various issues from broad research problem. Among various research issues one specific issue is selected for defining the research question. For example a manager is facing problem of decreased sales of the product of his company. This is known as managerial problem to be addressed by the manger. The concerning manager may be interested to explore the reasons for this problem. When the manager interacts with a researcher, the researcher will try to understand the problem from various perspectives such as

- What are the symptoms that considered by manager as "declining sales"
- What are the characteristics natures of the symptoms related to declining sales i.e seasonal, erratic or continuous
- The researcher may ask to the manager to be specific about the mode of investigation to know the reasons for declining sales.

Once the managerial question is understood and analysed properly by the researcher, it is converted into research question. Schindler & Cooper suggested following model for management –research hierarchy in formulation of research problem.



Research problem formulation is the first step toward achieving the goal of the research. The origin of research question may be attributed to search for the facts or motivation to solve the business or managerial problem. As Northrop (1966) writes, "Inquiry starts only when something is unsatisfactory, when traditional beliefs are inadequate or in question, when the facts necessary to resolve one's uncertainties are not known, when the likely relevant hypotheses are not even imagined. What one has at the beginning of inquiry is merely the problem" The formulation of research problems also has an important social function.

After reviewing various books and literatures it is found that there is little agreement among social scientists on the most effective procedure for formulating research problems. In particular, there has been considerable debate over whether or not it is important to define problems explicitly in research or not. Although it is advisable that the research problem as well as research question must be properly understood and defined by the research as well as the decision makers.

Considerations in selecting a research problem:

- **Interest**: a research endeavour is usually time consuming, and involves hard work and possibly unforeseen problems. One should select topic of great interest to sustain the required motivation.

- **Magnitude:** It is extremely important to select a topic that you can manage within the time and resources at your disposal. Narrow the topic down to something manageable, specific and clear.
- **Measurement of concepts**: Make sure that you are clear about the indicators and measurement of concepts (if used) in your study.
- **Level of expertise**: Make sure that you have adequate level of expertise for the task you are proposing since you need to do the work yourself.
- **Relevance**: Ensure that your study adds to the existing body of knowledge, bridges current gaps and is useful in policy formulation. This will help you to sustain interest in the study.
- **Availability of data**: Before finalizing the topic, make sure that data are available.
- **Ethical issues**: How ethical issues can affect the study population and how ethical problems can be overcome should be thoroughly examined at the problem formulating stage.

Practice Exercise: ***A company XYX Limited is facing stiff competition from the competitors for its product. The marketing manager is interested to know the customers satisfaction level with its specific brand of a product. The marketing manager hired you as researcher to investigate the problem. Identify the issues and research questions from managerial perspectives and researchers perspectives.***

19.9: RESEARCH PHILOSOPHY

A research philosophy is a belief about the way in which data about a phenomenon should be gathered, analysed and used. There are two terms that encompasses various philosophies of research approach. These are

- Epistemology: It deals with "what is known to be true"?
- Doxology: It deals with "what is believed to be true"?

Two major research philosophies have been identified in the western tradition namely **positivist** (sometimes called scientific) **and interpretivist** (also known as ant positivist)

- **Positivism:** Positivists believe that reality is stable and can be observed and described from an objective viewpoint (Levin, 1988), i.e. without interfering with the phenomena being studied. They contend that phenomena should be isolated and that observations should be repeatable. This often involves manipulation of reality with variations in only a single independent variable so as to identify regularities in, and to form relationships between, some of the constituent elements of the social world. Predictions can be made on the basis of the previously observed and explained realities and their inter-relationships. "Positivism has a long and rich historical tradition. It is so embedded in our society that knowledge claims not grounded in positivist thought are simply dismissed as a scientific and therefore invalid" (Hirschheim, 1985, p.33). This view is indirectly supported by Alavi and Carlson (1992) who, in a review of 902 IS research articles, found that all the empirical studies were positivist in approach. Positivism has also had a particularly successful association with the physical and natural sciences.

  There has, however, been much debate on the issue of whether or not this positivist paradigm is entirely suitable for the social sciences (Hirschheim, 1985), many authors calling for a more pluralistic attitude towards IS research methodologies (see e.g. Kuhn, 1970; Bjørn-Andersen, 1985; Remenyi and Williams, 1996). While we shall not elaborate on this debate further, it is germane to our study since it is also the case that Information Systems, dealing as it does with the interaction of people and technology, is considered to be of the social sciences rather than the physical sciences (Hirschheim, 1985). Indeed, some of the difficulties experienced in IS research, such as the apparent inconsistency of results, may be attributed to the inappropriateness of the positivist paradigm for the domain. Likewise, some variables or constituent parts of reality might have been previously thought unmeasurable under the positivist paradigm - and hence went unresearched, (after Galliers, 1991).

- **Interpretivism:** Interpretivists contend that only through the subjective interpretation of and intervention in reality can that reality be fully understood. The study of phenomena in

their natural environment is key to the interpretivist philosophy, together with the acknowledgement that scientists cannot avoid affecting those phenomena they study. They admit that there may be many interpretations of reality, but maintain that these interpretations are in themselves a part of the scientific knowledge they are pursuing. Interpretivism has a tradition that is no less glorious than that of positivism, nor is it shorter

19.10: RESEARCH PARADIGM

The most quoted definition of paradigm is given by Thomas Kuhn's (1962, 1970) concept in The Nature of Science Revolution. It is defined as the underlying assumptions and intellectual structure upon which research and development in a field of inquiry is based. According to Patton (1990) a paradigm is a world view, a general perspective, a way of breaking down the complexity of the real world.

Paradigm is also considered as an interpretative framework, which is guided by "a set of beliefs and feelings about the world and how it should be understood and studied." (Guba, 1990). Denzin and Lincoln (2001) listed three categories of those beliefs:

i. **Ontology**: In general, ontology refers to the study or concern about what kinds of things exist within society.

ii. **Epistemology**: Epistemology deals with the issue of knowledge, and specifically, who can be a 'knower'. It describes the relationship between the inquirer and the known. It is defined Gall, Borg, & Gall(1996) as the branch of philosophy that studies the nature of knowledge and the process by which knowledge is acquired and validated

iii. **Methodology**: It is concerned with "how do we know the world, or gain knowledge of it"? Methodology is a series of choices that describes followings

- Choices about what information and data to gather
- Choices about how to analyze the information and data that you gather
- Other methodological choices

When challenging the assumptions underlying positivism, Lincoln and Guba (2000) also identified two more categories that will distinguish different paradigms, i.e. beliefs in causality and **oxiology**. The assumptions of causality assert the position of the nature and possibility of causal relationship; oxiology deals with the issues about value. Specific assumptions about research include the role of value in research, how to avoid value from influencing research, and how best to use research products (Baptiste, 2000).

Dill and Romiszowski (1997) stated the functions of paradigms as follows:

- Define how the world works, how knowledge is extracted from this world, and how one is to think, write, and talk about this knowledge
- Define the types of questions to be asked and the methodologies to be used in answering
- Decide what is published and what is not published
- Structure the world of the academic worker
- Provide its meaning and its significance
-

19.11: RESEARCH APPROACHES

Research can have elements which are based upon a **non-empirical** approach, an **empirical approach**, or a combination of the two. For the empirical approach, there are three primary dimensions which can be evaluated for use:

1. Qualitative vs. quantitative
2. Deductive vs. inductive
3. Subjective vs. objective.

Before we discuss the approaches of research we have to group these approaches into two broad categories as empirical and non empirical.

- Empirical research is defined as research based on observed and measured Phenomena. It is based on actual observations or experiments using quantitative research methods and it may generate numerical data between two or more.
- Non empirical researches are not based on direct quantitative observation or experimentation. These are primarily based upon qualitative approaches and some degree of subjective investigations.

According to Hussey and Hussey (1997:10), "four different types of research purpose exist: exploratory, descriptive, analytical or predictive." Whatever the purpose of the research, empirical evidence is required. They define empirical evidence as, "data based on observation or experience."

Qualitative vs Quantitative approaches

Qualitative research focuses on qualitative representation of data and its interpretation. In contrast to qualitative research the quantitative research focuses on quantification of data and its interpretation in the research.

Myers (1997) distinguished between qualitative and quantitative research methods. According to him quantitative research methods were originally developed in the natural sciences to study natural phenomena and some examples of quantitative research are

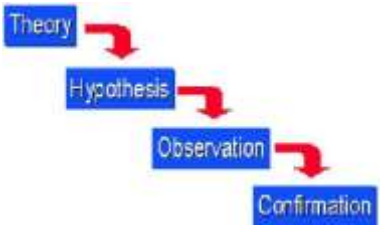
- **Survey research**
- **Laboratory experiments**
- **Formal methods**
- **Numerical methods such as mathematical modeling.**

Qualitative research methods were developed in the social sciences to enable researchers to study social and cultural phenomena and some examples of

- **Action research**
- **Case study research and ethnography**

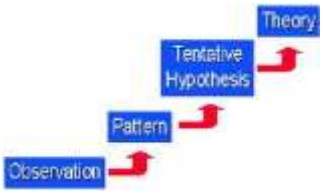
Qualitative data sources include observation and participant observation (fieldwork), interviews and questionnaires, documents and texts, and the researcher's impressions and reactions (Myers, 1997). This research would seek to understand, "people and the social and cultural contexts within which they live," (Myers, 1997). According to Hussey and Hussey's views (1997) qualitative research is "a subjective approach which includes examining and reflecting on perceptions in order to gain understanding of social and human activities."

Deductive vs Inductive approaches



Hussey and Hussey (1997) defined deductive research as "a study in which a conceptual and theoretical structure is developed which is then tested by empirical observation; thus particular instances are deducted from general influences." Deductive research is a study in which theory is tested by empirical observation. The deductive method is referred to as moving from the general to the particular.

Inductive research is a study in which theory is, "developed from the observation of empirical reality; thus general inferences are induced from particular instances, which is the reverse of the deductive method since it involves moving from individual observation to statements of general patterns or laws," (Hussey and Hussey, 1997). In research both approaches can be used and it defines logic to draw inferences and conclusion." The possibility of using both inductive and deductive approaches in the same case study has also been discussed by Perry (2001). He describes a continuum from pure induction **(theory-building)** to pure deduction (**theory-testing**). He advocates taking a balance between the two, striking the position of what he calls "**theory confirming/disconfirming**" approach.



Check Your progress Q 5- Which of the following logical process to derive conclusion is based on a knowledge of general premise or something known to be true?

- ***a.*** **Inductive Reasoning**
- ***b.*** ***Deductive Reasoning***
- ***c.*** ***Conclusive Reasoning***
- ***d.*** ***Predictive Reasoning***

Subjective vs objective approaches

Another significant choice which exists in the research paradigm to be adopted is the extent to which the researcher is subjective i.e. it involved in or has an influence on the research outcome or objective i.e. distanced from or independent in the execution of data collection and research. Easterby-Smith *et al.* (1991) discussed the **"traditional assumption that in science the researcher must maintain completes independence if there is to be any validity in the results produced**." The phenomenological research paradigm is, by its very nature, subjective. The use of this paradigm necessarily requires involvement in both real world circumstances as well as the involvement of the researcher himself. It is now accepted that such a subjective approach, as used in the research, requires the recognition of any influence or limitation such subjectivity may have on the conduct or findings of the research.

19.12: RESEARCH TYPES

As there are many ways of classifying research. However, studying the various characteristics of the different types of research helps us to identify and examine the similarities and differences. Research can be classified according to the:

- **Purpose of the research** – The reason why it was conducted
- **Process of the research** – The way in which the data were collected and analysed
- **Logic of the research** – Whether the research logic moves from the general to the specific or vice versa
- **Outcome of the research** – Whether the expected outcome is the solution to a particular problem or a more general contribution to knowledge. For example, the aim of your research project might be to describe a particular business activity (purpose)

  Table below shows the classification of the main types of research according to the above criteria.

Type of research	Basis of classification
Exploratory, descriptive, analytical or predictive research	Purpose of the research
Quantitative or qualitative research	Process of the research
Applied or basic research	Outcome of the research
Deductive or inductive research	Logic of the research

Type of research	Example
Exploratory	An interview survey among clerical staff in a particular office, department, company, group of companies, industry, region and so on, to find out what motivates them to increase their productivity (that is, to see if a research problem can be formulated).
Descriptive	A description of how the selected clerical staff are rewarded and what measures are used to record their productivity levels.
Analytical	An analysis of any relationships between the rewards given to the clerical staff and their productivity levels.
Predictive	A forecast of which variable(s) should be changed in order to bring about a change in the productivity levels of clerical staff.

After these analysis we can now discuss the characteristics of different types of research as follow

(i) Descriptive vs. Analytical: Descriptive research includes surveys and fact-finding enquiries of different kinds. The major purpose of descriptive research is description of the state of affairs as it exists at present. In social science and business research we quite often use the term Ex post facto research for descriptive research studies. The main characteristic of this method is that the researcher has no control over the variables; he can only report what has happened or what is happening. Most ex post facto research projects are used for descriptive studies in which the researcher seeks to measure such items as, for example, frequency of shopping, preferences of people, or similar data. Ex post facto studies also include attempts by researchers to discover causes even when they cannot control the variables. The methods of research utilized in descriptive research are survey methods of all kinds, including comparative and correlational methods. In analytical research, on the other hand, the researcher has to use facts or information already available, and analyze these to make a critical evaluation of the material.

(ii) Applied vs. Fundamental: Research can either be applied (or action) research or fundamental (to basic or pure) research. Applied research aims at finding a solution for an immediate problem facing a society or an industrial/business organisation, whereas fundamental research is mainly concerned with generalisations and with the formulation of a theory. Research concerning some natural phenomenon or relating to pure mathematics are examples of fundamental research. Similarly, research studies, concerning human behaviour carried on with a view to make **generalisations** about human behaviour, are also examples of fundamental research, but research aimed at certain conclusions facing a concrete social or business problem is an example of applied research. Thus, the central aim of applied research is to discover a solution for some pressing practical problem, whereas basic research is directed towards finding information that has a broad base of applications and thus, adds to the already existing organized body of scientific knowledge

(iii) Quantitative vs. Qualitative: Quantitative research is based on the measurement of quantity or amount. It is applicable to phenomena that can be expressed in terms of quantity. Qualitative research, on the other hand, is concerned with qualitative phenomenon, i.e., phenomena relating to or involving quality or kind. For instance, when we are interested in investigating the reasons for human behaviour (i.e., why people think or do certain things), we quite often talk of '**Motivation Research'**, an important type of qualitative research. This type of research aims at discovering the underlying motives and desires, using in depth interviews for the purpose. Attitude or opinion research i.e., research designed to find out how people feel or what they think about a particular subject or institution is also qualitative research. Qualitative research is specially important in the behavioural sciences where the aim is to discover the underlying motives of human behaviour. Through such research we can analyse the various factors which motivate people to behave in a particular manner or which make people like or dislike a particular thing.

(iv) Conceptual vs. Empirical: Conceptual research is that related to some abstract idea(s) or theory. It is generally used by philosophers and thinkers to develop new concepts or to reinterpret existing ones. On the other hand, empirical research relies on experience or observation alone, often without due regard for system and theory. It is data-based research, coming up with conclusions which are capable of being verified by observation or experiment. We can also call it as experimental type of research. In such a research it is necessary to get at facts firsthand, at their source, and actively to go about doing certain things to stimulate the production of desired

information. In such a research, the researcher must first provide himself with a working hypothesis or guess as to the probable results. He then works to get enough facts (data) to prove or disprove his hypothesis. He then sets up experimental designs which he thinks will manipulate the persons or the materials concerned so as to bring forth the desired information. Such research is thus characterized by the experimenter's control over the variables under study and his deliberate manipulation of one of them to study its effects. Empirical research is appropriate when proof is sought that certain variables affect other variables in some way. Evidence gathered through experiments or empirical studies is today considered to be the most powerful support possible for a given hypothesis

Besides these there are other types of specific researches, based on either the purpose of research, or the time required to accomplish research, on the environment in which research is done, or on the basis of some other similar factor. Form the point of view of time, we can think of research either as **one-time research or longitudinal research**. In the former case the research is confined to a single time-period, whereas in the latter case the research is carried on over several time-periods. Research can be **field-setting research** or **laboratory research** or **simulation research**, depending upon the environment in which it is to be carried out. Research can as well be understood as **clinical or diagnostic research**. Such research follows case-study methods or indepth approaches to reach the basic causal relations. Such studies usually go deep into the causes of things or events that interest us, using very small samples and very deep probing data gathering devices. The research may be exploratory or it may be formalized. The objective of exploratory research is the development of hypotheses rather than their testing, whereas formalized research studies are those with substantial structure and with specific hypotheses to be tested. **Historical research** is that which utilizes historical sources like documents, remains, etc. to study events or ideas of the past, including the philosophy of persons and groups at any remote point of time. Research can also be classified as **conclusion-oriented and decision-oriented.** While doing conclusion oriented research, a researcher is free to pick up a problem, redesign the enquiry as he proceeds and is prepared to conceptualize as he wishes. Decision-oriented research is always for the need of a decision maker and the researcher in this case is not free to embark upon research according to his own inclination. Operations research is an example of decision oriented research

since it is a scientific method of providing executive departments with a quantitative basis for decisions regarding operations under their control.

19.13: SCIENTIFIC THINKING AND LANGUAGE OF RESEARCH

According to Schieldler & Cooper, research involves a lot of reasoning and a researcher must develop a habit of sound reasoning for finding correct premises, testing the connection between facts and assumptions and making claims based on adequate evidences. Scientific thinking refers to the thought processes that are used in science, including the cognitive processes involved in theory generation, experiment design, hypothesis testing, data interpretation, and scientific discovery. Many of these aspects of scientific thinking involve cognitive processes that have been investigated in their own right, such as induction, deduction, analogy, expertise, and problem solving. In research scientific thinking is extensively required to maintain the quality and objectivity in the research. Next section discusses on the specific terms associated scientific thinking in the research.

When we do research, we seek to know what is in order to understand, explain, and predict phenomena. The phenomena are to be defined and explained. In this regard various terms are used.

Concept- A concept is a generally accepted collection of meanings or characteristics associated with certain events, objects, conditions, situations, and behaviors. Classifying and categorizing objects or events that have common characteristics beyond any single observation creates concepts. We use numerous concepts daily in our thinking, conversing, and other activities. In research, special problems grow out of the need for concept precision and inventiveness. We design hypotheses using concepts. We devise measurement concepts by which to test these hypothetical statements. We gather data using these measurement concepts. The success of research hinges on

(1) How clearly we conceptualize and

(2) How well others understand the concepts we use.

For example, when we survey people on the question of customer loyalty, the questions we use need to tap faithfully the attitudes of the participants. Attitudes are abstract, yet we must attempt to measure those using carefully selected concepts. The challenge is to develop concepts that others will clearly understand. We might, for example, ask participants for an estimate of their family's total income. This may seem to be a simple, unambiguous concept, but we will receive varying and confusing answers unless we restrict or narrow the concept by specifying:

- Time period, such as weekly, monthly, or annually.
- Before or after income taxes.
- For head of family only or for all family members.
- For salary and wages only or also for dividends, interest, and capital gains.
- Income in kind, such as free rent, employee discounts, or food stamps

Constructs- Concepts have progressive levels of abstraction, that is, the degree to which the concept does or does not have something objective to refer to. An abstraction like personality is much more difficult to visualize. Such abstract concepts are often called constructs. A construct is an image or abstract idea specifically invented for a given research and/or theory-building purpose. We build constructs by combining the simpler, more concrete concepts, especially when the idea or image we intend to convey is not subject to direct observation. Concepts and constructs are easily confused.

Definitions- Confusion about the meaning of concepts can destroy a research study's value without the researcher or client even knowing it. If words have different meanings to the stakeholders involved, then the parties are not communicating well. Definitions are one way to reduce this danger. Researchers struggle with two types of definitions: dictionary definition and operational definitions. In the more familiar dictionary definition, a concept is defined with a synonym. An operational definition is a definition stated in terms of specific criteria for testing or measurement. These terms must refer to empirical standards (i.e., we must be able to count, measure, or in some other way gather the information through our senses). Whether the object to be defined is physical (e.g., a can of soup) or highly abstract (e.g., achievement motivation), the definition must specify the characteristics and how they are to be observed. The specifications and

procedures must be so clear that any competent person using them would classify the object in the same way. Operational definitions may vary, depending on your purpose and the way you choose to measure them. Here are two different situations requiring different definitions of the same concepts:

Variables- In practice, the term variable is used as a synonym for construct, or the property being studied. In this context, a variable is a symbol of an event, act, characteristic, trait, or attribute that can be measured and to which we assign categorical values. There are various types of variables broadly classified as dependent and independent. Schiendler and coopers described the characteristics of these as follow.

