Directorate of Distance Education

UNIVERSITY OF JAMMU
JAMMU

SELF LEARNING MATERIAL
OF
BUSINESS STATISTICS
FOR
M.COM IST SEMESTER

For the examination to be held in 2023 onwards

Course No. : MCOM C154 **UNIT : I - IV**
LESSON NO. 1-20

Course Coordinator:
Prof. Sandeep Kour Tandon
Room No. 111, Ist Floor
DDE, University of Jammu.

http:/www.distanceeducationju.in
Printed and published on behalf of Directorate of Distance Education, University of Jammu, Jammu by the Director, DDE, University of Jammu, Jammu

BUSINESS STATISTICS

Wrtten & Reviewed by :

Dr. Ritika Sambyal
Department of Commerce
Udhampur Campus
University of Jammu

Proof Reading by :

Ms. Shriya Gupta
Teacher-in-Charge, M.Com
Room No. 205, IInd Floor,
DDE, University of Jammu

Printed by : Jandiyal Printing Press /2022 / 900

BUSINESS STATISTICS
M.COM SEMESTER - I (NON CBCS)
(CORE COURSES)

DETAILED SYLLABUS

Course No. : MCOM C154 **Maximum Marks : 100 Marks**
Credits : 4 **External : 80 Marks**
Time : 3.00 Hrs. **Internal : 20 Marks**

Syllabus for the Examination to be held in December 2023 onwards

COURSE OBJECTIVES

1. To discuss the role of business statistics.
2. To understand the relevance of' probability distributions and ANOVA.
3. To discuss the various methods of association of attributes and multiple regression equations.
4. To know the procedure of testing of hypothesis.

COURSE OUTCOMES

After the completion of this course, the students will be able to:

1. understand the relevance of statistics in business operations;
2. apply the probability and ANOVA for solving business problems;
3. able to use various methods of association of attributes, multiple correlation and regression analysis,
4. compute parametric and non-parametric tests for hypothesis testing; and
5. analyse the complex set of data through various statistical techniques for solving business problems.

UNIT I OVERVIEW OF SAMPLING

Concept: Features, functions and role of business statistics; Sampling: concept, need, essentials, principles and process of sampling; probability and non probability sampling techniques, sampling errors vs. non-sampling errors; effectiveness of Sampling. ; Determination of sample size; Sample distribution and standard error; Pilot and final Survey; Precautions in data collection

UNIT II PROBABILITY AND ANALYSIS OF VARIANCE

Concept and role of probability; Approaches of Probability: Classical, relative frequency; Subjective and axiomatic; Addition and multiplication theorem; Mathematical Expectation;

Normal distribution: Concept, importance properties and constants; binomial distribution: meaning, relevance, properties, constants; Poisson distributions: Meaning, co constants and applications; Analysis of variance: Concept; Assumptions, one way and two way classifications.

UNIT III ASSOCIATION OF ATTRIBUTES

Concept and terminology of association of attributes; Consistency of data; Association and disassociation; Methods of attributes: Comparison method, proportion method, Yule's coefficient of association, coefficient of colligation, coefficient of contingency; Partial correlation: Meaning. correlation: uses. limitations and computation of coefficients of partial correlation; Multiple correlation : Concept, advantages, disadvantages, computation of coefficients of multiple correlation; Concept and relevance of multiple regression, Computation of multiple regression equations, Shortcomings of regression analysis.

UNIT IV HYPOTHESIS TESTING

Concept, types and procedure of setting hypothesis; Type 1 and type II errors. Difference between Parametric and Nonparametric tests; Test of Significance for large samples; t - test - one sample t test, independent sample t test, dependent samples t test; Chi square test: Uses, steps and computation of chi square, Mann Whitney test and Kruskal Wallis test; Advantages and disadvantages of non-parametric tests.

SUGGESTIVE READINGS

1. Levin, R.I. Robin, D.S. Statistics for Management, Prentice-Hall of India. New Delhi.
2. Aczel, A. D. Sounderpandian, J. Complete Business Statistics, Mc Graw Hill Publishing. New Delhi.
3. Anderson, S., W. Statistics for Business and Economics, Cengage Learning. New Delhi.
4. Kazmeir L. J. Business Statistics, Tata Mc Graw Hill. New Delhi.
5. Vohra, N. D. Business Statistics. Tata Mc Graw Hill. New Delhi.
6. Freund, J. E., Williams, F.M. Elementary Business Statistics — The Modern Approach. Prentice Hall of India Private Ltd., New Delhi.
7. Clave, B. S. Statistics for Business and Economics - Prentice Hall Publication, New Delhi.