Independent Variable	Dependent Variable
Predictor	Criterion
Presumed cause	Presumed effect
Stimulus	Response
Predicted from…	Predicted to…
Antecedent	Consequence
Manipulated	Measured outcome

Many textbooks use the term predictor variable as a synonym for **independent variable (IV).** This variable is manipulated by the researcher, and the manipulation causes an effect on the dependent variable. We recognize that there are often several independent variables and that they are probably at least somewhat "correlated" and therefore not independent among themselves. Similarly, the term criterion variables used synonymously with **dependent variable (DV).** This variable is measured, predicted, or otherwise monitored and is expected to be affected by manipulation of an independent variable.

Proposition and Hypothesis- We define a proposition as a statement about observable phenomena (concepts) that may be judged as true or false. When a proposition is formulated for empirical testing, we call it a hypothesis. As a declarative statement about the relationship between two or more variables, a hypothesis is of a tentative and conjectural nature. Hypotheses have also been described as statements in which we assign variables to cases. A case is defined in this sense as

the entity or thing the hypothesis talks about. The variable is the characteristic, trait, or attributes that, in the hypothesis, is imputed to the case.

Theory: A theory is a set of systematically interrelated concepts, definitions, and propositions that are advanced to explain and predict phenomena (facts). In this sense, we have many theories and use them continually to explain or predict what goes on around us. To the degree that our theories are sound and fit the situation, we are successful in our explanations and predictions. In marketing, the product life cycle describes the stages that a product category goes through in the marketplace.

Model: A Research Model refers to the design of research to be tested and assumptions of complex relationships among the variables which are to be investigated. The term model is used as representation of a system i.e. constructed to study some aspect of the system or the system as a whole. The term model is also used in business research and other fields of business to represent phenomena through the use of analogy. A model is defined here as a representation of a system that is constructed to study some aspect of that system or the system as a whole. Models differ from theories in that a theory's role is explanation whereas a model's role is representation. Model defines the relationships among variables which are to be tested and it may be used for applied or highly theoretical research.

For a business research following types of research models are suggested :-

(i) Descriptive model

(ii) Explicative model

(iii) Simulation model

- Static
- Dynamic

These classifications are based on functions of models.

- Descriptive model describes the behavior of elements in a system where theory is inadequate and non-existent

- Explicative model extent the application of well developed theories or improve our understanding of their key concepts.
- Simulation model clarify the structural relations of concepts and attempt to reveal the processes and relationships among them.

The term static and dynamic defines behavior of systems with time. If the research model is concerned with the study and representation of system at a time, it is called static model. If the research model is defining the research study for the system over a period of time, it is called as dynamic system.

19.14: CHARACTERISTICS OF A GOOD RESEARCH

The eight most widely agreed upon characteristics of research are as follows:

- Systematic procedures
- Controlled procedures,
- Validity,
- Rigorousness,
- Logicality,
- Critical thought,
- Objectivity and
- Accuracy.

Some important terms defining the characteristics are explained here-

- Systematic: The research should use valid procedures and principles.
- Reproducibility: The design should be valid with clear procedures so that others can test the findings
- Controlled: It refers to how variables are manipulated and controlled

- Empirical and objective: It should be based on primary findings and direct observations
- Analytical and critical: Refers to valid logic and reasoning
- Accuracy: The findings must be valid and data must be accurate without any manipulations
- Originality: Contribute significantly by innovative and new thoughts.

19.15: PROBLEMS WITH RESEARCH IN INDIA

With the time the research become applied in nature as the developed theories and construct are helpful in explaining the existing phenomena. Now a day, the quality of research is going down as most of the researches are being deviated from the criteria of a good research. This is leading to problems with the research in current scenario. Some problems which are associated with the research are listed here.

- Accuracy and Reliability of data are questionable.
- Biasness of the researcher.
- Scientific collection of data is questionable.
- Misinterpretation of data
- Time consuming
- Not in the interest of the person
- Lack of clarity of data
- Lack of valid and objective conclusion
- Lack of knowledge
- High cost
- Quantity of data
- Improper sampling and its size

- Source of data is questionable
- Lack of information

Above problems re leading to substandard research and compromise in quality of research. With the advent of internet and communication technology theft of research data (conducted by other researcher) is a major problem of research leading to ethical issues.

19.16: SUMMARY

In this unit we have discussed about fundamental concept of research and research methodology. Various terms associated with research methodology are discussed in this unit so that the students can develop a solid foundation is self paced learning. Research has been defined as a process of systematic data collection and analysis to get new insight into the problem

19.17: GLOSSARY

Ontology: In general, ontology refers to the study or concern about what kinds of things exist within society

Paradigm: It refers to the school of thought which are followed to carry out the research

Epistemology: Epistemology deals with the issue of knowledge, and specifically, who can be a 'knower'.

Concept- A concept is a generally accepted collection of meanings or characteristics associated with certain events, objects, conditions, situations, and behaviors

Theory: A theory is a set of systematically interrelated concepts

Model: A Research Model refers to the design of research to be tested and assumptions of complex relationships among the variables which are to be investigated.

19.18 CHECK YOUR PROGRESS

Q1: Doing research requires drafting a working outline, which is

a. Having a predefined and clear-cut objective(s).
b. Planning to get answers for what, why & where questions.
c. Having a clear idea about the research problem solution.
d. None of the above

Q2: The main objectives behind doing research are to

a. Study and explore knowledge.
b. Get new ideas.
c. Define clear objectives.
d. All the above

- Q3:In order to begin research, one must

a. Start with a number of clear goals.
b. Start with a number of predefined objectives.
c. Have a well defined research method.
d. Solve the research problem

Q4:- Research philosophies refers to

a. Approaches and the discipline.
b. Correct procedures in the discipline.
c. Ideas to discover.
d. Objectives to consider in the research process.

Q 5- Which of the following logical process to derive conclusion is based on a knowledge of general premise or something known to be true?

a. Inductive Reasoning
b. Deductive Reasoning
c. Conclusive Reasoning
d. Predictive Reasoning

Q6: Which of the following research leads to new insight into the existing knowledge and not based on specific situations of managerial problem?

a. Pure Research
b. Applied Research

c. Field Research

d. Historical Research

Q7: In research, the variable which leads to change in other variable is called as

a. Dependent variable

b. Independent variable

c. Moderating Variable

d. Intervening Variable

Q8- Descriptive research studies is a category of research that aims to

a. Achieve new insights of a concept.
b. Analyze characteristics of something.
c. Determine the frequency with which something occurs.
d. Test the relationship between variables

19.19: ANSWERS TO CHECK YOUR PROGRESS

Q1-b, Q2-d, Q3-b, Q4-b, Q5-b, Q6- b, Q7- b, Q8-b

19.20: TERMINAL QUESTIONS

Q1. Define the following terms:

i. Research

ii. Research Problems

iii. Research Methods

iv. Research Methodology

v. Research Paradigm

vi. Research Process

vii. Research Variables

viii. Research Design

Q2.Explain various objectives of a research?

Q3.Distinguish between Research Methods and Research Methodology.

Q4.What are the characteristics of a research? Explain its significance in modern times?

Q5.Discuss the factors motivating research.

Q6.Explain the principles of a good research

Q7.What is the scope of research in the present context of opening of national economy and globalization of markets?

Q8.Explain the problem and limitations faced in conduct of research.

Q9. What do you mean by Model in a research? What are the types of research model?

Q10. What do you mean by Exploratory and Descriptive Research?

Q11.What do you mean by scientific thinking? Why scientific thinking is required for a good research?

Q12. What do you mean by deductive and inductive reasoning? What are the importances of these concepts in research?

19.21: SUGGESTED READINGS

1.Kothari, C R (2008), Research Methodology & techniques, New Age Publication, Delhi

2.Bajpai, Naval (2012) Business Research Methods, Pearson, Delhi

3.Schiendler & Cooper(2009) : Business Research Methods, TMH, New Delhi

4.Marketing Research : N K Malhotra , Pearsons Education Asia publication

BLOCK 6: RESEARCH METHODOLOGY

UNIT 20: RESEARCH DESIGN

STRUCTURE

OBJECTIVES

After reading this unit you would be able to understand

- What is a research design?
- What are the steps in research design?
- What are the types of research design?
- What are the criteria of a good research design?

20.1: INTRODUCTION

In the previous unit, you have learned the meaning and concepts associated with research. In this section, we will discuss the specific approaches for designing research. As the research has different interpretations and connotations for different people, the work of researchers occurs through research methods, selected by the researcher in the research design.

In the most elementary sense, the research design is a logical sequence of activities related to the research process. It starts with the conceptualization of the research problem and ends with findings and conclusions. Yin (1994) specified that a research design is a blueprint of research, dealing with the following.

- What questions to study?
- What data are relevant?
- What data to collect?
- How to analyze the results?

Research design is much more than a work plan because the main purpose is to help to avoid the situation in which the evidence does not address the initial research questions. Hence the research design deals with a logical problem and also specifies how the investigator will address the critical issues. Designs in research describe the data collection method or research approach that is used in a study. It defines the various ways by which information is gathered for evaluation or

assessment. There are various designs in research, and each is used for a different purpose. It is not uncommon to use a mix of two or more research designs in certain studies.

20.2: DEFINING RESEARCH DESIGN

There are many definitions of research design, but no one definition covers the full range of important aspects. Some definitions for research design are quoted here.

According to David J Luck and Ronald S Rubin, "A research design is the determination and statement of the general research approach or strategy adopted for the particular research project".

Kerlinger defines research design as "the plan, structure, and strategy of investigation conceived to obtain answers to research questions and to control variance".

According to Green and Tull, "A research design is the specification of methods and procedures for acquiring the information needed".

Cooper & Schindler say that "the research design constitutes the blueprint for the collection, measurement, and analysis of data."

Green defines research design as "the specification of methods and procedures for acquiring the information needed. It is the overall operational pattern of a framework of the project that stipulates what information is to be collected from which sources by what procedures".

Research design could be described as the plan and structure of investigation so conceived as to obtain answers to research questions. The plan is the overall scheme or program of the research. It includes an outline of what the investigator will do from writing hypotheses and their operational implications to the final analysis of data. A structure is the framework to organize the research process and configurations of the research. A good research design will ensure that the information obtained is relevant to the research questions and that it was collected by objective and economical procedure. Specifically a research design could be described as the overall plan for connecting the conceptual research problems to the pertinent (and achievable) empirical research. It articulates what data is required, what methods are going to be used to collect and analyse this data, and how

all of this is going to answer your research question. Different design logics are used for different types of study.

The research design also reflects the **purpose of the inquiry**, which can be characterized as Exploration, Description, Explanation, Prediction, Evaluation, and History. The table below summarizes the types of questions addressed in different groups.

Question type	Question	Examples
Exploratory questions	What is the case? What are the key factors?	What are the critical success factors of a profitable company? What are the distinguishing features of a good leader? What are the reasons for the carnage on South African roads?
Descriptive questions	How many? What is the incidence of x? Are x and y related?	How many people died of AIDS in South Africa last year? Is there a correlation between parental support and scholastic achievement?
Causal questions	Why? What are the causes of y?	What are the main causes of malnutrition in a rural community? Is smoking the main cause of lung cancer?
Evaluative questions	What was the outcome of x? Has P been successful?	Has the new TB awareness programme produced a decline in reportable TB cases? Has the introduction of a new refrigeration technology led to more cost-effective production?
Predictive questions	What will the effect of x be on y?	What effect will the introduction of a new antibiotic have on population P?
Historical questions	What led to y happening? What were the events that led up to y? What caused y?	What caused the demise of socialism in Central Europe during the late eighties? What led NATO countries to decide to start aerial bombing of Kosovo?

Practice Exercise: Visit the library of your university and access the research journals available in the library. Read the research articles which are published in that journal. Try to understand the research design and research methodology followed by author(s) in that article.

20.3: CHARACTERISTICS OF THE RESEARCH DESIGN

After analyzing the definitions and explanation of the research design, we can now summarize the characteristics of the research design as follows:

- Research design is a set of interconnected and sequential activities and time time-based plan.
- Research design depends upon the research question and research framework.
- Research design guides the selection of the source of information for research.
- Research design specifies the relationship of the variables to be studied.
- Research design outlines and specifies the procedures to be followed in various research activities
- Research design also specifies the tools and instruments to be used in the research.

According to Cooper & Schindler (2009) a research design is intended to answers for various questions such as

- What techniques will be used to gather data?
- What method of analysis has to be followed?
- What kind of sampling will be used?
- How will time and cost constraints be dealt with?

20.4: RESEARCH DESIGN AND RESEARCH METHODOLOGY

There are basic differences between research method and research methodology, and research design. At this stage you must understand the difference between these two concepts. Schiendler and Cooper have compared the research design research methodology as follow.

Research design	Research methodology
Focuses on the **end-product**: What kind of study is being planned and what kind of results are aimed at. E.g. Historical - comparative study, interpretive approach OR exploratory study, inductive and deductive etc.	Focuses on the **research process** and the kind of tools and procedures to be used. E.g. Document analysis, survey methods, analysis of existing (secondary) data/statistics etc)
Point of departure (driven by) = Research problem or question.	Point of departure (driven by) = Specific tasks (data collection or sampling) at hand.
Focuses on the logic of research: What evidence is required to address the question adequately?	Focuses on the individual (not linear) steps in the research process and the most 'objective' (unbiased) procedures to be employed.

20.5: CRITERIA FOR A GOOD RESEARCH DESIGN

According to Chawla & Sondhi(2011) a research design must endure following basic tenets.

- Able to convert the research question and the stated assumptions/hypothesis into operational variables that can be measured.
- Must specify the process that would be followed to complete the above task, as efficiently and economically as possible.
- Specify the control mechanism that would be used to ensure that the effect of other variables that could not impact the outcome of the study has been controlled.

Based on the above discussions, we can now summarise the criteria for a good research design as follows.

- **Simplicity**: A research design should be simple and understandable.
- **Economical**: Research design must be economical. The technique selected must be cost-effective and less time-consuming.

- **Reliability**: A good research design must ensure to reduce the possibilities of various errors. This should have the minimum bias and have the reliability of data collected and analysed.
- **Workability:** A good research design must be workable, pragmatic, and practicable.
- **Flexibility**: A good research design must be flexible enough to accommodate the consideration of many different aspects of the research problem and phenomena that may appear during the research process.
- **Accuracy**: A good research design must lead to accurate and objective results to draw valid conclusions.

20.6: STEPS TO BE FOLLOWED IN RESEARCH DESIGN

The steps for formulating the research design are cyclic processes. These processes are interrelated in such a way that the feedback of one is required for correcting the previous one. The research design formulation sometimes used interchangeably with designing the research processes, which is not true. Research design for each research process is unique and it requires concrete focus of specifications of the tasks to be carried out in term of why, how and when.

The research design formulation starts with specifying the objectives and scope of the research. The researcher has to select the mode of inquiry that defines what kind of knowledge is possible and legitimate. A large number of different terms have been used to refer to designing in research and these terms are often used synonymously as methodologies, approaches, perspectives, and philosophies as if they are all comparable , but they are different.

The mode of inquiry and investigation of research problem defines epistemological position of a researcher. **Epistemology** deals with the basic issue on knowledge exploration as "how knowledge is derived and it should be tested and validated". There are three suggested modes of inquiry in research which guides the research design.

1. **Positivistic** – In this mode the empirical and scientific investigation is carried out. It requires control and statistical mode of analysis

2. **Constructivists**- It emphasizes on qualitative mode of investigation and argues that it is the best choice for research in social science as compared to quantitative method
3. **Triangulation** – It suggests simultaneous and sequential uses of qualitative and quantitative methods of investigation

After selection of epistemological position from any one of the above stated approach, the research questions are subjected to thorough and comprehensive theoretical review. It helps the researcher to develop a practical and manageable perspective for the research question and method of investigation.

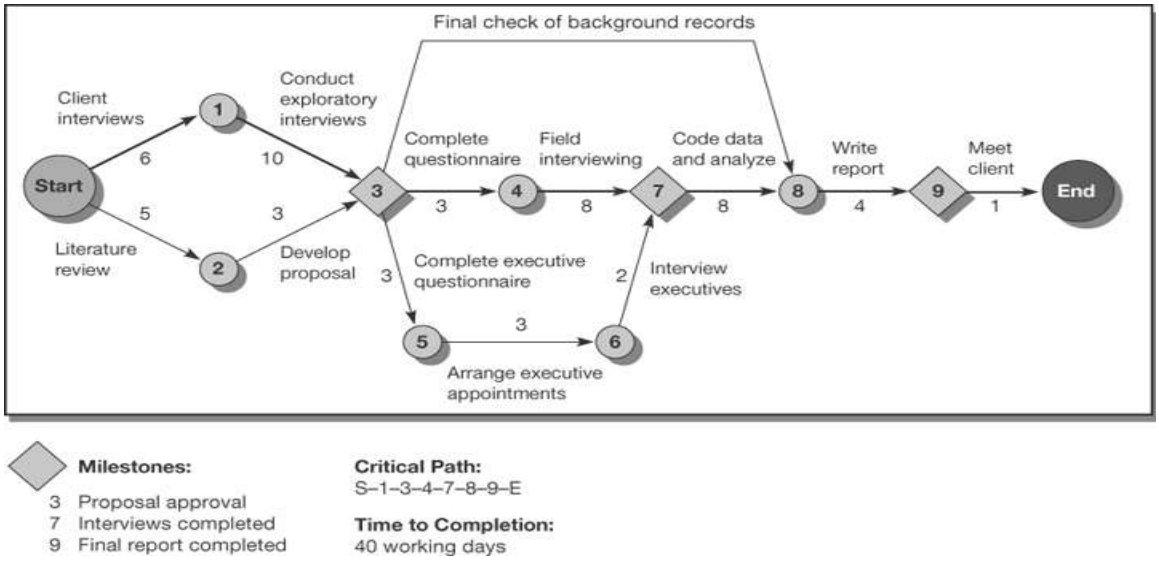
After this step a conceptual research design is developed by the researcher in which variables of study are identified and mode of study and analysis are selected.

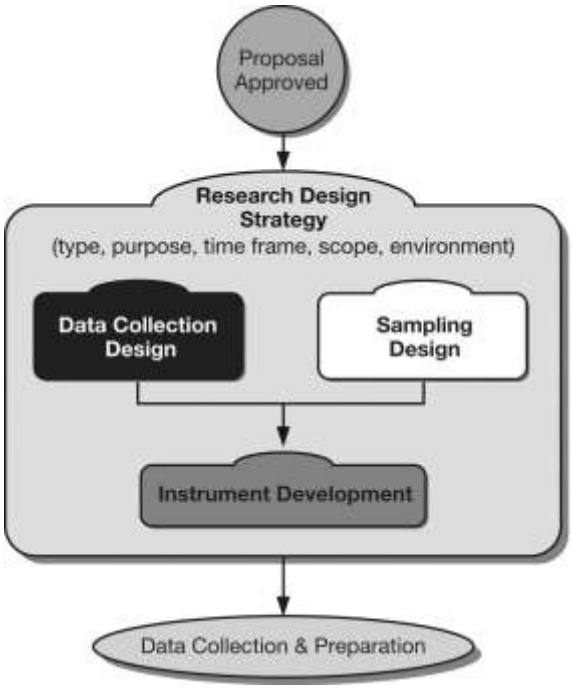
After above discussions now you are able to understand about the aspects of research design. More specifically a research design consists of following generic steps.

1. **Selection and Definition of a problem**: The problem selected for study should be defined clearly in operational terms so that researcher knows positively what facts he is looking for and hat is relevant to the study.
2. **Source of Data:** Once the problem is selected it is the duty of the researcher to state clearly the various sources of information such as library, personal documents, field work, a particular residential group etc.
3. **Nature of Study:** The research design should be expressed in relation to the nature of study to be undertaken. The choice of the statistical, experimental or comparative type of study should be made at this stage so that the following steps in planning may have relevance to the proposed problem.
4. **Object of Study**: Whether the design aims at theoretical understanding or presupposes a welfare notion must be explicit at this point. Stating the object of the study helps not only in clarity of the design but also in a sincere response from the respondents.
5. **Social-Cultural Context:** The research design must be set in the social-cultural context. For example in a study of the fertility rate in a people of „backward" class the context of the so-called backward class of people and the conceptual reference must be made clear.

Unless the meaning of the term is clearly defined there tends to be a large variation in the study because the term backward could have religious, economic and political connotations.

6. **Temporal context:** The geographical limit of the design should also be referred to at this stage that research related to be hypothesis is applicable to particular social group only.
7. **Dimension:** It is physically impossible to analyze the data collected from a large universe. Hence the selection of an adequate and representative sample is a by-word in any research.
8. **Basis of Selection:** The mechanics of drawing a random, stratified, and purposive, double cluster or quota sample when followed carefully with produce a scientifically valid sample in an unbiased manner
9. **Technique of Data Collection**: relevant to the study design a suitable technique has to be adopted for the collection of required data. The relative merit of observation, interview and questionnaire, when studied together will help in the choice of suitable technique. Once the collecting of data is complete, analysis, coding and presentation of the report naturally follow. Schieldler and Cooper presented the steps of research design in following models





Practice Exercise: ***Read the performance report of any company published in newspaper of available on the website. Understand the mode of data collection, its analysis and presentation to get insight as how this has been compiled?***

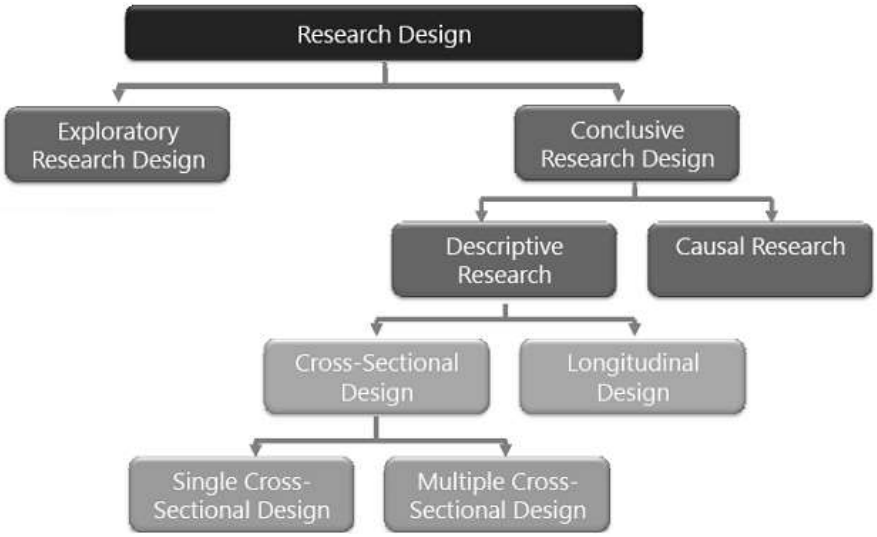
20.7: CRITERIA FOR CLASSIFYING RESEARCH DESIGNS

Various types of research design are classified on the basis of following criteria.

- **The degree to which the research question has been crystallized**- On this basis we have Exploratory study and Formal study
- **The method of data collection**- On the basis of this criteria the research designs are classified as Monitoring and Communication Study
- **The power of the researcher to produce effects in the variables under study**- Under this category the research design are classified as Experimental and Ex post facto
- **The purpose of the study**- This basis leads to categories the research design as Reporting, Descriptive, Causal-Explanatory, Causal-Predictive

- **The time dimension**- On time dimension the research designs are classified as Cross-sectional and Longitudinal
- **The topical scope, breadth and depth of the study-** On the basis of this we can classify the research design as Case study and Statistical study
- **The research environment**- The research environment basis of classification defines research design as Field setting, Laboratory research, Simulation
- **The participants' perceptional awareness of the research activity**- It classifies the research designs as Actual routine research design and Modified routine research design.

Malhotra in his book marketing research developed the relationships among different types of research design Figure on right side depicts the classification and relationships of various research design



20.8 TYPES OF RESEARCH DESIGN

After previous discussion, we can conclude that there are various ways by which research processes could be designed and perceived. Even after so many types of research design, most of the books prefer to discuss the research design by classifying them into following there types

1. Exploratory research design
2. Descriptive research design
3. Causal research design

We will have detail discussions on various these types of research design now.

20.8.1: EXPLORATORY RESEARCH DESIGN

Exploratory research is unstructured, informal research undertaken to gain background information about the general nature of the research problem. This type of research design is specifically suitable when the research objective is to provide insights into

- Identifying the problems or opportunities
- Defining the problem more precisely,
- Gaining deeper insights into the variables operating in a situation
- Identifying relevant courses of action
- Establishing priorities regarding the potential significance of a problems or opportunities
- Gaining additional insights before an approach can be developed and
- Gathering information on the problems associated with doing conclusive research

Much research has been of an exploratory nature; emphasizing on finding practices or policies that needed changing and on developing possible alternatives. Exploratory research could also be used in conjunction with other research. As mentioned below, since it is used as a first step in the

research process, defining the problem, other designs will be used later as steps to solve the problem. For instance, it could be used in situations when a firm finds the going gets tough in terms of sales volume, the researcher may develop use exploratory research to develop probable explanations. Analysis of data generated using exploratory research.

On examination of the objectives of exploratory research, it is well understood that it could be used at the initial stages of the decision-making process. It allows the marketer to gain a greater understanding of something that the researcher doesn‟t know enough about. This helps the decision maker and the researcher in situations when they have inadequate knowledge of the problem situation and/or alternative courses of action. In short, exploratory research is used in the absence of tried models and definite concepts.

Exploratory research could also be used in conjunction with another research. As mentioned below, since it is used as a first step in the research process, defining the problem, other designs will be used later as steps to solve the problem. For instance, it could be used in situations when a firm finds the going gets tough in terms of sales volume, the researcher may develop use exploratory research to develop probable explanations. Analysis of data generated using exploratory research is essentially abstraction and generalization. Abstraction refers to the translation of the empirical observations, measurements etc. into concepts; generalization means arranging the material so that it focuses on those structures that are common to all or most of the cases. The exploratory research design is best characterized by its flexibility and versatility. This is so because of the absence of the non-imperativeness of a structure in its design. It predominantly involves the imagination, creativity, and ingenuity of the researcher.

Exploratory research are conducted by following methods

- **Secondary data analysis:** Secondary data refers to the process of searching for and interpreting existing info relevant to the research problem (e.g., census data, articles in journals, newspapers, etc.).
- **Experience (Expert) surveys:** Refers to gathering info from those thought to be knowledgeable on the issues relevant to the problem (i.e., ask experts).
- **Case Analysis:** Uses past situations that are similar to the present research problem.

- **Focus groups**: Involves small (8-12) groups of people brought together and guided by a moderator through unstructured, spontaneous discussion.

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.

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. For example, a children's book publisher gained valuable information about a sales decline by talking with librarians and schoolteachers who indicated that more and more people were using library facilities and presumably buying fewer books for their children

Focus group interviews are among the most commonly 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.

Case Analyses is Intensive study of selected examples of the phenomenon of interest 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 fi eld 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 fi eld with the sales force, this insight probably wouldn't have been uncovered.

Practically, exploratory research design could be used for the following purposes.(Malhotra, 2001)

- Formulate a problem more precisely
- Identify an alternative course of action
- Develop a hypothesis
- Isolate key variables and relationships for further examination
- Gain insights for developing an approach to the problem
- Establish priorities for further research

20.8.2: DESCRIPTIVE RESEARCH DESIGN

Descriptive research provides answers to the questions of who, what, when, where, and how. It is important to note here that we cannot conclusively ascertain answers to WHY using descriptive studies.

As the name suggests, descriptive study involves describing some event or phenomena on investigation and study under research. For example, a descriptive research design for market conditions may consider the following situations for descriptions under who, what, when, where, and how.

- Characteristics or functions
- Estimate the percentage of customers in a particular group exhibiting the same purchase behaviour;
- Perceptions of product characteristics; and
- To predict the pattern of behaviour of characteristic versus the other

In descriptive research, the data is collected for a definite purpose and involves analysis and interpretation by the researcher. The major difference between exploratory and descriptive research are as follows

- Descriptive research is characterized by the formulation of specific objectives.

- Descriptive studies restrict flexibility and versatility as compared to exploratory research.
- It involves a higher degree of formal design specifying the methods for selecting the sources of information and for collecting data from those sources.

Formal design is required in order to ensure that the description covers all desired phases. It is also required to restrain the collection of unnecessary data

While designing descriptive research, the researcher should also have sufficient knowledge of the nature and type of statistical techniques he/she is going to use. This will greatly help to have the right design in place. Mostly descriptive studies are conducted using questionnaires, structured interviews and observations. The results of descriptive studies are directly used for managerial decisions

Descriptive research design could be classified into broad categories as follows.

- Cross-sectional descriptive research design
- Longitudinal descriptive research design

Longitudinal research relies on panel data and panel methods. It involves fixing a panel consisting of fixed sample of subjects that are measured repeatedly. The panel members are those who have agreed to provide information at a specific intervals over an extended period. For example, data obtained from panels formed to provide information on market shares are based on an extended period of time, but also allow the researcher to examine changes in market share over time. New members may be included in the panel as an when there is a dropout of the existing members or to maintain representativeness. Panel data is analytical and possess advantages with respect to the information collected in the study. They are also considered to be more accurate than cross sectional data because panel data better handle the problem associated with the errors that arise in reporting past behaviour and the errors that arise because of the necessary interaction between interviewer and respondent. Some examples of descriptive research are as follow

- Study Measuring Various Attributes of Salespeople, a Training Program, or a Retailing Situation
- Measuring how salespeople or customers behaved, as well as what happened to sales volume

- Learn about characteristics of people shopping at a particular store
- Satisfaction Study taken at multiple times throughout the year

Cross-sectional research is the most predominantly and frequently used descriptive research design in marketing. It involves a sample of elements from the population of interest. The sample elements are ensured on a number of characteristics. There are two types of cross-sectional studies

- Field studies and
- Surveys

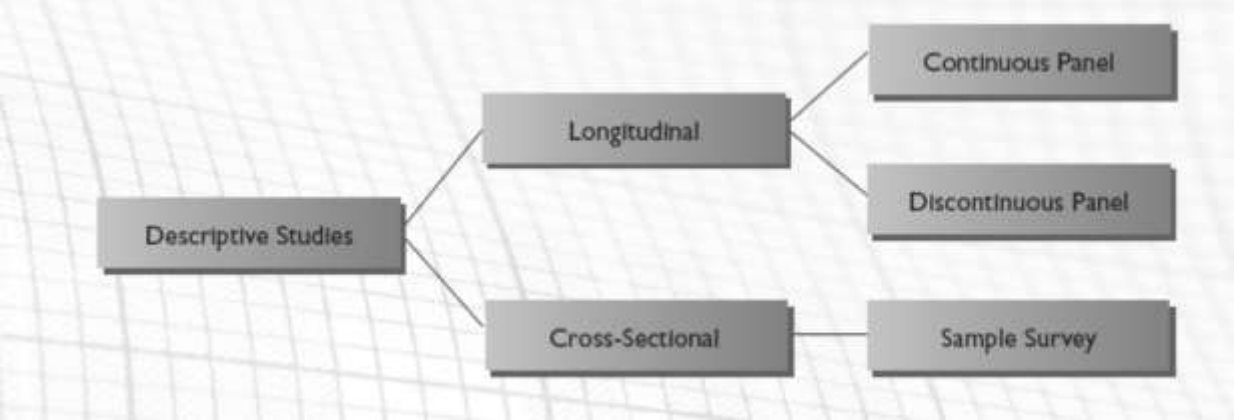
It may appear that field studies and surveys are no different but the same. However, for practical reasons, they are classified into two categories cross sectional research. The fundamental difference lies in the depth of what these research cover. While survey has a larger scope, field study has greater depth. Survey attempts to be representative of some known universe and filed study is less concerned with the generation of large representative samples and is more concerned with the in-depth study of a few typical situations. Cross-sectional design may be either single or multiple cross-sectional design depending on the number of samples drawn from a population. In single cross-sectional design, only one sample of respondents is drawn whereas in multiple cross-sectional designs, there are two or more samples of respondents. A type of multiple cross-sectional design of special interest is Cohort analysis.

Descriptive research does not fit neatly into the definition of either quantitative or qualitative research methodologies, but instead it can utilize elements of both, often within the same study. The term descriptive research refers to the type of research question, design, and data analysis that will be applied to a given topic. Descriptive statistics tell what is, while inferential statistics try to determine cause and effect.

The type of question asked by the researcher will ultimately determine the type of approach necessary to complete an accurate assessment of the topic at hand. Descriptive studies, primarily concerned with finding out "what is," might be applied to investigate the following questions: Descriptive research can be either quantitative or qualitative. It can involve collections of quantitative information that can be tabulated along a continuum in numerical form, such as scores on a test or the number of times a person chooses to use a-certain feature of a multimedia program,

or it can describe categories of information such as gender or patterns of interaction when using technology in a group situation. Descriptive research involves gathering data that describe events and then organizes, tabulates, depicts, and describes the data collection (Glass & Hopkins, 1984). It often uses visual aids such as graphs and charts to aid the reader in understanding the data distribution. Because the human mind cannot extract the full import of a large mass of raw data, descriptive statistics are very important in reducing the data to manageable form. When in-depth, narrative descriptions of small numbers of cases are involved, the research uses description as a tool to organize data into patterns that emerge during analysis. Those patterns aid the mind in comprehending a qualitative study and its implications.

Descriptive research design could be classified as follow



The above exhibit is an overview of various 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.

Practice Exercise: ***Write some research articles under the guidance of faculty members of your university. Try to understand the steps followed in defining and formulating the research topic and data collection. Send these articles for publication also.***

20.8.3: CAUSAL RESEARCH DESIGN

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 the 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. 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. Sometimes we conduct experiments in "fake" or "sterile" environments so that we can carefully control exactly what research participants (called experimental subjects) see and experience. This allows us to

observe the effect of the manipulated variables while the effect of other factors is minimized. The experiments could be conducted as laboratory experiments or field experiments:

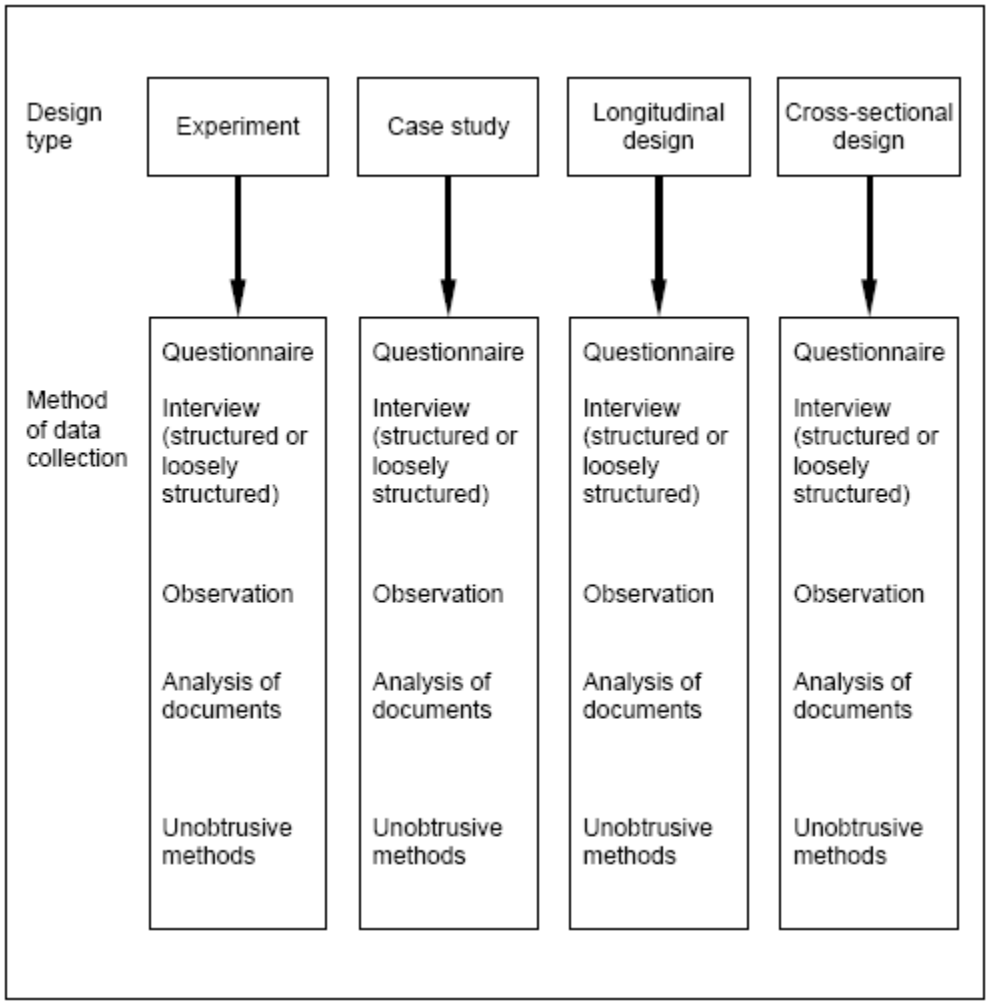
Laboratory experiments allow us to be almost certain that the variables we manipulate produce the outcomes we observe because we can hold all other factors constant.

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

Example: The researchers studying consumer preferences for clustered (versus non-clustered) trip chains also conducted a field experiment. In this case, the experiment was conducted with residents who actually lived in the area that had been mapped for subjects in the lab experiment. For the field study, however, researchers used a telephone survey and based the study on the subjects' home address and actual locations of retailers who were known to the subjects. They asked them to imagine that they needed to make trips to the two kinds of retailers and then presented them with two alternative routes (one that was clustered and one that was non-clustered). As in the laboratory experiment, subjects expressed a preference for the clustered trip chain compared with the non-clustered trip chain, even though the overall travel distance was about the same

Practice Exercise: ***A company is facing a problem as sales of its product are declining. The sales manager of this company is worried about decreasing sales revenue and has called a meeting of sales executives. In this meeting, various reasons for declining sales were explained. The sales manager is not satisfied with the explanations and decided conduct research on this problem. Which type of research design do you suggest? Justify your answer.***

Illustration: Suggested Methods of data collection used in different types of research design



20.9: ERRORS IN A RESEARCH DESIGN

Now we have developed an understanding of the types of research design. Specifically, research design depends upon the formulation of the research problem by the researcher and the selection of the mode of investigation. Even in the broad categorization of various types of research design, it is a subjective process yet to be standardized. A wrong selection of any step may lead to errors in the research design, which may be due to the followings

1. Errors in Population Specification: This type of error occurs when the researcher selects an inappropriate population or universe from which to obtain data.

Example: *Packaged goods manufacturers often conduct surveys of housewives because they are easier to contact, and it is assumed they decide what is to be purchased and also do the actual purchasing. In this situation there often is a population specification error. The husband may purchase a significant share of the packaged goods, and have significant direct and indirect influence over what is bought. For this reason, excluding husbands from samples may yield results targeted to the wrong audience.*

2. Errors in Sampling: Sampling error occurs when a probability sampling method is used to select a sample, but the resulting sample is not representative of the population of concern. Unfortunately, some element of sampling error is unavoidable. This is accounted for in confidence intervals, assuming a probability sampling method is used**.**

Example: *Suppose that we collected a random sample of 500 people from the adult population to gauge their entertainment preferences. Then, upon analysis, found it to be composed of 70% females. This sample would not be representative of the general adult population and would influence the data. The entertainment preferences of females would hold more weight, preventing accurate extrapolation to the general adult population. Sampling error is affected by the homogeneity of the population being studied and sampled from and by the size of the sample.*

3. Errors in Selection: Selection error is the sampling error for a sample selected by a nonprobability method.

Example*: Interviewers conducting a mall intercept study have a natural tendency to select those respondents who are the most accessible and agreeable whenever there is latitude to do so. Such samples often comprise friends and associates who bear some degree of resemblance in characteristics to those of the desired population.*

4. Non-responsive errors: Non-response error can exist when an obtained sample differs from the original selected sample.