Note : Latest edition of the books may be preferred.

NOTE FOR PAPER SETTING

The paper consists of two sections. Each section will cover the whole of the syllabus without repeating the question in the entire paper.

Section A: It will consist of eight short answer questions, selecting, two from each unit. A candidate has to attempt any six and answer to each question shall be within 200 words. Each question carries four marks and total weightage to this section shall be 24 marks.

Section B: It will consist of six essay type questions with answer to each question within 800 words. One question will be set atleast from each unit and the candidate has to attempt four. Each question will carry 14 marks and total weightage shall be 56 marks.

Contents

OVERVIEW OF SAMPLING
CONCEPT, FEATURES, FUNCTIONS AND ROLE OF BUSINESS STATISTICS

STRUCTURE

1.1 INTRODUCTION

Business statistics like many areas of study has its own language. It is important to begin our study with an introduction of some basic concepts in order to understand and communicate about the subject. We begin with a discussion of the word statistics. The word statistics has many different meanings in our culture. Webster's Third New International Dictionary gives a comprehensive definition of statistics as a "science dealing with the collection, analysis, interpretation, and presentation of numerical data". Also, the study of business statistics is important, valuable, and interesting. However, because it involves a new language of terms, symbols, logic, and application of mathematics, it can be at times overwhelming.

The study of statistics can be organised in a variety of ways. One of the main ways is to subdivide statistics into two branches: descriptive statistics and inferential statistics. If a business analyst is using data gathered on a group to describe or reach conclusions about that same group, the statistics are called descriptive statistics. For example, if an instructor produces statistics to summarise a class's examination effort and uses those statistics to reach conclusions about that class only, the statistics are descriptive. Many of the statistical data generated by businesses are descriptive. They might include number of employees on vacation during June, average salary at the Denver office, corporate sales for 2009, average managerial satisfaction score on a company-wide census of employee attitudes, and average return on investment for the Lofton Company for the years 1990 through 2008.

Another type of statistics is called inferential statistics. If a researcher gathers data from a sample and uses the statistics generated to reach conclusions about the population from which the sample was taken, the statistics are inferential statistics. The data gathered from the sample are used to infer something about a larger group. Inferential statistics are sometimes referred to as inductive statistics. The use and importance of inferential statistics continue to grow. One application of inferential statistics is in pharmaceutical research. Some new drugs are expensive to produce, and therefore tests must be limited to small samples of patients. Utilising inferential statistics, researchers can design experiments with small randomly selected samples of patients and attempt to reach conclusions and make inferences about the population.

Market researchers use inferential statistics to study the impact of advertising on various market segments. Suppose a soft drink company creates an advertisement depicting a dispensing machine that talks to the buyer, and market researchers want to measure the impact of the new advertisement on various age groups. The researcher could stratify the population into age categories ranging from young to old, randomly sample each stratum, and use inferential statistics to determine the effectiveness of the advertisement for the various age groups in the population. The advantage of using inferential statistics is that they enable the researcher to study effectively a wide range of phenomena without having to conduct a census.

1.2 OBJECTIVES

After going through this lesson, you would be able to:

- Define the concept of business statistics.
- Explain the role and importance of business statistics.
- Elaborate the essential features of business statistics.

1.3 CONCEPT OF BUSINESS STATISTICS

In the beginning, it may be noted that the word 'statistics' is used rather curiously in two senses plural and singular. In the plural sense, it refers to a set of figures or data. In the singular sense, statistics refers to the whole body of tools that are used to collect data, organise and interpret them and, finally, to draw conclusions from them. It should be noted that both the aspects of statistics are important if the quantitative data are to serve their purpose. If statistics, as a subject, is inadequate and consists of poor methodology, we could not know the right procedure to extract from the data the information they contain. Similarly, if our data are defective or that they are inadequate or inaccurate, we could not reach the right conclusions even though our subject is well developed.