Example*: In telephone surveys, some respondents are inaccessible because they are not at home for the initial call or call-backs. Others have moved or are away from home for the period of the survey. Not-at-home respondents are typically younger with no small children, and have a much higher proportion of working wives than households with someone at home. People who have moved or are away for the survey period have a higher geographic mobility than the average of* the population. Thus, most surveys can anticipate errors from non-contact of respondents. Online

surveys seek to avoid this error through e-mail distribution, thus eliminating not-at-home respondents.

5. Measurement errors: Measurement error is generated by the measurement process itself and represents the difference between the information generated and the information wanted by the researcher.

Example: *A retail store would like to assess customer feedback from at-the-counter purchases. The survey is developed, but fails to target those who purchase in the store. Instead, results are skewed by customers who bought items online.*

20.10: TYPES OF DATA AND DATA COLLECTION METHODS

Research data may be collected in various ways. Some of these methods depend on the methodology and the theoretical assumptions used in the research. *" Triangulation"*- a notion introduced from military studies by Denzin (1978) (as quoted by Tomkin & Groves, 1983), has been suggested as a way to make research studies more robust and rigorous by verifying results through different methods, thus ensuring that the results are not a function of the research method.

There are two types of data used in research

1. Primary Data
2. Secondary Data

Data that are observed or collected directly from first-hand experience are called as primary data. Published data and the data collected in the past or other parties is called secondary data. Common sources of secondary data for include censuses, organizational records, and data collected through qualitative methodologies or qualitative research. Primary data, by contrast, are collected by the investigator conducting the research.

Questionnaires and the survey method:

Questionnaires have, according to Sharp & Howard (1996, p 145), "over the past century, become a common method of gathering information." It can be defined as "a pre-formulated written set

of questions to which participants record their answers, usually within largely closely defined alternatives." (Sekaran, 1992, p 200). Survey is a method where questionnaire is used to collect data. Creswell (1994) define a survey as **"the data collection process of asking questions, provides a quantitative or numeric description of some fraction of the population i.e. a sample which can be in turn generalised to the population from which the sample was drawn."**

Interviews:

Nachmias & Nachmias (1996,) define an interview as a **"face-to-face, interpersonal role situation in which an interviewer asks participants questions designed to elicit answers pertinent to the research hypotheses**". However, Sekaran (1992) reminds us that interviews need not be face-to-face, as it can be conducted through the telephone or can even be computer-assisted. Interview could be classified as *structured or unstructured* (or non-directive interview), although Nachmias & Nachmias (1996) identify a third category- *the focused interview*, which is a variation of the structured interview.

In the structured interview, the format is more rigid and assumes that the researcher knows exactly what information is needed and has a list of pre-determined questions he intends to ask of the participants. The same questions are administered to every interviewee, although in certain cases, depending on the circumstances or participants' answers, the researcher may elicit additional information by asking additional questions not on his schedule. "**Through this process, new factors might be identified and a deeper understanding might result**" (Sekaran 1992)

In the ***nonstructured or non-directive interview***, the researcher does not have a schedule listing a set of pre-specified questions, nor are the questions asked in a specific order. The researcher does not direct the interviewee, and thus the interviewee is encouraged to relate his or her experiences and to reveal their attitudes and perceptions on the topic of interest. In this method, the interviewer has an opportunity to probe various areas and to raise specific queries during the interviews.

Observation:

Observation is a way of gathering data by watching behavior, events, or noting physical characteristics in their natural setting. Observations can be overt (everyone knows they are being observed) or covert (no one knows they are being observed, and the observer is concealed). The benefit of covert observation is that people are more likely to behave naturally if they do not know they are being observed. However, you will typically need to conduct overt observations because of ethical problems related to concealing your observation. Observations can also be either direct or indirect. Direct observation is when you watch interactions, processes, or behaviors as they occur; for example, observing a teacher teaching a lesson from a written curriculum to determine whether they are delivering it with fidelity. Indirect observations are when you watch the results of interactions, processes, or behaviors; for example, measuring the amount of plate waste left by students in a school cafeteria to determine whether a new food is acceptable to them

Practice Exercise: Which technique of data collection will you use in the following situation?

- ***Evaluating the customers' satisfaction***
- ***Pre-Poll survey for the election of the president of the students' union***
- ***Work stress management techniques of the employees***
- ***Understanding the opinion of executives on new RBI initiatives for controlling inflation.***

20.11: VARIABLES IN RESEARCH DESIGN

In research variable is a measurable characteristic that varies. It may change from group to group, person to person, or even within one person over time. There are six common variable types:

Independent Variables: These are variables that the researcher has control over. This "control" may involve manipulating existing variables (e.g., modifying existing methods of instruction) or introducing new variables (e.g., adopting a totally new method for some sections of a class) in the research setting. Whatever the case may be, the researcher expects that the independent variable(s) will have some effect on (or relationship with) the dependent variables

Dependent Variables: It shows the effect of manipulating or introducing the independent variables. For example, if the independent variable is the use or non-use of a new language teaching procedure, then the dependent variable might be students' scores on a test of the content taught using that procedure. In other words, the variation in the dependent variable depends on the variation in the independent variable

Intervening Variables: It refer to abstract processes that are not directly observable but that link the independent and dependent variables. In language learning and teaching, they are usually inside the subjects' heads, including various language learning processes which the researcher cannot observe. For example, if the use of a particular teaching technique is the independent variable and mastery of the objectives is the dependent variable, then the language learning processes used by the subjects are the intervening variables.

Moderator variables: These are such variables that affect the relationship between the independent and dependent variables by modifying the effect of the intervening variable(s). Unlike extraneous variables, moderator variables are measured and taken into consideration. Typical moderator variables in TESL and language acquisition research (when they are not the major focus of the study) include the sex, age, culture, or language proficiency of the subjects.

Control Variables: Language learning and teaching are very complex processes. It is not possible to consider every variable in a single study. Therefore, the variables that are not measured in a particular study must be held constant, neutralized/balanced, or eliminated, so they will not have a biasing effect on the other variables. Variables that have been controlled in this way are called control variables.

Extraneous Variables: These are those factors in the research environment that may have an effect on the dependent variable(s) but which are not controlled. Extraneous variables are dangerous. They may damage a study's validity, making it impossible to know whether the effects were caused by the independent and moderator variables or some extraneous factor. If they cannot be controlled, extraneous variables must at least be taken into consideration when interpreting results.

20.12: SAMPLING AND ITS USES

In any research conducted, people, places, and things are studied. The opportunity to study the entire population of those people, places, and things is an endeavor that most researchers do not have the time and/or money to undertake. This limitation is overcome by sampling and statistical techniques. In research design, we have to specify the source of data also for primary data. Sampling is concerned with the selection of a subset of individuals from within a statistical population to estimate characteristics of the whole population. A sample is a representative group of the population.

The sampling process consists of the following steps:

1. Defining the population
2. Specifying a sampling frame,
3. Specifying a sampling method
4. Determining the sample size
5. Implementing the sampling plan
6. Sampling and data collecting

There are a variety of different sampling methods available to researchers to select individuals for a study. Sampling method fall into two categories:

1. **Probability sampling:** Every individual in the population is known and each has a certain probability of being selected. A random process decides the sample based on each individual's probability. Some examples of probability sampling are

    a. **Simple random sample:** Each unit in the population is identified, and each unit has an equal chance of being in the sample. The selection of each unit is independent of the selection of every other unit. The selection of one unit does not affect the chances of any other unit

b. **Systematic random sampling**: Each unit in the population is identified, and each unit has an equal chance of being in the sample.
c. **Stratified random sampling**: Each unit in the population is identified, and each unit has a known, non-zero chance of being in the sample. This is used when the researcher knows that the population has sub-groups (strata) that are of interest.
d. **Cluster sampling**: cluster sampling views the units in a population as not only being members of the total population but as members also of naturally-occurring in clusters within the population. For example, city residents are also residents of neighborhoods, blocks, and housing structures.

2. **Nonprobability sampling:** The population is not entirely known; thus, individual probabilities cannot be known. Common sense or ease is used to choose the sample, but efforts are made to avoid bias and keep the sample representative. Some examples of non-probability sampling are

a. **Convenience sample**: also called an "accidental" sample or "man-in-the-street" samples. The researcher selects units that are convenient, close at hand, easy to reach, etc.
b. **Purposive sample**: the researcher selects the units with some purpose in mind, for example, students who live in dorms on campus, or experts on urban development.
c. **Quota sample**: The researcher constructs quotas for different types of units. For example, to interview a fixed number of shoppers at a mall, half of whom are male and half of whom are female.

20.13: INTRODUCTION OF HYPOTHESIS

Research design is based on some assumption(s) about the research conclusion. This assumption is to be tested by using standard processes and data. Hypothesis is the research assumption that guides the research design. A supposition or explanation (theory) that is provisionally accepted in order to interpret certain events or phenomena and to provide guidance for further investigation is called a hypothesis. A hypothesis may be proven correct or wrong, and must be capable

of refutation. If it remains unrefuted by facts, it is said to be verified or corroborated. Statistically, it is defined as an assumption about certain characteristics of a population. If it specifies values for every parameter of a population, it is called a simple hypothesis; if not, a composite hypothesis. If it attempts to nullify the difference between two sample means (by suggesting that the difference is of no statistical significance), it is called a null hypothesis.

22.14: SUMMARY

In this unit we learned about the concept of research design. Various types of research design were also studied. The criteria of a good research design and types of variables also discussed. Data collection method is an integral part of any research design. We learned about survey, interview and observation as three important methods of data collection.

22.15: GLOSSARY

Epistemology deals with the basic issue on knowledge exploration as "how knowledge is derived and it should be tested and validated

Constructivists- It emphasizes on qualitative mode of investigation and argues that it is the best choice for research in social science as compared to quantitative method

Triangulation – It suggests simultaneous and sequential uses of qualitative and quantitative methods of investigation

Analysis: Uses past situations that are similar to the present research problem.

Focus groups: Involves small (8-12) groups of people brought together and guided by a moderator through unstructured, spontaneous discussion.

Hypothesis : An assumption about certain characteristic of population to be tested statistically

20.16 CHECK YOUR PROGRESS

Q1: What is a research design?

a) A way of conducting research that is not grounded in theory

b) The choice between using qualitative or quantitative methods

c) The style in which you present your research findings, e.g. a graph

d) A framework for every stage of the collection and analysis of data

Q2: If a study is "reliable", this means that:

a) It was conducted by a reputable researcher who can be trusted

b) The measures devised for concepts are stable on different occasions

c) The findings can be generalized to other social settings

d) The methods are stated clearly enough for the research to be replicated

Q3: Which of the following research design is used to get insight into the problem?

a. Exploratory

b. Descriptive

c. Causal

d. None of these

Q4: Which research design is correctly described by the statement "The introduction of planned changes on one or more variables, measurement on small number of variables and control over other variables"

a. Survey

b. Ethnography

c. Experimental

d. Case

Q5: Which of the following error is affected by homogeneity of population being studied and sample size?

a. Population Specification Error

b. Selection Error

c. Sampling Error

d. Non-Response Error

Q6: Which of the following method is used for primary data collection?

a. Survey

b. Interview

c. Observation

d. All of the above

Q7. Why sampling is done in research?

a. To reduce the cost of research

b. To complete the research within time

c. To remove the limitation of accessibility of all members of the population.

d. All of the above

20.17 ANSWERS TO CHECK YOUR PROGRESS

Q1-d, Q2-d, Q3-a, Q4-c, Q5-c, Q6-d, Q7-d

20.18: TERMINAL QUESTIONS

Q1. What is research design? Explain its importance.

Q2. Explain the features of a good research design

Q3. Describe the steps of conducting a research design.

Q4. What is objectivity? State its need in the Research design

Q5. What do you mean by variables in research design? Discuss

Q6. What do you mean by causal studies?

Q7. What do you mean by focus group in research design?

Q8. Explain the exploratory research design and describe its characteristics.

Q9. Distinguish between exploratory and descriptive research design.

Q10. Discuss the major components of a research design.

Q11. What do you mean by interview? What are the types of interview? Discuss

Q12. What is an observation? Discuss its merits and limitations in business research.

Q13. What are the potential sources of errors in research? Discuss with an example.

Q14. What do you mean by experimental design? Discuss its characteristics.

20.19: SUGGESTED READINGS

1. Cresswell, John, W.,(2008). Research Design; Qualitative, Quantitative and Mixed Methods Approaches. Newbury Park, CA: Sage Publication.
3. Marczyk, G.R, DeMatteo, D. & Festinger D., (2005). Essentials of Research Design and Methodology, New York City, NY: Wiley.
4. Ethridge, Don E., (2004). Research Methodology in Applied Economics. Daryaganj, ND: Wiley – Blackwell,
5. Bergh, D. and Ketchen, D. (2009) Research Methodology in Strategy and Management. Binglay, UK: Emarald Group Publishing. Research Methodology: C R Kothari (New Age)
6. Marketing Research : N K Malhotra (Pearsons)

BLOCK 6: RESEARCH METHODOLOGY

UNIT 21: MEASUREMENT AND SCALING TECHNIQUES

STRUCTURE

OBJECTIVES

After reading this unit you would be able to

- Understand the concept of Measurement in research
- Understand the types of measurement scales
- Understand the process of development of scale
- Understand the criteria for a good measurement scale

21.1: INTRODUCTION

An important aspect of research is measurement of variables which are identified in research design. Measurement is the process of describing some property of a phenomenon of interest or variable of study usually by assigning numbers in a reliable and valid way. The numbers convey information about the property being measured. In this way we are able to quantify the qualitative aspects of the attributes of a variable. 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. It is to be noted here that what we measure in research is not the object rather the characteristics of that object. For example if a research design is oriented to measure the satisfaction level of the consumers with products of the company, we measure the satisfaction level not the consumer.

The basic question arises, how to measure? In order to measure the attributes of a variable we need suitable scale. The scales are created specifically in business research for measurement. The process of creating of scale for measurement is called scaling. Scaling is the generation of a broadly defined continuum on which measured objects are located (Peterson, 2000). Scale is a device providing a range of values that correspond to different values in a concept being measured. Correspondence rules indicate the way that a certain value on a scale corresponds to some true value of a concept.

For example, in a survey question if a researcher asks from sales manager a question about the trustworthiness of the sales representative. The response is to be recorded by the answer of a question-

"Assign the numbers 1 through 7 according to how much trust that you have in your sales representative. If the sales representative is perceived as completely untrustworthy, assign the numeral 1, if the sales rep is completely trustworthy, assign a 7." The respondent selects suitable number from 1 to 7 to rank the trust of a sales representative. In this scale we are measuring the trust of sales executive towards sales manager.

Measurement can also be illustrated by thinking about the way instructors assign students' grades. A grade represents a student's performance in a class. Students with higher performance should receive a different grade than do students with lower performance.

As stated earlier, to make the measurement process effective, the relationships existing among the objects or events in the empirical system should directly correspond to the rules of the number system. If this correspondence is misrepresented, measurement error has occurred. The term number indicates the application of numbers to various aspects measured in the measurement process. Data analysis is a statistical process done on the data generated using scales. Hence, all measures should be converted into quantitative terms by applying numbers. However, the definition of measurement imposes certain restrictions on the type of numerical manipulations admissible. The further sections of this unit explain the types and methods of measurements.

21.2: WHAT IS MEASUREMENT IN RESEARCH?

A researcher has to know what to measure before knowing how to measure something. The problem definition process should suggest the concepts that must be measured. As discussed in earlier unit, **concept** can be thought of as a generalized idea that represents something of meaning. Concepts such as age, sex, education, and number of children are relatively concrete properties. They present few problems in either definition or measurement. Other concepts are more abstract. Concepts such as loyalty, personality, channel power, trust, corporate culture, customer satisfaction, value, and so on are more difficult to both define and measure. For example, loyalty has been measured as a combination of customer share (the relative proportion of a person's purchases going to one competing brand/store) and commitment (the degree to which a customer will sacrifice to do business with a brand/store).Thus, we can see that loyalty consists of two

components: the first is behavioral, and the second is attitudinal. While selecting the variable to be measured, it is necessary to understand the concept underlying that variable.

Variables are things that we measure, control, or manipulate in research. Before variables can be measured, they must be defined in concepts. As discussed earlier in previous units, there are the following types of definitions used to describe the variables;

- **Theoretical Definition:** This is the definition used in the dictionary to describe theory, concept, or construct and is commonly used. These types of definitions are generic in nature and used to understand the concept.
- **Operational Definition:** This definition explains how the variable is to be measured in research design. It assigns a meaning to a concept or variable by specifying the operations needed to measure it.

Sometimes, a single variable cannot define a single concept alone. Using multiple variables to measure one concept can often provide a more complete account of some concept than could any single variable. In this perspective construct is used. A **construct is a term used for concepts that are measured with multiple variables.** For instance, when a business researcher wishes to measure the customer orientation of a salesperson, several variables like these may be used, each captured on a 1–5 scale:

1. Offer the product that is best suited to a customer's problem.

2. A good employee has to have the customer's best interests in mind.

3. Try to find out what kind of products will be most helpful to a customer.

Practice Exercise: Create different types of questions to assess the impacts of absenteeism on the functioning of organization. Ask these questions from HR managers. After receiving the response get some explanations from the respondents for the responses given by them.

21.3: WHAT IS SCALING?

Scaling is a process of creating a continuum on which the characteristics of measured objects are located. Researchers measure concepts through a process known as operationalization. This

process involves identifying scales that correspond to variance in the concept. Scales provide a range of values that correspond to different values in the concept being measured. In other words, scales provide correspondence rules. For example, a scaling technique might involve estimating individuals' levels of extraversion or the perceived quality of products. Certain methods of scaling permit estimation of magnitudes on a continuum, while other methods provide only for the relative ordering of the entities.

The steps that are to be followed for developing the scale are:

1. Definition of the concept or concepts to be measured;
2. Identification of the components of the concept;
3. Specification of a sample of observable and measurable items (indicators or proxy variables) that represent the components of the concept;
4. Selection of the appropriate scales to measure the concept;
5. Combination of the items into a composite scale, referred to as an instrument, which in turn serves as a means of measuring the concept.
6. Administer the instrument to a sample and assess respondent understanding
7. Assess reliability and validity.
8. Revise the instrument as needed.

21.4: CRITERIA OF A GOOD SCALE

The five major criteria for analyzing the goodness of measurement are

- **Reliability**: Reliability is an indicator of a measure's internal consistency. Consistency is the key to understanding reliability. A measure is reliable when different attempts at measuring something converge on the same result. For example, consider an exam that has three parts: 25 multiple-choice questions, 2 essay questions, and a short case. If a student gets 20 of the 25 (80 percent) multiple-choice questions correct, we would expect she would also score about 80 percent on the essay and case portions of the exam. Further, if a professor's research tests are reliable, a student should tend toward consistent scores on all tests. In other words, a student who makes an 80 percent on the first test should make scores close to 80 percent on

all subsequent tests. This is the "consistency" of the test--does repeated applications of the test yield similar scores? Reliability is typically evaluated by calculating correlations, and we are always looking for strong, positive **r-values**

- **Test-retest reliability** involves correlating the scores of two administrations of the test to one group of people (people who score high the first time should obtain similar high scores the second time, and vice versa)
- **Parallel forms reliability** involves correlating scores on two different versions of the same test (people who score high on Version A should also score high on Version 2)
- **Split-half reliability** involves correlating scores on one half of the test (e.g., the odd-numbered items) with scores on the other half (e.g., the even-numbered items) --high scores on the first half should correlate positively with high scores on the second half.
- **Statistical Reliability** refers to the probability that the results occurred due to "chance." If the probability is low (the **"magic number:" p < .05),** then we say it is a **"real"** result (i.e., not due to chance; therefore, it is likely to occur again if the experiment were repeated). Be on the lookout for "significant" results in the literature you read (e.g., a **"significant difference"** between two groups is a **"statistically-reliable" difference**).

- **Validity**: **Validity** is the accuracy of a measure or the extent to which a score truthfully represents a concept. In other words, are we accurately measuring what we think we are measuring? This is the "meaningfulness," or **accuracy** of a test--does the test measure what it intends to measure?
  - This is assessed by identifying some **criterion** (e.g., a behavior that the test should predict)
  - The **predictive validity** of the test is determined by correlating scores on the test with scores on the criterion
  - e.g., we could correlate scores on a test measuring the trait of "aggressiveness" (the **predictor variable)** with teachers' records of the number of fights a child has at school (the **criterion variable)**
  - If a strong, positive correlation is obtained between our trait measure and the teachers' records, then we can say that our scale has "predictive validity."

- **Sensitivity:** The sensitivity of a scale is an important measurement concept, particularly when changes in attitudes or other hypothetical constructs are under investigation. **Sensitivity** refers to an instrument's ability to accurately measure variability in a concept. A dichotomous response category, such as "agree or disagree," does not allow the recording of subtle attitude changes. A more sensitive measure with numerous categories on the scale may be needed. For example, adding "strongly agree," "mildly agree," "neither agree nor disagree," "mildly disagree," and "strongly disagree" will increase the scale's sensitivity.

21.5: CLASSIFICATION OF MEASUREMENT SCALES

While conducting research on business-related issues, a researcher has to initially define what is to be measured, how it will be measured, and also the concept that needs to be measured. In the early 1940s, Harvard psychologist S. S. Stevens coined the terms **nominal, ordinal, interval, and ratio** to classify the scales of measurement (Stevens, 1946). Scales of measurement are rules that describe the properties of numbers. These rules imply that a number is not just a number in science. Instead, the extent to which a number is informative depends on how it was used or measured. In this section, we discuss the extent to which data are informative. In all, scales of measurement are characterized by three properties: order, differences, and ratios. Each property can be described by answering the following questions:

1. Order: Does a larger number indicate a greater value than a smaller number?

2. Differences: Does subtracting two numbers represent some meaningful value?

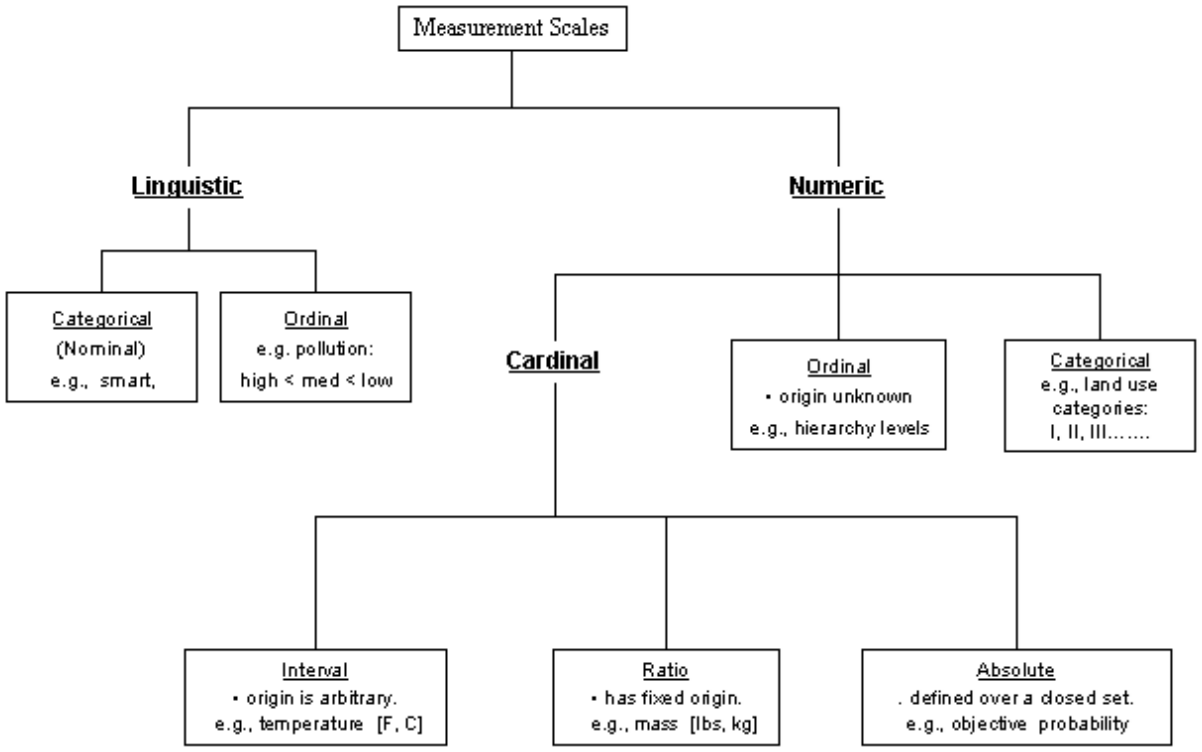
3. Ratio: Does dividing (or taking the ratio of) two numbers represent some meaningful value?

Scales of measurement refer to how the properties of numbers can change with different uses. Measurement scales are normally categorized into four types, namely,

1. Nominal scale
2. Ordinal scale
3. Interval scale and

4. Ratio scale.

Business researchers use many scales or number systems. Not all scales capture the same richness in a measure. Not all concepts require a rich measure. Traditionally, the level of scale measurement is seen as important because it determines the mathematical comparisons that are allowable. The four levels or types of scale offer the researcher progressively more power in analyzing and testing the validity of a scale.



Measurement Scales

Practice Exercise: Discuss with the subject teachers of your university and ask the methods by which they measure the performance of students in class and examinations. Try to relate it with the discussions with different types of scales.

21.5.1: NOMINAL SCALES

The nominal scale represents the most elementary level of measurement. A nominal scale assigns a value to an object for identification or classification purposes only. The value can be, but does not have to be, a number because no quantities are being represented. In this sense, a nominal scale is truly a qualitative scale. Nominal scales are extremely useful, and are sometimes the only appropriate measure, even though they can be considered elementary. Nominal scaling is arbitrary i.e, each label can be assigned to any of the categories without introducing error. Nominal scales are categorical scales used to identify, label, or categorize objects or persons, or events. Nominal scales are the lowest form of measurement. The simple rule is to be followed while developing a nominal scale. In business research, nominal scales are used substantially on many occasions. For example, a nominal scale is used to identify and classify brands, sales regions, awareness of brands, working status of women etc.,

On data generated using a nominal scale, the types of statistical analysis appropriate are mode, percentages, and the chi-square test. Mode alone could be used as a measure of central tendency. Mean and median could be employed on nominal data since they involve higher-level properties of the number system. Researchers should be careful enough to identify the type of scales before they apply any statistical technique. The researcher may not be able to make any meaning inference from the mean or median value obtained from nominal data.

21.5.2: ORDINAL SCALE

An ordinal scale is a ranking scale that indicates an ordered relationship among the objects or events. It involves assigning numbers to objects to indicate the relative extent to which the objects possess some characteristic. It measures whether an object or event has the same characteristic as some other object or event. It is an improvement over the nominal scale in that it indicates an order.

However, this scale does not indicate on how much more or less of the characteristic various objects or events possess. The term how much refers to ranks that it do not indicate if the second rank is a close second or a poor second to the first rank. Data generated using an ordinal scale appears as ranks where the object that has ranked first has more of the characteristic as compared to those objects ranked second or third. Hence, the important feature of the ordinal scale over the nominal scale is that it indicates relative position, not the magnitude of the difference between the objects. In research, ordinal scales are used to measure relative attitudes, opinions, perceptions, etc. Most data collected by the process of interrogating people have ordinal properties. To illustrate, a marketer may be interested in knowing the preference of the customers across various brands. The customers may be requested to rank the products in terms of their preference for the products.

The numbers assigned to a particular object or event can never be changed in ordinal scales. Any violation of this principle would result in confounding results by the researcher. Mean is not an appropriate statistic for an ordinal scale.

21.5.3: INTERVAL SCALE

An interval scale is otherwise called a rating scale. It involves the use of numbers to rate objects or events. In interval scales, numerically equal distances on the scale represent equal values in the Interval scale is otherwise called a rating scale. It involves the use of numbers to rate objects or events. In interval scales, numerically equal distances on the scale represent equal values in the characteristic being measured. An interval scale is an advancement over the ordinal scale in that it has all the properties of an ordinal scale, plus it allows the researcher to compare the differences between objects. It also possesses the property of equality of difference between each level of measurement. The feature of this scale is that the difference between any two scale values is identical to the difference between any other two adjacent values of an interval scale. Examples of interval scales are the Fahrenheit and Celsius scales. Interval scales also place restrictions on the assignment of values to the scale points. The zero that could be assigned is a arbitrary zero rather than natural. Arbitration involves the freedom to place the zero value on any point. There is a constant or equal interval between scale values. In research, most of the research on attitudes,

opinions, and perceptions is done using scales treated as interval scales. All statistical techniques that are employed on nominal and ordinal scales could also be employed on data generated using interval scales.

21.5.4: RATIO SCALES

Ratio scales differ from interval scales in that it has a natural/absolute zero. It possesses all the properties of the normal, ordinal, and interval scales. Data generated using ratio scales may be identified, classified into categories, ranked, and compared with other properties. It could also be expressed in terms of relativity in that one can be expressed in terms of a division of the other. Hence, it may be called as relative scales. Ratio scales have a great many numbers of applications in research. They include sales, market share, costs, ages, and number of customers. In all these cases, natural zero exists. All statistical techniques can be applied to ratio data.

	Scale of Measurement			
Property	Nominal	Ordinal	Interval	Ratio
Order	No	Yes	Yes	Yes
Difference	No	No	Yes	Yes
Ratio	No	No	No	Yes

Scales could also be classified broadly as follow:

- **Metric** scales include summated ratings, numerical scales, semantic differentials, and graphic ratings scales.
- **Nonmetric** scales include categorical, rank order, sorting, constant sum, and paired comparisons.

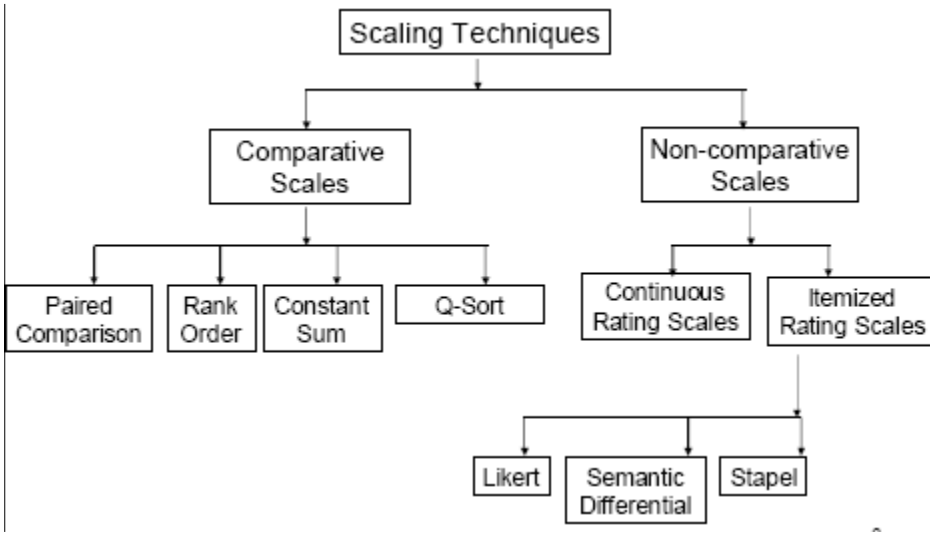
21.6: SCALING TECHNIQUES

This section continues our discussion of how scales are developed and how some of the more common scaling techniques and models can be used. It focuses on broad concepts of attitude scaling—the study of scaling for the measurement of managerial and consumer or buyer perception, preference, and motivation. All attitude (and other psychological) measurement procedures are concerned with people i.e, consumers, purchasing agents, marketing managers, or

whoever responds to certain stimuli according to specified sets of instructions. The stimuli may be alternative products or services, advertising copy themes, package designs, brand names, sales presentations, and so on. The response may involve which copy theme is more pleasing than another, which package design is more appealing than another, what do each of the brand names means, which adjectives best describe each salesperson, and so on. Scaling procedures can be classified in terms of the measurement properties of the final scale (nominal, ordinal, interval, or ratio), the task that the subject is asked to perform, or in still other ways, such as whether the emphasis is to be placed on subject, stimuli, or both.

A well-designed research problem constitutes a well-designed measurement process. The process of measurement is a fundamental aspect of any research. This is the step where you actually try to find out the reality by measuring it. Decision makers are more interested as the steps prior to this step are purely descriptive, and, this is the step where actual quantification happens. The measures should be devoid of measurement errors. There may be disastrous situations where the marketer may be confused with the findings of the data. If he is well aware of the confounding results, then he may discard the findings that emerge from the data analysis. This requires lot of wisdom and knowledge in identifying if the data that resulted from the measurement is consistent, unambiguous etc., but unfortunately, marketers may not be interested in knowing or rather would not know the type of scales used to measure the aspects involved in the marketing problem. Any decision made based on the findings would lot of negative implications on the organisation. Hence, it is very imperative that the researcher is wise enough to develop measurement scales that capture the right property with appropriately.

The scaling techniques employed in research could be broadly classified into comparative and non comparative scale. Comparative scales as its name indicate derive their name from the fact that all ratings are comparisons involving relative judgements. It involves direct comparison of stimulus objects. It contains only ordinal or rank order properties. It is also otherwise called non metric scales in that it does not allow any numerical operations on it against all that could be applied on interval and ratio scales. Comparative scales involve the direct comparison of stimulus objects.



21.6.1: COMPARATIVE SCALING TECHNIQUES

Comparative scaling techniques consist of:

a) Paired comparison scaling

b) Rank order scaling

c) Constant sum scaling and

d) Q-sort.

Paired Comparison Scaling

Paired comparison scaling as its name indicates involves presentation of two objects and asking the respondents to select one according to some criteria. The data are obtained using ordinal scale. For example, a respondent may be asked to indicate his/her preference for TVs in a paired manner. Paired comparison data can be analysed in several ways. In the above example, the researcher can calculate the percentage of respondents who prefer one particular brand of TV over the other.

Under the assumption of transitivity, data generated using paired comparison technique could be converted to a rank order. Transitivity of preference implies that if a respondent prefers brand X over brand Y, and brand Y is preferred to Z, then brand X is preferred to Z. This may be done by determining the number of times each brand is preferred by preference, from most to least preferred.

Paired comparison technique is useful when the number of brands is limited, as it requires direct comparison and overt choice. However, it is not so, that possible comparison could not be made, but comparisons would become so much unwieldy. The most common method of taste testing is done by paired comparison where the consumer may be, for example, asked to taste two different brands of soft drinks and select the one with the most appealing taste.

Rank Order Scaling

This is another popular comparative scaling technique. In rank order scaling is done by presenting the respondents with several objects simultaneously and asked to order or rank them based on a particular criterion. For example, the customers may rank their preference for TVs among several brands. In this scaling technique, ordinal scale is used. The consumers may be asked to rank several brands of television in an order, 1 being the most preferred brand, followed by 2, 3 and so on. Like paired comparison, it is also comparative in nature. Data generated using this technique are employed with conjoint analysis because of the discriminatory potential of the scaling, stimulating the consumers to discriminate one brand from the other. Under the assumptions of transitivity, rank order can be converted to equivalent paired comparison data, and vice versa.

Constant Sum Scaling

This technique allows the respondents to allocate a constant sum of units, such as points, rupees or among a set of stimulus objects with respect to some criterion. The technique involves asking the respondents to assign 10 points to attributes of a sports utility vehicle. If the attribute is unimportant, then the respondents would want to enter zero. The attributes are scaled by counting the points assigned to each one by all the respondents and dividing by the number of respondents. This predominantly uses ordinal because of its comparative nature and the resulting lack of generalisability. Constant sum scaling has an advantage in that it allows for discrimination among stimulus objects without requiring too much time. Its advantage involves the allocation of more or fewer units than those specified.

Q-Sort

Q-sort refers to discriminating among a relatively large number of objects quickly. This technique uses a rank order procedure in which objects are sorted into piles based on similarity with respect to some criterion. A typical example quoted in Malhotra (2004) is as follows. Respondents are given 100 attitude statements on individual cards and asked to place them into 11 piles, ranging from „most highly agreed with" to „least highly agreed with". The number of objects to be sorted should not be less than 60 nor more than 140: 60 to 90 objects is a reasonable range. The number of objects to be placed in each pile is pre-specified, often to result in a roughly normal distribution of objects over the whole set.

21.6.2 NON-COMPARATIVE SCALING TECHNIQUES

Non-comparative scales or otherwise called as nomadic scales, because only one object is evaluated at a time. Researchers use this scale, allowing respondents to employ whatever rating standard seems appropriate to them and not specified by the researcher. The respondents do not compare the object being rated either to another object or to some specified standard set by the researcher. Non-comparative techniques use continuous and itemized rating scales. In such scales, each object is scaled independently of the other objects in the stimulus set, the resulting data is generally assumed to be interval or ratio scale.

Continuous Rating Scale

This is also otherwise called as a graphic rating scale. This is a type of scale that offers respondents a form of continuum (such as a line) on which to provide a rating of an object. Researchers develop a continuous rating scale allowing the respondents to indicate their rating by placing a mark at the appropriate point on a line that runs from one end of the criterion variable to the other or a set of predetermined response categories. Here, the respondents need not select marks already set the researcher. Several variations are possible. The line may be vertical or horizontal; it may be unmarked or marked; if marked, the divisions may be few or as many as in the thermometer scale; the scale points may be in the form of numbers or brief descriptions. Three versions are normally used as given in the table below:

Examples of continuous rating scale:

Please evaluate the service quality of a restaurant by placing an x at the position on the horizontal line that most reflects your feelings

Empathy

The worst ----------

The best

Continuous rating scales are easy to construct, however, the scoring may be cumbersome and unreliable. With the advent of computers in research, they are increasingly used, though they otherwise provide little new information.

Itemized Rating Scales

This scale is similar to the graphic scale in that the individuals make their judgement independently, without the benefit of direct comparison. The respondents are provided with a scale that has a number or brief description associated with each category. This scale allows the respondents to choose from a more limited number of categories, usually five to seven, although 10 or more are rarely used. The categories are ordered in terms of scale position, and the respondents are required to select the specified category that best describes the object being rated. The categories are given verbal descriptions, although this is not absolutely necessary. These scales are widely used in research and nowadays, more complex types such as multi-item rating scales are used. There are few variants among itemised rating scales. They are Likert, Semantic differential and Stapel scales.

Likert Scale

This scale is named after Renis Likert. This is the most widely used scale in research, in particular, in testing models. Several research studies are done using Likert scale. The respondents require to indicate a degree of agreement of disagreement with each of a series of statements about the stimulus objects. Example of a portion of a popularly used Likert scale to measure tangibility of service is given below.

Listed below are the tangibility of service rendered by a bank is listed below. Please indicate how strongly you agree or disagree with each by using the following scale

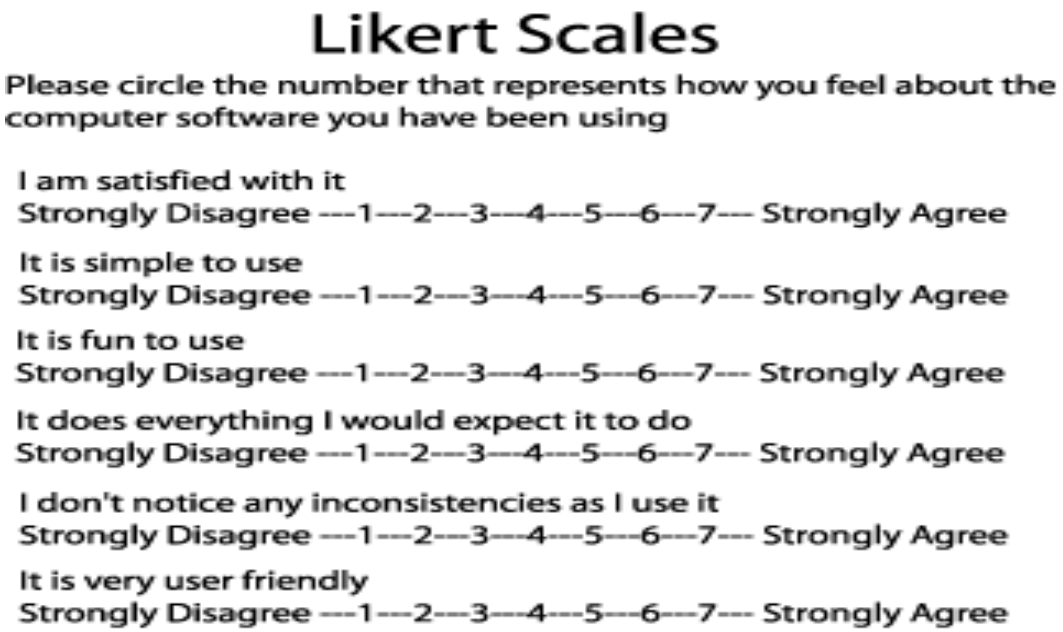
1 = Strongly disagree

2 = Disagree

3 = Neither agree nor disagree

4 = Agree

5 = Strongly agree



To analyse the data generated using this scale, each statement is assigned a numerical score, ranging either from -2 to +2 through a zero or 1 to 5. The analysis can be conducted item-wise or a total score (summated) or a mean can be calculated for each respondent by summing or averaging across items. It is important in Likert scale that a consistent scoring procedure so that a high score reflects favourable response and a low score reflects an unfavourable response. Any deviation in the form of reverse coding, where the lowest value is given to a favourable response and the highest value is given to an unfavourable response should be clearly specified by the researcher. Usually, reverse coding is used when the statements indicate a negative concept, and when used with other statements, reverse coding would give a positive effect.