A.L. Bowley has defined statistics as: (i) statistics is the science of counting, (ii) Statistics may rightly be called the science of averages, and (iii) statistics is the science of measurement of social organism regarded as a whole in all its manifestations. Boddington defined as:

Statistics is the science of estimates and probabilities. Further, **W.I. King** has defined Statistics in a wider context, the science of Statistics is the method of judging collective, natural or social phenomena from the results obtained by the analysis or enumeration or collection of estimates. Seligman explored that statistics is a science that deals with the methods of collecting, classifying, presenting, comparing and interpreting numerical data collected to throw some light on any sphere of enquiry. **Spiegal** defines statistics highlighting its role in decision-making particularly under uncertainty, as follows: statistics is concerned with scientific method for collecting, organising, summa rising, presenting and analyzing data as well as drawing valid conclusions and making reasonable decisions on the basis of such analysis. According to Prof. **Horace Secrist**, Statistics is the aggregate of facts, affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a pre-determined purpose, and placed in relation to each other.

1.3.1 Features of Business Statistics

(i) Statistics are the aggregates of facts. It means a single figure is not statistics. For example, national income of a country for a single year is not statistics but the same for two or more years is statistics**.**

(ii) Statistics are affected by a number of factors. For example, sale of a product depends on a number of factors such as its price, quality, competition, the income of the consumers, and so on.

(iii) Statistics must be reasonably accurate. Wrong figures, if analysed, will lead to erroneous conclusions. Hence, it is necessary that conclusions must be based on accurate figures.

(iv) Statistics must be collected in a systematic manner. If data are collected in a haphazard manner, they will not be reliable and will lead to misleading conclusions.

(v) Collected in a systematic manner for a pre-determined purpose

(vi) Lastly, Statistics should be placed in relation to each other. If one collects data unrelated to each other, then such data will be confusing and will not lead to any logical conclusions. Data should be comparable over time and over space**.**

Most of the information around us is determined with help of statistics; e.g., weather forecasts, medical studies, quality testing, stock markets etc. Therefore, Business Statistics involves the application of statistical tools in the area of marketing, production, finance, research and development, manpower planning etc. to extract relevant information for the purpose of decision making. Business managers use statistical tools and techniques to explore almost all areas or business operations of public and private enterprises.

1.3.2 Functions of Statistics

The functions of statistics are as follows:

1. **Statistics simplifies complexity:** Statistic consists of aggregate of numerical facts. Huge facts and figures are difficult to remember. The complex mass of figures can be made simple and understandable with the help of statistical methods. Statistical techniques such as averages, dispersion, graph, diagram etc. make huge mass of figures easily understandable. So, the function of statistics is to reduce the complexity of the huge mass of figures to a simpler form.

2. **Statistics presents fact in a definite form:** One of the important functions of statistics is to present the general statements in a precise and definite form. The conclusion stated numerically is definite and hence more convincing than the conclusions stated qualitatively. This fact can readily be understood by the following example: "The population of Nepal is 1981 has been increased than in 1971". There will be no clear idea about this statement. Everybody wants to know to what extent the population of Nepal has increased. But the statement that "the population of Nepal has increased from 11555983 in 1971 to 15022839 in 1981" is a definite form.

3. **Statistics facilities comparison:** The science of statistics does not mean only counting but also comparison. Unless the figures are compared with other figures with the same kind, they are meaningless. Statistical methods such as averages, ratios, percentages, rates, coefficients etc. offer the best way of comparison between two phenomena which will enable to draw valid conclusion. So, statistics helps in the comparison of two phenomena. For example: The statement that "the per capita

income of Nepal is $160" is not so clear unless it is compared with the per capita income of any other country.

4. **To help in formulation of policies:** Statistics helps in formulating the policies in different fields mainly in economics, business etc. The government policies are also framed on the basis of statistics. In fact, without statistics, suitable policies cannot be framed. For example: The quantity of food grains to be imported in a particular year depends upon the expected internal production and the expected consumption. That is if the expected wheat production in the particular year be 701 thousands metric tons and that of consumption 710 thousand metric tons so we must import 9 thousand metric tons of food grains.

5. **Statistics helps in forecasting:** While preparing suitable policies and plans, it is necessary to have the knowledge of future tendency. This is mostly in case of industry, commerce and so on. Statistical methods provide helpful means in forecasting the future by studying and analyzing the tendencies based on passed records. For example: Suppose a businessman wants to know the expected sales of T.V. for the next year, the better method for him would be to analyze the sales data of the past years for the estimation of the sales volume for the next year.

6. **Statistics helps in formulating and testing hypothesis:** Statistical methods are helpful not only in estimating the present forecasting the future but also helpful in formulating and testing the hypothesis for the development of new theories. Hypothesis like 'whether a particular fertilizer is effective for the production of a particular commodity' 'whether a dice is biased or not' can be tested with the help of statistical tools.