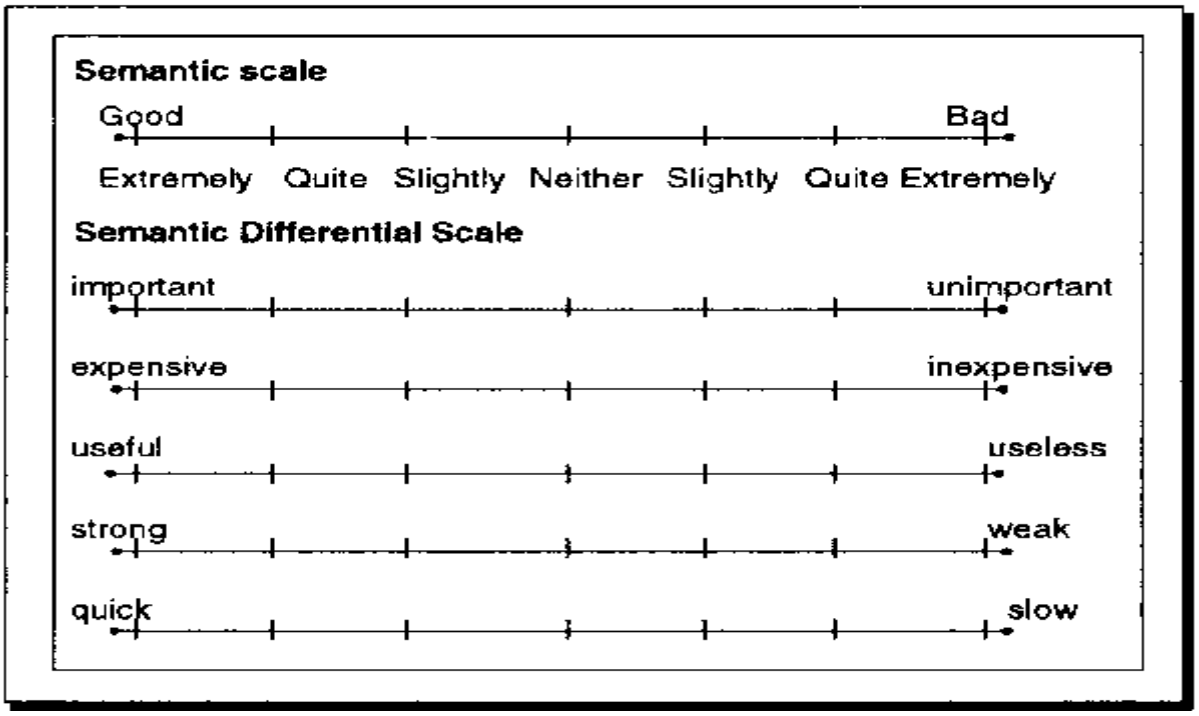
Semantic Differential Scale

Semantic differential scale is a popular scaling technique next to Likert scale. In this scale, the respondents associate their response with bipolar labels that have semantic meaning. The respondents rate objects on a number of itemised, seven point rating scales bounded at each end by one of two bipolar adjectives such as "Excellent" and "Very bad". The respondents indicate their response choosing the one that best describes their choice. The points are marked either from - 3 to +3 through a zero or from 1 to 7. The middle value may be treated as a neutral position. The value zero in the first type is the neutral point and 4 in the second type is the neutral point. The resulting data are commonly analysed through profile analysis. In such analysis, the means or median values on each rating scale are calculated and compared by plotting or statistical analysis. This would help the researcher to determine the overall differences and similarities among the objects. To assess differences across segments of respondents, the researcher can compare mean responses of different segments. This data generated using this scale could be employed with summary statistics such mean, though, there is a controversy on the employment of mean on this scale. Mean is typical of Interval and ratio scales whereas this scale theoretically is an ordinal scale. However, looking beyond this objection by statisticians, researchers invariably apply all statistical techniques on this scale. The following example illustrates semantic differential scales

1) Pleasant ---------- unpleasant

2) Aggressive ---------- submissive

3) Exciting ---------- unexciting



Stapel Scale

This scale is named after Jan Stapel, who developed it. This is a unipolar rating scale with in general 10 categories number from -5 to +5 without a neutral point (zero). This scale is usually presented vertically and respondents choose their response based on how accurately or inaccurately each item describes the object by selecting an appropriate numerical response category. The higher number indicates more accurate description of the object and lower number indicates lower description of the object. An example is given below:

+ 5

+4

+3

+2

+1

High tangibility of service

-1

-2

-3

-4

-5

The data generated using staple scale could be analysed in the same way as semantic differential scale. The main advantage of Stapel Scale is that it does not require a retest of the adjectives or phrases to ensure true bipolarity, and it can be administered over the telephone.

Mathematical and Statistical Analysis on the data of different Scales

Sl No	Scales	Basic Characteristics	Examples	Permitted Statistics	
				Descriptive	Inferential
1	Nominal	Numbers identify and classify objects	Types of store, Yes or no choice, gender	Percentage, Mode	Chi Square, Binomial test
2	Ordinal	Numbers indicates the relative positions of the objects but not the magnitude of difference between them	Preference ranking, Quality Ranking	Percentile, Median	Rank Order, Correlation, ANOVA
3	Interval	Difference between objects can be compared; Zero point is arbitrary	Attitude, Opinions, Index Numbers	Range, Mean, Standard Deviation	Moment, Correlation, t-test, ANOVA, regression and factor analysis
4	Ratio	Zero Point is fixed; ratios of scale values can be computed	Age, Income, Costs, Sales, Market Shares	Geometric Mean, harmonic mean	Coefficient of variation

Practice Exercise. Indicate whether the following measures use a nominal, ordinal, interval, or ratio scale:

a. Prices on the stock market

b. Marital status, classified as "married" or "never married"

c. A yes/no question asking whether a respondent has ever been unemployed

d. Professorial rank: assistant professor, associate professor, or professor

e. Grades: A, B, C, D, or F

When we try to develop a measurement scale in business research various considerations are required. Followings are the necessary decisions when developing the scales:

- **Number of scale categories** – Larger the number, the greater precision of the measurement scale is required.
- **Number of items to measure a concept** – What should be the numbers of items? Minimum of 3 items is necessary to achieve acceptable reliability, but it is common to see at least 5 to 7 items and sometimes more.
- **Odd or even number of categories** – The mid-point typically represents a neutral position when an odd number of categories are used in a scale. If researcher wants to force a choice on a particular issue, than an even number should be used.
- **Balanced or unbalanced scales** – Balanced means the number of favorable and unfavorable categories is equal, and unbalanced means they are not. Unbalanced scales are used when the researcher expects responses to be skewed toward one end of the scale.
- **Forced or non-forced choice** – With forced-choice scales, respondents are forced to make a choice. There is no mid-point that can be considered a neutral or no opinion category. If a respondent selects the middle category when they have 'no opinion' or are 'neutral' this will cause error in the responses, so it's better to use a forced-choice scale and provide a 'no opinion' option.

- **Category labels for scales** – Three types of category labels are verbal, numerical, and unlabeled choices. Numerical and unlabeled scales are used when researchers have difficulty in developing appropriate verbal descriptions for the middle categories.

21.7: ERRORS IN MEASUREMENT

Measurement error is defined as the difference between the distorted information and the undistorted information about a measured product. A variety of sources can cause measurement error, including response styles, specifically acquiescence, disacquiescence, extreme response, response range, midpoint responding, and non-contingent responding (Baumgartner & Steenkamp, 2001; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). Some types of errors are discussed in this section.

- **Error due to agreement bias**: Agreement bias is a tendency to agree with statements, irrespective of the content of the item. Also referred to as acquiescence response style.
- **Error due to location bias:** It occurs when individuals differ in the manner in which they use response scale categories (e.g., a tendency to scale upward or use extremes).
- **Leniency error:** It is the tendency of a respondent to rate too high or too low.
- **Severity error:** It is the opposite of leniency.
- **Midpoint responding error**: It is a tendency to use the middle scale point irrespective of content (Baumgartner & Steenkamp, 2001), which may be caused by evasiveness, indecision, or indifference (Messick, 1968; Schuman & Presser, 1981).
- **Extreme response error:** It is a style that refers to choosing extreme responses irrespective of content

21.8: QUESTIONNAIRE DESIGNING

In previous sections we have studied fundamental concepts related to measurement and scaling. In this section we will apply the learned concept in form of questionnaire designing. Although there are no set rules for developing a questionnaire, the collective experience of numerous researchers offer a broad set of guidelines for minimising the likelihood and the severity of data validity

problems in designing questionnaires. On this topic, an excellent reference remains Boyd and Westfall (1992). A seven-step procedure is proposed to assist the design of a questionnaire.

Step 1: Determine the information required

Since the questionnaire is the link between the information needs and the data to be collected, the researcher must have a **detailed listing of the information needs** as well as a clear identification of the respondent group. This step is normally the result of exploratory research and of the hypotheses development phases. The different forms of market response described in the preceding chapter will help the analyst to identify the concepts to be measured.

Step 2: Determine the type of questionnaire to be used

Data collection can be made by personal interview, mail or telephone. The choice among these alternatives is largely determined by the type of information to be obtained. It is necessary to decide on the type of questionnaire at this point since the content and wording of the questions, the length of the questionnaire, and the sequence of questions will all be influenced by this decision. A decision to use conjoint analysis, for example, would preclude the use of a telephone interview. Thus, at this stage, the market analysis must specify precisely how the primary data needed will be collected and also the type of analysis to be made with the data.

Step 3: Determine the content of individual questions

Once the information needed is known and the data collection method decided, the researcher is ready to begin formulating the questions. Several points should be reviewed systematically once the content of the questions is determined.

- **Is the question necessary?** Avoid including interesting questions that are not directly related to the information needed.
- **Are several questions needed instead of one?** Some questions may have two or more elements, and if these are left in one question, interpretation becomes impossible. This is typically the case for the 'why' question.
- **Does the respondent have the information requested?** Three sub-questions can be examined to determine this:
    - Is the point raised within the respondent's experience?
    - Can the respondent remember the information?
    - Will the respondent have to do a lot of work to get the information?

- **Will respondents give the information?** Even though they know the answer, respondents will sometimes not answer questions, because:
  - They are unable to phrase their answer
  - They do not want to answer.

Step 4: Determine the type of question to use

In forming the actual questions, the researcher has the choice between three major types of questions.

- An **open-ended question** requires the respondents to provide their own answers to the question.
- A **multiple-choice question** requires the respondent to choose an answer from a list provided with the question. The respondent may be asked to choose one or more of the alternatives presented.
- A **dichotomous question** is an extreme form of the multiple-choice question, which allows the respondent only two responses, such as yes/no, agree/disagree, and so on.

In a multiple-choice question, when the proposed answers are ranked, the objective is not simply to identify a category (as in a nominal scale), but rather to 'measure' a level of agreement, a degree of importance, or a level of preference.

Step 5: Decide on the wording of questions

The problem at this stage is to phrase the questions in a way that:

• The respondent can easily understand

• Does not give the respondent a clue as to how he or she should answer.

Consider the following points to make sure questions are both comprehensible and unbiased.

• **Define the issue clearly.** Each question should be checked on six points – who, where, when, what, why and how – to be sure that the issue is clear.

• **Decide whether to be subjective or objective.** A subjective question puts the question in terms of the individual while objective phrasing tends to refer to what people in general think. Subjective questions tend to give more reliable results.

• **Use simple words.** Words used in questionnaires should have only one meaning, and everyone should know the meaning. There are many examples of misunderstanding of what seem to be everyday words. In particular, the technical jargon of marketing (brand image,

positioning, etc.) should be avoided. The pre-test of the questionnaire is very useful in overcoming this difficulty.

• **Avoid ambiguous questions.** Ambiguous questions mean different things to different people. Indefinite words include 'often', 'occasionally', 'frequently', 'many', 'good', 'fair', 'poor', and so on; these may have many different meanings. For example, 'frequent reading' of a monthly magazine may be six or seven issues a year for one person and twice a year for another.

• **Avoid leading or one-sided questions**. A leading question is one that may steer respondents towards a certain answer. One-sided questions present only one aspect of an issue. A question should be constructed in as neutral a way as possible, by avoiding the name of a brand or a company, or by presenting all the sides of an issue.

• **Avoid double-barrelled questions**. A double-barreled question calls for two responses and thereby creates confusion for the respondents. Such questions should be divided into two separate parts.

• **Use split-ballot wherever possible**. There is no 'correct' wording for a question. When there are two wordings from which to choose, but no basis on which to select one in preference to the other, one wording can be adopted on half of the questionnaires and the other on the other half.

Step 6: Decide the sequence of questions

There are generally three major sections in a questionnaire:

• The basic information sought

• The socio-demographic information is useful in obtaining the profile of the respondent

• The identification sections to be used by the interviewer.

The general rule is to put the sections in that order: the main body of the questionnaire should go in first position and the socio-demographic questions at the end (unless they serve as filter questions to qualify respondents for the survey).

The following points should also be considered in order to determine the sequence of questions to maximum effect:

• **Use simple and interesting opening questions**. If the opening questions are interesting, simple to comprehend and easy to answer, the respondent's co-operation will be gained.

- **Use the funnel approach**. The funnel approach involves beginning with a very general question on a topic and gradually leading up to a narrowly focused question on the same topic.
- **Arrange questions in logical order**. Sudden changes in subject confuse the respondent and cause indecision.
- **Place difficult or sensitive questions near the end**. Sensitive questions should be relegated towards the end of the questionnaire, once the respondent has become involved in the study.

A mail questionnaire raises specific sequence problems since it must sell itself. It is particularly important that the opening questions capture the respondent's interest. Questions should then proceed in logical order. However, in a mail questionnaire, it is not possible to take advantage of sequence position in the same way as in personal interviews since it is the respondent who will decide the order of response. The layout and physical attractiveness is particularly important for a self-administered questionnaire.

Step 7: Pre-test the questionnaire

Before a questionnaire is ready for the field it needs to be pre-tested under field conditions. Pre-testing involves administering the questionnaire to a limited number of potential respondents selected on a convenience basis, but not too divergent from the target population. It is not necessary, however, to have a statistical sample for pre-testing. The pre-testing process allows the researcher to determine whether the respondents have any difficulty in understanding the questionnaire and whether there are ambiguous or biased questions. Tabulating the results of the pre-test is also very useful to ensure that all the required information will be obtained.

21.9: SUMMARY

In this unit we have learned the concept of measurement and scaling. It is important to understand that measurement is assigning numerical values to variables or psychological constructs to quantify the decisions. Various levels of measurement scales are nominal, ordinal, ratio and interval. Some standard measurement scales are developed like likert, semantic differential etc.

21.10: GLOSSARY

Measurement: A process of assigning numbers to a concept or variable in a scientific way

Construct: An operationally defined concept to be measured

Variables: Any construct that varies in research design

Scaling: Scaling is a process of creating a continuum on which the characteristics of measured objects are located

Validity: A measure of a good scale to ensure "measures what is intended to measure"

Sensitivity: It refers to an instrument's ability to accurately measure variability in a concept.

21.11: CHECK YOUR PROGRESS

Q1: Which of the following is/are the measurable concepts in business research?

a. Satisfaction level

b. Attitude

c. Both

d. None of these

Q2: Which of the following is used to describe the characteristic of a measurement scale as "to measure what it is supposed to measure"?

a. Reliability

b. Validity

c. Sensitivity

d. None of these

Q3: Which of the following orders of scale represents increasing complexity?

a. Ordinal, Nominal, Interval, Ratio

b. Nominal, Ordinal, Interval, Ratio

c. Interval, Nominal, Ordinal, Ratio

d. Ration, Ordinal, Interval, Nominal

Q4: Which scale involves assigning numbers to objects to indicate the relative extent to which the objects possess some characteristic?

a. Ratio

b. Ordinal

c. Interval

d. Nominal

Q5: All psychological variables like attitude, belief, and satisfaction levels are?

a. Unipolar

b. Bipolar

c. Neutral

d. None of these

Q6: In which case principle of transitivity could be applied?

a. Nominal

b. Ordinal

c. Ratio

d. None of these

e.

21.12: ANSWER TO CHECK YOUR PROGRESS

Q1-a, Q2-b, Q3-a, Q4-b, Q5-b, Q6-b

21.13: TERMINAL QUESTIONS

Q1. What do you mean by Measurement in business research?

Q2. What do you mean by scales in measurement? What are the levels of the scales?

Q3. What are the difficulties one faces in measurement?

Q4. What do you mean by reliability? Discuss

Q5. What do you mean by validity? Discuss

Q6. What are the criteria for a good measurement scale?

Q7. What do you mean by open-ended and closed-ended questions in a questionnaire?

Q8. Discuss the characteristics of the Likert Scale with an example. How is it different from the Semantic Differential Scale?

Q9. What different types of measurement scales with examples? What are the permitted statistical operations associated with each scale?

Q10. Develop a Likert scale, Semantic Differential Scale, and Stapel Scale for measuring loyalty of customers towards a shopping mall.

Q11. What are the steps to be followed for the creation of a good questionnaire? Discuss

Q12. What are the errors in measurement? How could these errors be reduced?

Q13. Comment on the validity and reliability of the following:

a. A respondent's report of an intention to subscribe to Consumer Reports is highly reliable. A researcher believes this constitutes a valid measurement of dissatisfaction with the economic system and alienation from big business.

b. A general-interest magazine claimed that it was a better advertising medium than television programs with similar content. Research had indicated that for soft drinks and other test products, recall scores were higher for the magazine ads than for 30-second commercials.

c. A respondent's report of the frequency of magazine reading consistently indicates that she regularly reads Good Housekeeping.

21.14: SUGGESTED READINGS

1.Cresswell, John, W.,(2008). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Newbury Park, CA: Sage Publication.

2.Marczyk, G.R, DeMatteo, D. & Festinger D., (2005). Essentials of Research Design and Methodology, New York City, NY: Wiley.

1. Ethridge, Don E., (2004). Research Methodology in Applied Economics. Daryaganj, ND: Wiley–Blackwell,
2. Bergh, D. and Ketchen, D. (2009) Research Methodology in Strategy and Management. Binglay, UK: Emarald Group Publishing.
5. Research Methodology: C R Kothari (New Age)
6. Marketing Research: N K Malhotra (Pearsons)

BLOCK 6: RESEARCH METHODOLOGY

UNIT 22: INTERPRETATION, REPORT WRITING & COMPUTER APPLICATION IN RESEARCH

OBJECTIVES

After reading this chapter, the student should be able to:

- Understand the process of research report preparation and presentation
- Understand the basic requirements of report preparation, including report format, report writing, graphs, and tables.
- Understand the nature and scope of the written and oral presentation of the report.
- Identify the ethical issues related to the interpretation and reporting of the research process and findings to the client.
- Explain the use of the Internet and computers in report preparation and presentation.

22.1: INTRODUCTION

The results of research must be effectively and meaningfully communicated to the intended users. It helps in taking decisions. A research report is a systematic, well-organized document which presents the complete activities, research design, processes, interpretations and findings of a research work in a systematic and logical writing. It may be presented orally with the help of audio visual tools and computers after getting it compiled in written form. It starts with recording and interpreting the sequence of events followed during research work and ends with conclusions and references. Presenting the findings of a research study to users involves a formal written report as well as an oral presentation. The report and its presentation are extremely important. Preparing a research report involves other activities besides writing; in fact, writing is actually the last step in the preparation process. Before writing can take place, the results of the research project must be fully understood and thought must be given to what the report will say. Thus, preparing a research report involves three steps: understanding, organizing and writing. Further sections of this unit will discuss in details about various aspects of preparation of research report. In this unit we will learn how to write a research report. To check your progress of learning various types of practice exercises and questions are included.