1.3.3 Role of Business Statistics

Statistics play an important role in business. A successful businessman must be very quick and accurate in decision making. He knows that what his customers wants, he should therefore, know what to produce and sell and in what quantities. Statistics helps businessman to plan production according to the taste of the customers, the quality of the products can also be checked more efficiently by using statistical methods. Hence, all the activities of

businessman based on statistical information. He can make correct decision about the location of business, marketing of the products, financial resources etc.

1. In Business – It helps to make swift decisions by providing useful information about customer trends and variations, cost customer trends and variations, price customer trends and variations etc.
2. In Mathematics – It helps in describing measurements and providing accuracy of theories.
3. In Economics – It helps to find relationship between two variables like demand and supply, cost and revenue, imports and exports and helps to establish relationship between inflation rate, per capita income, income distribution etc.
4. In Accounts – It helps to discover trends and create projections for next year.
5. In Physics – It helps to compute distance between objects in space.
6. Research – It helps in formulating and testing hypothesis.
7. Government – Government takes help of statistics to make budgets, set minimum wages, estimate cost of living etc.

1.4 IMPORTANCE OF BUSINESS STATISTICS

These days statistical methods are applicable everywhere. There is no field of work in which statistical methods are not applied. According to A .L. Bowley, "knowledge of statistics is like a knowledge of foreign languages or of Algebra, it may prove of use at any time under any circumstances". The importance of the statistical science is increasing in almost all spheres of knowledge, e g., astronomy, biology, meteorology, demography, economics and mathematics.

Economic planning without statistics is bound to be baseless. Statistics serve in administration, and facilitate the work of formulation of new policies. Financial institutions and investors utilise statistical data to summaries the past experience. Statistics are also helpful to an auditor, when he uses sampling techniques or test checking to audit the accounts of his client. The importance of business statistics can be summarised through the following points:

1. Deal with uncertainties by forecasting seasonal, cyclic and general economic fluctuations.
2. Helps in Sound Decision making by providing accurate estimates about costs, demand, prices, sales etc.
3. Helps in business planning on the basis of sound predictions and assumptions.
4. Helps in measuring variations in performance of products, employees, business units etc.
5. It allows comparison of two or more products, business units, sales teams etc. Helps in identifying relationship between various variables and their effect on each other like effect of advertisement on sales.
6. Helps in validating generalisations and theoretical concepts formulated by managers.

1.5 LIMITATIONS OF BUSINESS STATISTICS

The scope of the science of statistic is restricted by certain limitations:

1. **Statistics deals only with quantitative characteristics**: Statistics are numerical statements of facts. Data which cannot be expressed in numbers are incapable of statistical analysis. Qualitative characteristics like honesty, efficiency, intelligence etc. cannot be studied directly.
2. **Statistics deals with aggregates not with individuals:** Since statistics deals with aggregates of facts, the study of individual measurements lies outside the scope of statistics.
3. **Statistical laws are not perfectly accurate:** Statistics deals with such characteristics which are affected by multiplicity of causes and it is not possible to study the effect of these factors. Due to this limitation, the results obtained are not perfectly accurate but only an approximation.
4. **Statistical results are only an average:** Statistical results reveal only the average behaviour. The Conclusions obtained statistically are not universally true but they are true only under certain conditions.

5. **Statistics is only one of the methods of studying a problem:** Statistical tools do not provide the best solution under all circumstances.

6. **Statistics can be misused:** The greatest limitation of statistics is that they are liable to be misused. The data placed to an inexperienced person may reveal wrong results. Only persons having fundamental knowledge of statistical methods can handle the data properly.

1.6 SUMMARY

Statistics play an important role in business. A successful businessman must be very quick and accurate in decision making. So, all the activities of the businessman based on statistical information. He can make correct decision about the location of business, marketing of the products, financial resources etc.

Statistical studies are extremely important in our everyday life. Statistics are the method of conducting a study about a particular topic by collecting, organising, interpreting, and finally presenting data. Some major areas relying on statistics include government, education, science, and large companies.