22.2: ANALYSIS AND INTERPRETATION OF DATA

Before we write a research report it is necessary to interpret the data. Interpretation refers to the task of drawing inferences from the collected facts after an analytical and experimental study. Usefulness and utility of research findings lie in proper interpretation. Data analysis and interpretation is the process of assigning meaning to the collected information and determining the conclusions, significance, and implications of the findings. The steps involved in data analysis are the function of the type of information collected.

From a managerial perspective, data can be viewed as recorded information useful for making decisions. Completed questionnaires or other measurement instruments must be edited, coded, entered into a data set for processing by computer, and carefully analyzed before their complete meanings and implications can be understood.

Depending upon the types of research, data analysis could be of quantitative as well as qualitative data analysis. There is an important distinction between quantitative data analysis and qualitative data analysis. In quantitative research, the information obtained from the participants is expressed in numerical form. Studies in which we record the number of items recalled, reaction times, or the numbers of aggressive acts are all examples of quantitative research. In qualitative research, on the other hand, the information obtained from participants is not expressed in numerical form. The emphasis is on the stated experiences of the participants and on the stated meanings they attach to themselves, to other people, and to their environment. Those carrying out qualitative research sometimes make use of direct quotations from their participants, arguing that such quotations are often very revealing

Some characteristics of qualitative and quantitative data analysis are discussed here.

Quantitative data analysis- The analysis of quantitative data is represented in mathematical terms. The most common types of statistical analysis for quantitative data analysis are as follow

- **Mean** – The mean score represents a numerical average for a set of responses.

- **Standard deviation** – The standard deviation represents the distribution of the responses around the mean. It indicates the degree of consistency among the responses. The standard deviation, in conjunction with the mean, provides a better understanding of the data. For example, if the mean is 3.3 with a standard deviation (SD) of 0.4, then two-thirds of the responses lie between 2.9 (3.3 – 0.4) and 3.7 (3.3 + 0.4).
- **Frequency distribution** – Frequency distribution indicates the frequency of each response. For example, if respondents answer a question using an agree/disagree scale, the percentage of respondents who selected each response on the scale would be indicated. The frequency distribution provides additional information beyond the mean, since it allows for examining the level of consensus among the data.

Besides these we may calculate Median, Mode and Variance also.

Higher levels of statistical analysis (e.g., t-test, factor analysis, regression, ANOVA) can also be conducted on the data which depends upon the research design and intentions of the researcher. The discussions on these aspects are beyond the scope of your syllabus.

Qualitative Data Analysis- The analysis of qualitative data is conducted by organizing the data into common themes or categories. It is often more difficult to interpret narrative data since it lacks the built-in structure found in numerical data. Initially, the narrative data appears to be a collection of random, unconnected statements. The assessment purpose and questions can help direct the focus of the data organization. The following strategies may also be helpful when analyzing narrative data.

The cardinal principle of qualitative analysis is that causal relationships and theoretical statements be clearly emergent from and grounded in the phenomena studied. The theory formulation after causal relationship emerges from the data; it is not imposed on the data.

Practice Exercise : In small groups of three or four people, consider how you might conduct a study to analyse the number of aggressive acts witnessed by children when they watch television cartoon programmes designed for a child audience. Would quantitative methods be most appropriate? How would you ensure that your results were as reliable as possible?

Analysis can be viewed as the categorization, the aggregation into constituent parts, and the manipulation of data to obtain answers to the research question or questions underlying the research project. A special aspect of analysis is interpretation. The process of interpretation involves taking the results of analysis, making inferences relevant to the research relationships studied, and drawing managerially useful conclusions about these relationships. The analysis of obtained data represents the end of the research process except for the preparation of the report, and everything done prior to this stage has been done so that the proper analysis can be

22.3: PROCESS OF DATA ANALYSIS

The overall process of analyzing and making inferences from sample data can be viewed as a process of refinement that involves a number of separate and sequential steps that may be identified as part of three broad stages:

1. **Tabulation**: It involves identifying appropriate categories for the information desired, sorting the data into them, making the initial counts of responses, and using summarizing measures to provide economy of description and so facilitate understanding. In the tabulation process, appropriate categories are defined for coding the information desired, making the initial counts of responses, and preparing a descriptive summary of the data.
2. **Formulating hypotheses**: It involves formulation of assumption using the inductions derived from the data concerning the relevant variables, their parameters, their differences, and their relationships. The hypothesis is to be tested statistically to draw conclusions which are statistically significant.
3. **Making inferences**: It involves reaching conclusions about the variables that are important, their parameters, their differences, and the relationships among them.

22.4: CLASSIFICATION AND TABULATION OF DATA

The most fundamental step in data analysis and interpretation is classification and tabulation. Classification is the way of arranging the data in different classes in order to give a definite form and a coherent structure to the data collected, facilitating their use in the most systematic and effective manner. It is the process of grouping the statistical data under various understandable, homogeneous groups for the purpose of convenient interpretation. A uniformity of attributes is the basis criterion for classification; and the grouping of data is made according to similarity. Classification becomes necessary when there is diversity in the data collected for meaningful presentation and analysis. However, in respect of homogeneous presentation of data, classification may be unnecessary.

Objectives of classification of data:

- To group heterogeneous data under the homogeneous group of common characteristics;
- To facility the similarity of various group;
- To facilitate effective comparison;
- To present complex, haphazard, and scattered dates in a concise, logical, homogeneous, and intelligible form;
- To maintain the clarity and simplicity of complex data;
- To identify independent and dependent variables and establish their relationship;
- To establish a cohesive nature for the diverse data for effective and logical analysis;
- To make logical and effective quantification
- A good classification should have the characteristics of clarity, homogeneity, and equality of scale, purposefulness, accuracy, stability, flexibility, and unambiguity.

Classification is of two types, viz., **quantitative classification,** which is on the basis of variables or quantity; and **qualitative classification**, which is according to attributes. The former is the way of grouping the variables, say quantifying the variables in cohesive groups, while the latter group

the data on the basis of attributes or qualities. Classifications may be of **multiple classification** or **dichotomous classification.** The former is the way of making many (more than two) groups on the basis of some quality or attributes, while the latter is the classification into two groups on the basis of the presence or absence of a certain quality.

Grouping the workers of a factory under various income (class intervals) groups comes under multiple classifications; and making two groups into skilled workers and unskilled workers is dichotomous classification. The tabular form of such classification is known as statistical series, which may be inclusive or exclusive.

Practice Exercise: You have to conduct survey on the satisfaction level with after sales services of a consumer durable company. List down all the characteristics of respondents on which you feel that satisfaction level depends. Classify each characteristic of respondent in quantitative, qualitative, multiple and dichotomous.

The classified data may be arranged in tabular forms called as tables in the intersections of columns and rows. Tabulation is the simplest way of arranging the data, so that anybody can understand it most easily. It is the most systematic way of presenting numerical data in an easily understandable form. It facilitates a clear and simple presentation of the data, a clear expression of the implication, and an easier and more convenient comparison. There can be simple or complex tables, and general purpose or summary tables. Classification and tabulation are interdependent events in research.

Following steps are involved in the process of data tabulation:

- **Categorizing.** Define appropriate categories for coding the information collected.
- **Coding.** Assign codes to the respondent's answers.
- **Create Data File.** Enter the data into the computer and create a data file.
- **Error Checking**. Check the data file for errors by performing a simple tabulation analysis to identify errors in coding or data entry. Once errors are identified, data may be recoded to collapse categories or combine or delete responses.
- **Generate New Variables**. New variables may be computed by data manipulations that multiply, sum, or otherwise transform variables.

- **Weight Data Subclasses.** Weights are often used to adjust the representation of sample subgroups so that they match the proportions found in the population.
- **Tabulate**. Summarize the responses to each variable included in the analysis.

22.5: BASIC STATISTICS FOR ANALYSIS AND INTERPRETATION OF DATA

Statistics provides fundamental tools for analysis and interpretation of the data. The data could also be analyzed without help of statistical tools also, but the generalization of findings may be invalid and not reliable. The analysis of the data via statistical measures and/or narrative themes provides answers to the research questions. Interpreting the analyzed data from the appropriate perspective helps in determining the significance and implications of the inference. In this section, we will discuss some important statistical measures for data analysis and interpretation.

Measure of Central Tendencies: Measures of central tendency describe how the data cluster together around a central point. There are three main measures of central tendency: **the mean, the median and the mode.**

Mean: The **mean** in each group or condition is calculated by adding up all the scores in a given condition, and then dividing by the number of participants in that condition. Suppose that the scores of the nine participants in the no-noise condition are as follows: 1, 2, 4, 5, 7, 9, 9, 9, and 17. The mean is given by the total, which is 63, divided by the number of participants, which is 9. Thus, the mean is 7.

The main advantage of the mean is the fact that it takes all the scores into account. This generally makes it a sensitive measure of central tendency, especially if the scores resemble the **normal distribution**, which is a bell-shaped distribution in which most scores cluster fairly close to the mean. However, the mean can be very misleading if the distribution differs markedly from the normal and there are one or two extreme scores in one direction. Suppose that eight people complete one lap of a track in go-karts. For seven of them, the times taken (in seconds) are as follows: 25, 28, 29, 29, 34, 36, and 42. The eighth person's go-kart breaks down, and so the driver has to push it around the track. This person takes 288 seconds to complete the lap. This produces

an overall mean of 64 seconds. This is clearly misleading, because no one else took even close to 64 seconds to complete one lap and it does not represent central tendency.

Median: Another way of describing the general level of performance in each condition is known as the **median**. If there is an odd number of scores, then the median is simply the middle score, having an equal number of scores higher and lower than it. In the example with nine scores in the no-noise condition (1, 2, 4, 5, 7, 9, 9, 9, 17), the median is 7. Matters are slightly more complex if there is an even number of scores. In that case, we work out the mean of the two central values.

For example, suppose that we have the following scores in size order: 2, 5, 5, 7, 8, 9. The two central values are 5 and 7, and so the median is

$$\frac{5+7}{2} = 6$$

The main advantage of the median is that it is unaffected by a few extreme scores, because it focuses only on scores in the middle of the distribution. It also has the advantage that it tends to be easier than the mean to work out. The main limitation of the median is that it ignores most of the scores, and so it is often less sensitive than the mean. In addition, it is not always representative of the scores obtained, especially if there are only a few scores.

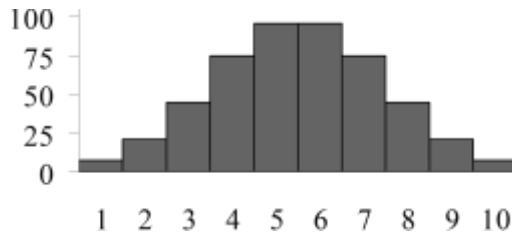
Mode: The final measure of central tendency is the **mode**. This is simply the most frequently occurring score. The main advantages of the mode are that it is unaffected by one or two extreme scores and that it is the easiest measure of central tendency to work out.

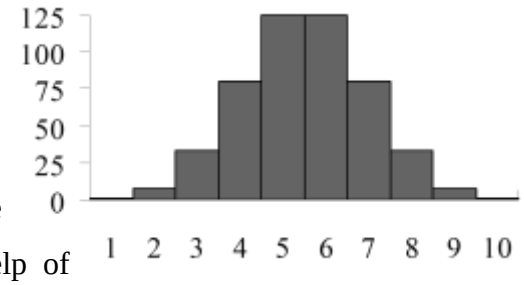
Measures of Dispersions:

Measures of dispersion measure how spread out a set of data is. Measures of dispersion are descriptive statistics that describe how similar a set of scores are to each other. The more similar the scores are to each other, the lower the measure of dispersion will be. The less similar the scores are to each other, the higher the measure of dispersion will be. In general, the more spread out a distribution is, the larger the measure of dispersion..

Information about the dispersion in a sample can be presented in several ways. If it is presented in a graph or chart, this may make it easier for people to understand what has been found, compared to simply presenting information about the central tendency and dispersion.

Practice Exercise 3: In which of the following case is the dispersion more and why?





The dispersion is normally measured with the help of variance or standard deviation. The range is defined as the difference between the largest score in the set of data and the smallest score in the set of data, X_L - X_S

Practice Exercise : What is the range of the data 4 8 1 6 6 2 9 3 6 9?

Answer: The largest score (X_L) is 9; the smallest score (X_S) is 1; the range is X_L - X_S = 9 - 1 = 8

Variance is defined as the average of the square deviations. It is represented as

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

Standard deviation = √variance

Variance = standard deviation^2

22.6: RESEARCH REPORT PREPARATION

Depending upon the purpose and types of reports, a research report has identifiable and specific components. The complete report has to be compiled in different chapters with chapter headings. Each chapter must be divided into subheadings or sections. The level of elaborations of headings and subheadings of chapter in a research report depends upon the requirements of technical or business report.

The research report consists of the following broad sections

1. **Preliminary Section-** It consists of a title page, letter of transmittal, a letter of authorization, an executive summary, a table of content, list of table, list of figures & charts, a declaration and an acknowledgment.
2. **Introductory Section-** It consists of problem statement, research objectives, introduction and scope, contribution of the research work etc.
3. **Background Section-** It consists of review of literatures, theoretical framework, preliminary results of explorations from experience survey etc.
4. **Methodology Section**- It consists of research setting and design, sampling design, data collection methods & tools, method used to analyse the data, etc
5. **Findings-** It consists compilation of interpretations and findings after analysis of the data. It explains the facts interpreted after data analysis
6. **Conclusion-** It represents the inferences after analysis and findings. The conclusion may be presented in tabular form or in summary statements.
7. **Bibliography-** In this section the sources of secondary data and literature reviews are compiled in proper citation, style and format.
8. **Appendices-** In this section supporting documents, statistical tests, complex tables, forms and questionnaires are attached.

There are, however, different interpretations of what a report should look like. Whatever be the purpose of research, a research reports must always be accurate, focused and well structured.

Specifically, a research report consists of the compilation of the following

- Introduction and background of the subject or research question
- Details on the objectives & scopes of study in research work
- Literature review and theoretical framework to get insights for conducting the research
- Details of the research methodology followed to carry out the research
- Details of sampling and data collection methods used in the research
- Logical descriptions and analysis of data

- Interpretation, findings, and conclusions after data analysis
- Limitations faced by the researcher during research work
- Suggestions & recommendations to the indented users of the research report
- References and bibliographies
- Annexure

While preparing a research report, the researcher should have considerations for the following

- The purpose of the research work must be clearly defined and detail in the report.
- The objectives and scope must be detailed in the report
- The expectations of the users of the research report must be taken care of.
- The report must be well organized.
- The report must be properly formatted.

Researchers differ in the way they prepare a research report. The personality, background, expertise, and responsibility of the researcher, along with the decision maker (DM) to whom the report is addressed, interact to give each report a unique character. Yet, there are guidelines for formatting and writing reports and designing tables and graphs (Malhotra, 1999). The following is intended as a guideline from which the researcher can develop a format for the research project

Most research reports include the following elements:

Title page

II. Letter of transmittal

III. Letter of authorization

IV. Table of contents

V. List of tables

VI. List of graphs

VII. List of appendices

VIII. List of exhibits

IX. Executive summary

Major findings

Conclusions

Recommendations

X. Problem definition

Background to the problem

Statement of the problem

XI. Approach to the problem

XII. Research design

Type of research design

Information needs

Data collection from secondary sources

Data collection from primary sources

Scaling techniques

Questionnaire development and pretesting

Sampling techniques

Field work

XIII. Data analysis

Methodology

Plan of data analysis

XIV. Results

XV. Limitations and caveats

XVI. Conclusions and recommendations

XVII. Exhibits

Questionnaires and forms

Statistical output

Lists

This format is the standard format of any research report. The results may be presented in several chapters of the report. For example, in a national survey, data analysis may be conducted for the overall sample, and then the data for each of the four geographic regions may be analyzed separately. If so, the results may be presented in five chapters instead of one chapter. In further section, we will understand the details of each section of the research report.

Title Page. The title page should include the title of the report, information (name, address, and telephone) about the researcher or organization conducting the research, the name of the client for whom the report was prepared, and the date of release.

Letter of Transmittal. A formal report generally contains a letter of transmittal that delivers the report to the client and summarizes the researcher's overall experience with the project, without mentioning the findings. The letter should also identify the need for further action on the part of the client, such as implementation of the findings or further research that should be undertaken. **Letter of Authorization.** A letter of authorization is written by the client to the researcher before work on the project begins. It authorizes the researcher to proceed with the project and specifies its scope and the terms of the contract. Often, it is sufficient to refer to the letter of authorization in the letter of transmittal. However, sometimes it is necessary to include a copy of the letter of authorization in the report.

Table of Contents. The table of contents should list the topics covered and the appropriate page numbers. In most reports, only the major headings and subheadings are included. The table of contents is followed by a list of tables, list of graphs, list of appendices, and list of exhibits. **Executive Summary.** The executive summary is an extremely important part of the report, as this is often the only portion of the report that executives read. The summary should concisely describe the problem, approach, and research design that was adopted. A summary section should be devoted to the major results, conclusions, and recommendations. The executive summary should be written after the rest of the report.

Problem Definition. This section of the report gives the background to the problem, highlights the discussions with the decision makers and industry experts, discusses the secondary data analysis, the qualitative research that was conducted, and the factors that were considered.

Moreover, it should contain a clear statement of the management decision problem and the marketing research problem

Approach to the Problem. This section should discuss the broad approach that was adopted in addressing the problem. This section should also contain a description of the theoretical foundations that guided the research, any analytical models formulated, research questions, hypotheses, and the factors that influenced the research design.
Research Design. The section on research design should specify the details of how the research was conducted. This should include the nature of the research design adopted, information needed, data collection from secondary and primary sources, scaling techniques, questionnaire development and pre-testing, sampling techniques, and field work. These topics should be presented in a non-technical, easy-to-understand manner. The technical details should be included in an appendix. This section of the report should justify the specific methods selected.

Data Analysis. This section should describe the plan of data analysis and justify the data analysis strategy and techniques used. The techniques used for analysis should be described in simple, non-technical terms.

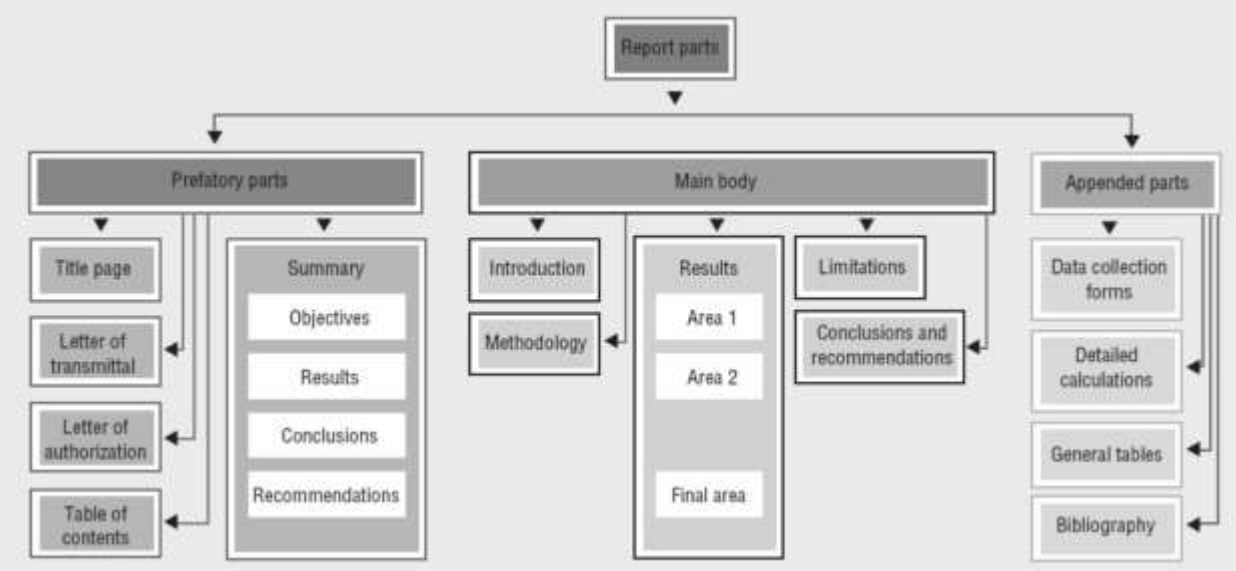
Results. This section is normally the longest part of the report and may comprise several chapters. Often, the results are presented not only at the aggregate level but also at the subgroup (market segment, geographical area, etc.) level. The results should be organized in a coherent and logical way. For example, in a health care marketing survey of hospitals, the results were presented in four chapters. One chapter presented the overall results, another examined the differences between geographical regions, a third presented the differences between for-profit and nonprofit hospitals, and a fourth presented the differences according to bed capacity. The presentation of the results should be geared directly to the components of the marketing research problem and the information needs that were identified. The details should be presented in tables and graphs, with the main findings discussed in the text.

Limitations and **Caveats.** All research projects have limitations caused by time, budget, and other organizational constraints. Furthermore, the adopted research design may be limited in terms of the various types of errors and some of these may be serious enough to warrant discussion. This section should be written with great care and a balanced perspective. On the one hand, the

researcher must make sure that management does not overly rely on the results or use them for unintended purposes, such as projecting them to unintended populations. On the other hand, this section should not erode their confidence in the research or unduly minimize its importance.

Conclusions and Recommendations. Presenting a mere summary of the statistical results is not enough. The researcher should interpret the results in light of the problem being addressed to arrive at major conclusions. Based on the results and conclusions, the researcher may make recommendations to the decision makers. Sometimes researchers are not asked to make recommendations because they research only one area, but do not understand the bigger picture at the client firm. If recommendations are made, they should be feasible, practical, actionable, and directly usable as inputs into managerial decision making. Research in Practice contains guidelines on conclusions and recommendations.

The major part of a research report could be depicted with the help of the following diagram



While writing the research report following considerations must be taken into account

- A report should be written for a specific reader or readers
- The report should take into account the readers' technical sophistication and interest in the project, as well as the circumstances under which they will read the report and how they will use it.

- Technical terminology and jargon should be avoided. As expressed by one expert, "The readers of your reports are busy people, and very few of them can balance a research report, a cup of coffee, and a dictionary at one time." If some technical terms cannot be avoided, briefly define them in an appendix.
- Often the researcher must cater to the needs of several audiences with different levels of technical sophistication and interest in the project. Such conflicting needs may be met by including different sections in the report for different readers or preparing entirely separate reports.

22.7: TYPES OF RESEARCH REPORT

There are various ways by which research report could be classified. Schiendler & Cooper suggested following types of research report.

- **Short Report** – It is also called as informal report. The short report may range from a short statement of facts presented on a single page to a longer presentation taking several pages. The short, informal, report is usually submitted in the form of a letter or memorandum. It does not carry a cover, table of contents or any special display. In style, this short report is personal, informal and relaxed. It is written in the first person, unlike the formal report in which the use of first person is usually for the sake of complete objectivity. This types of report is appropriate when the problem is well defined with limited scope and well structured simple methodology. A letter may be considered as a form of short report and its tone has to be informal
- **Long Report-** It is normally formal in nature. The report is always a long one and consists of all or only some of these parts: cover, title page, contents page, and letter of transmittal (covering letter), summary, introduction, and the body of report, conclusion with or without recommendations, appendix, bibliography, and index. It is even printed sometime and bound in hard covers like a book. When it happens to be very long, a summary of its main points is given after the introduction.
  In style, the long or formal report is impersonal and restrained in tone. The writer or writers

generally do not use the first person (I or WE), but used third-person reference in some such ways. "It was found" and "the writers are of the opinion" etc.

The long reports are of following two types

- **Technical Report-** These types of research report is prepared for academicians or researchers. It consist the full documentation in details. The analysis and interpretation in this report are comprehensive.
- **Management Report-** It is also called as business report and prepared for those who are not supposed to be aware about the research methodology. These types of report specifically focuses on specific problem and less concerned with methodology. It is prepared in such a way that it encourages quick readings with findings and recommendations.

22.8: CHARACTERISTICS OF A GOOD REPORT

A good research report must meet following criteria in its presentation in written.

- Information collected in the report must be **relevant and focused** to derive desired results
- Report should follow the exact **predefined goals and objectives**
- The report should always contain the executive **summary of the work**
- It should also contain the **methodology of the research**
- The report should contain the **description of the questionnaires**
- The report should be **flexible** enough to be changed accordingly
- Clear Information has to be understood at the first reading. The report has to be **easy to read** with legible writing and a clear message
- It must be **complete and concise** also
- It must be **correct** for every piece of information and **verifiable.**

22.9: WRITING FINDINGS AND CONCLUSIONS

The purpose of a conclusion is to tie together, or integrate the various issues, research, etc., covered in the body of the report, and to make comments upon the meaning of all of it. This includes noting any implications resulting from your discussion of the topic, as well as recommendations, forecasting future trends, and the need for further research.

The conclusion should:

- Be a logical ending to what has been previously discussed. It must pull together all of the parts of your argument and refer the reader back to the focus you have outlined in your introduction and to the central topic. This gives your essay a sense of unity.
- Never contains any new information.
- Usually be only a paragraph in length, but in an extended essay (3000+ words) it may be better to have two or three paragraphs to pull together the different parts of the essay.
- Add to the overall quality and impact of the essay. This is your final statement about this topic; thus it can make a great impact on the reader.

The conclusion may include:

- A summary of the arguments presented in the body and how these relate to the essay question
- A restatement of the main point of view presented in the introduction in response to the topic
- The implications of this view or what might happen as a result.

It begins with a sentence that refers to the main subject that was discussed in the body in the report. It is necessary to make sure that these sentences also link to the preceding paragraph. It may be supported by a brief summary of your argument and identify the main reasons/causes/factors that relate to the question you have been asked to address. If there are two or more parts to the question, be sure to include responses to each part in your conclusion. Finally, it is a good idea to add one or two sentences to reinforce the conclusive statements which was used in the introduction. This shows the reader that you have done what you said you would do and gives a sense of unity in writing. Additional elements that may be added include recommendations for future action and speculations on future trends. Generally, although a short pithy quote can sometimes be used to spice up your conclusion, the conclusion should be in your own words. Try to avoid direct quotations, or references to other sources.

22.11: WRITING SUGGESTIONS, RECOMMENDATIONS AND LIMITATIONS

Recommendations regarding actions that should be taken or considered in light of the research results are also considered as an important aspect of a research report. It is subjective statements that helps the managers to decide about the suggested course of action to be taken based on findings of the research work. The recommendation may be related to add or drop a product, advertising positioning, market segments to select as primary targets, how to price the product etc. While writing suggestions or recommendations, it is necessary to logically validate them with the data and its interpretations. Limitations refer to the constraints in research work that may have created impacts on findings and conclusions.

22.12: WRITING REFERENCES, END NOTES, FOOT NOTES

In research, we have to use secondary sources and literature review to get some insights in the related areas of knowledge. The literature review is a critical look at the existing research that is significant to the work that you are carrying out. In terms of a literature review, "the literature" means the works you consulted to understand and investigate your research problem. Some important sources of literature are-

- **Journal articles-** They are frequently used in literature reviews because they offer a relatively concise, up-to-date format for research, and because all reputable journals are refereed.
- **Books**- Textbooks are unlikely to be useful for including in your literature review as they are intended for teaching, not for research, but they do offer a good starting point from which to find more detailed sources.
- **Conference proceedings**- These can be useful in providing the latest research or research that has not been published.
- **Government/corporate reports-** Many government departments and corporations commission or carry out research. Their published findings can provide a useful source of information, depending on your field of study.
- **Newspapers**: Often newspapers are more helpful as providers of information about recent trends, discoveries, or changes, e.g., announcing changes in government policy, but you should then search for more detailed information in other sources.
- **Theses and dissertations**: These can be useful sources of information.
- **Internet:** It is the fastest-growing source of information now but we have to cautious for filtering the relevant information.

This literature are to be referred in a research report in the proper way. Whenever you have taken something from another author (that is to say, you have taken an author's theory, opinion, idea, example, conclusion, or findings), you must say who you took it from, and where the original can be found. In other words, you must acknowledge and cite your sources. This is important whether or not you use the author's own words.

The followings are reasons to acknowledge the sources of secondary data and findings in your research:

- To show that you have read and understood the research published in your area of interest
- To lend authority to what you are writing;
- To strengthen your argument;
- To support your own ideas;
- To provide details or background to what you are writing;
- To provide interest, and
- To avoid the charge of plagiarism.

If you don't acknowledge sources, you may be accused of plagiarism. **Plagiarism** is the act of using another person's ideas as if they are your own. It's a very serious breach of academic etiquette. Your assignment will be given a failing mark, and in extreme cases, you may be booked under the laws of violation of copyright. It doesn't matter whether the original words or ideas are those of a published writer or those of another student; you must not copy without giving your source.

There is more than one way to acknowledge sources. The commonest systems are the footnoting system and the Author-Date system (often known as the Harvard system). Most research reports prefer the Author-Date system. But whichever system you use, you must follow it consistently.

Sometimes **word bibliography** is also used instead of **references.** Although technically a reference list is not the same as a bibliography, there is usually no difference between them as far as the study is concerned. They are the same thing with different names. Some people call it a reference list, and others call it a bibliography, but the same rules apply. Your reference list/bibliography must provide full and accurate details, as it is how the reader can follow up on your sources. Some examples of standard referencing in a research report are as follows.

1. Dyer, C. (1995). Beginning research in psychology. Oxford: Blackwell.
2. Eysenck, M.W. (1994). Individual differences: Normal and abnormal. Hove, UK: Psychology Press.

3. Lovland, J. (1976). Doing social life: The qualitative study of human interaction in natural settings. New York: Wiley.

Common guidelines for writing references are as follows:

- An entry must consist of author(s), date of publication (full date for daily or weekly publications, year only for others), title details, and publisher details.
- Entries must be in alphabetical order of surname.
- Titles of books and journals should be in italics (or underlined where italic font is not available).
- Titles of books and journals should be in Title Case (i.e. all important words have a capitalised initial letter).
- Titles of articles or chapter headings should be in Sentence case (i.e. only the first word or proper nouns should have a capital).
- Book title must include edition (other than first) and any other details given on the title page (eg, series, translator, original title).
- Journal title must include volume, number, and page numbers of the article.

Within these standard conventions of referencing, there are many style differences:

- Date may or may not be in brackets.
- Punctuation between items may differ.
- Article titles may or may not use inverted commas.
- First names may be given in full or reduced to initials.
- Names of joint authors may be separated by 'and' or by an ampersand (&).

These are only a few of the stylistic differences that individuals or organisations may choose while citing references in the research report.

22.13: COMPUTER APPLICATIONS IN RESEARCH

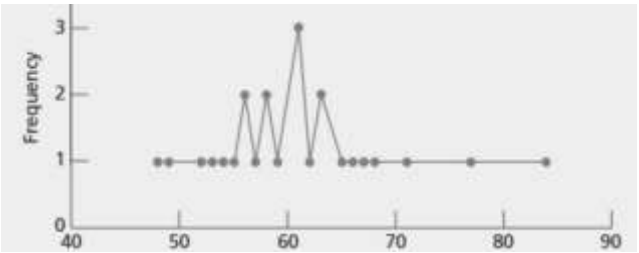
Use of computer in research is so extensive that it is difficult to conceive a scientific research project without computer. Many research studies cannot be carried out without use of a computer particularly those involving complex computations, data analysis and modelling. Computer in research is used at all stages of study-from the proposal/budget stage to the submission/presentation of findings.

We can summarize the uses of computers and information technology in research as follows-

- Computers help in searching and storing information and kinds of literature
- Several software programs are available for statistical analysis and sample size determination
- It helps in the preparation of professional reports by using various tools and software

The computer provides various tools and software for the graphical presentation of data as charts and graphs, as well as for statistical analysis. Some important charts that are generated by computer software and used in research are discussed here:

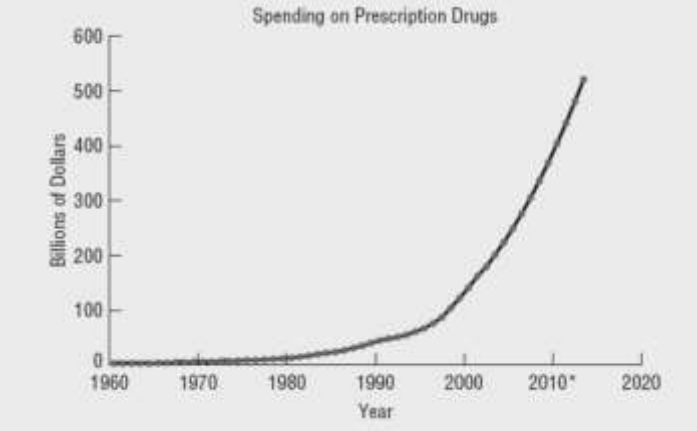
Frequency polygon:



One way of summarizing these data is in the form of a frequency polygon. This is a simple form of chart in which the scores from low to high are indicated on the x-axis or horizontal axis, and the frequencies of the various scores (in terms of the numbers of individuals obtaining each score) are indicated on the y-axis or vertical axis.

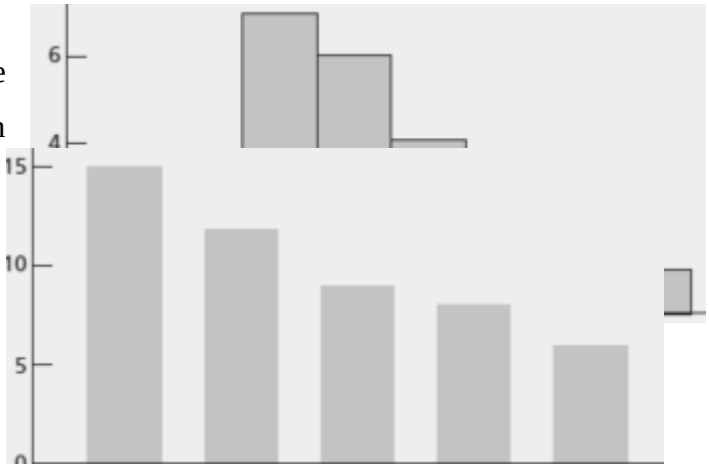
Line graphs:

Line graphs are useful for showing the relationship of one variable to another. The dependent variable generally is shown on the vertical axis, and the independent variable on the horizontal axis. The most common independent variable for such charts is time, but it is by no means the only one.



Histogram:

In a histogram, the scores are indicated on the horizontal axis and the frequencies are shown on the vertical axis. In contrast to a frequency polygon, however, the frequencies are indicated by rectangular columns. These columns are all the same width but vary in height in accordance with the corresponding frequencies.



In a bar chart, the categories are shown along the horizontal axis, and the frequencies are indicated on the vertical axis. In contrast to the data contained in histograms, the categories in bar charts cannot be ordered numerically in a meaningful way. However, they can be arranged in ascending (or descending) order of popularity. Another difference from histograms is that the rectangles in a bar chart do not usually touch each other.

SPSS and MS EXCEL in Research

SPSS is a statistical package used for statistical analysis of the data. SPSS (Statistical Package for the Social Sciences) is a computer software that provides tools for statistical analysis of data. It allows for in-depth data access and preparation, analytical reporting, graphics and modelling. SPSS has numbers of statistical and mathematical functions, numbers of statistical procedures, and a

very flexible data handling capability. It can read data in almost any format (e.g., numeric, alphanumeric, binary, dollar, date, time formats), and can read files created using spread sheet/data base software. It also has excellent data manipulation utilities.

The following is a brief overview of some of the functionalities of SPSS:

- Data transformations
- Data Examination
- Descriptive Statistics
- Contingency tables
- Reliability tests
- Correlation
- T-tests
- ANOVA
- MANOVA
- General Linear Model (Release 7.0 and higher)
- Regression
- Nonlinear Regression
- Logistic Regression
- Loglinear Regression
- Discriminant Analysis
- Factor Analysis
- Cluster anlaysis
- Multidimensional scaling
- Probit analysis

- Forecasting/Time Series
- Survival analysis
- Nonparametric analysis
- Graphics and graphical interface.

MS Excel also provides various tools for generation of graphs and charts on the basis of data. It also provides tools for statistical analysis of the data. Microsoft Excel is a spreadsheet application developed by Microsoft for Microsoft Windows and Mac OS X. It features calculation, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications.

Practice Exercise: On a computer system learn the uses of MS Excel software. Try to create different types of graphs and charts. Learn to use various statistical functions and Data analytic tools of MS Excel.

22.13: ETHICS IN RESEARCH REPORT WRITING

Report preparation and presentation involve many issues pertaining to research integrity. These issues include defining the research problem to suit hidden agendas, compromising the research design, deliberately misusing statistics, falsifying figures, altering research results, misinterpreting the results to support a personal or corporate point of view, and withholding information. The researcher must address these issues while preparing the report and presenting the findings. The dissemination of the research results to the client and other stakeholders, as appropriate, should be honest, accurate, and complete.
The researcher should be objective throughout all phases of the marketing research process. Some research procedures and analyses may not reveal anything new or significant.

Ethical dilemmas can arise in these instances if the researcher nevertheless attempts to conclude from wrong analyses. Such temptations must be resisted to avoid unethical conduct. Likewise, clients also have the responsibility for complete and accurate disclosure of the research findings and are obligated to use the research results ethically.

22.14: SUMMARY

In this section, we learned about the objectives of research report preparation and presentation. The uses of computers are also discussed, along with the functionality of relevant software. The format of a research report was described in terms of its major sections

22.15: GLOSSARY

Plagiarism: Copying the secondary data without acknowledging the sources

SPSS: A software for statistical analysis

Classification: Grouping of data based on homogeneity

Mean: A measure of central tendency

Standard Deviation: A measure of dispersion

22.16: CHECK YOUR PROGRESS

Q1: What preparations are required before we start writing a research report?

a. The data must be analyzed and understood
b. The researcher must understand the conclusion
c. The framework for compiling the research report is known
d. All of these

Q2: If we classify the population based on gender as Male and Female, then it comes under which of the following classifications?

a. Multiple Classifications
b. Dichotomous Classifications
c. Both
d. None of these

Q3: What is the basic criterion for the classification of data?

a. Heterogeneity

b. Homogeneity

c. Consistency

d. Verifiability

Q4: Which of the following is NOT a measure of central tendency of the data?

a. Mean

b. Median

c. Mode

d. Standard Deviation

Q5: In which of the following sections of the research report should we add the questionnaire of the research work?

a. Introduction

b. Appendices

c. Data Analysis

d. Conclusions

Q6: What do you understand by the word "verifiable" as a characteristic of a good research report?

a. The sources of secondary data could be checked

b. The research design mentioned in the research report could be reproduced

c. Both a and b

d. None of these

Q7: Which of the following should be avoided while writing the conclusion?

a. Quotations

b. References to other sources

c. Deviation from the intended task of research

d. All of the above

Q8: Which of the following is considered a serious breach of academic etiquette?

a. Uses of Secondary data with acknowledging the source

b. Plagiarism

c. Collection of Primary Data

d. None of the above

22.17: ANSWER TO CHECK YOUR PROGRESS

Q1- d, Q2- c, Q3 – a, Q4-d, Q5-b, Q6-c, Q7-d, Q8-b

22.18: TERMINAL QUESTIONS

Q1. What do you understand by the interpretation and analysis of data? What tools are used?

Q2. What do you mean by classification and tabulation? How does it help in managerial decisions?

Q3. What do you understand by quantitative and qualitative data analysis? Discuss with examples.

Q4. What are the major sections of a research report? Discuss

Q5. What should be included in the Appendices part of a research report? What are its uses?

Q6. What ethics are to be followed in preparing a research report?

Q7. What are the uses of computers in research? Discuss in detail.

22.19: SUGGESTED READINGS

1.Cresswell, John, W.,(2008). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Newbury Park, CA: Sage Publication.

2.Marczyk, G.R, DeMatteo, D. & Festinger D., (2005). Essentials of Research Design and Methodology, New York City, NY: Wiley.

3. Ethridge, Don E. (2004). Research Methodology in Applied Economics. Daryaganj, ND: Wiley – Blackwell,

4. Bergh, D. and Ketchen, D. (2009) Research Methodology in Strategy and Management. Bingley, UK: Emarald Group Publishing.

5. Research Methodology: C R Kothari (New Age)

6. Marketing Research: N K Malhotra (Pearson's)