1.7 GLOSSARY

- **Business** : Business is the activity of making one's living or making money by producing or buying and selling products (such as goods and services).
- **Statistics:** Statistics is a branch of mathematics dealing with data collection, organization, analysis, interpretation and presentation. In applying statistics to, for example, a scientific, industrial, or social problem, it is conventional to begin with a statistical population or a statistical model process to be studied.
- **Inferential Statistics:** Inferential statistics consists of methods that are used for drawing inferences, or making broad generalisations, about a totality of observations on the basis of knowledge about a part of that totality.
- **Descriptive Statistics:** The term descriptive statistics deals with collecting, summarizing, and simplifying data, which are otherwise quite unwieldy and voluminous.

- **Business Statistics:** Business statistics is the science of good decision making in the face of uncertainty and is used in many disciplines such as financial analysis, econometrics, auditing, production and operations including services improvement and marketing research.
- **Data:** Data are individual pieces of factual information recorded and used for the purpose of analysis. It is the raw information from which statistics are created.

1.8 SELF ASSESSMENT QUESTIONS

A. Multiple Choice Questions:

1. Which of the following is a branch of statistics:

    a. Descriptive
    b. Inferential
    c. Industrial
    d. Both a & b

2. Which of the following values is used as a summary measure for a sample, such as a sample mean?

    a. Sample Statistics
    b. Population Parameters
    c. Sample Parameter
    d. Population Mean

3. Review of performance appraisal, labour turnover rates, planning of incentives, and training programs is the examples of which of the following?

    a. Statistics in Production
    b. Statistics in Marketing
    c. Statistics in Finance
    d. Statistics in Personnel Management

1.9 LESSON END EXERCISE

1. Define business statistics? Explain its functions and importance.

2. Differentiate between descriptive and inferential statistics.

3. What factors contribute to the increasing importance of quantitative approach to management?

4. Discuss various limitations of business statistics.

5. Differentiate between statistical and operations research techniques.

6. Discuss the role of business statistics.

1.10 SUGGESTED READING

- Levin, R.I. Robin, D.S. Statistics for Management, Prentice-Hall of India. New Delhi.
- Aczel, A. D. Sounderpandian, J. Complete Business Statistics, Mc Graw Hill Publishing. New Delhi. downloaded
- Anderson, S., W. Statistics for Business and Economics, Cengage Learning. New Delhi.
- Kazmeir L. J. Business Statistics, Tata Mc Graw Hill. New Delhi.
- Vohra, N. D. Business Statistics. Tata Mc Graw Hill. New Delhi.

CONCEPT, NEED, ESSENTIALS, PRINCIPLES AND PROCESS OF SAMPLING

STRUCTURE

2.1 INTRODUCTION

The way in which we select a sample of individuals to be research participants is critical. How we select participants will determine the population to which we may generalise our

research findings. The procedure that we use for assigning participants to different treatment conditions will determine whether bias exists in our treatment groups? We address the concept of sampling in this lesson. Further, the current lesson explores the process as well as principles and essential features of sampling. Why do researchers often take a sample rather than conduct a census? This lesson addresses these questions about the sampling. Sampling is widely used in business as a means of gathering useful information about a population. Data are gathered from samples and conclusions are drawn about the population as a part of the inferential statistics process.

Therefore, in this lesson, we will discuss concept, need, essential features, principles as well as process of sampling.

2.2 OBJECTIVES

After studying this lesson, you will be able to:

- Define the concept of census and sampling.
- Contrast sampling to census.
- Know the essentials of sampling.
- Explain the principles and process of sampling.

2.3 CENSUS

Sometimes it is preferable to conduct a census of the entire population rather than taking a sample. There are at least two reasons why a business researcher may opt to take a census rather than a sample, providing there is adequate time and money available to conduct such a census:

1) To eliminate the possibility that by chance a randomly selected sample may not be representative of the population.

2) For the safety of the consumer. Even when proper sampling techniques are implemented in a study, there is the possibility that a sample could be selected by chance that does not represent the population. For example, if the population of interest is all truck owners in the state of Colorado, a random sample of truck owners

could yield mostly ranchers when, in fact, many of the truck owners in Colorado are urban dwellers. If the researcher or study sponsor cannot tolerate such a possibility, then taking a census may be the only option. In addition, sometimes a census is taken to protect the safety of the consumer. For example, there are some products, such as airplanes or heart defibrillators, in which the performance of such is so critical to the consumer that 100% of the products are tested, and sampling is not a reasonable option.

Every research study has a target population that consists of the individuals, institutions, or entities that are the object of investigation. The sample is taken from a population list, map, directory, or other source used to represent the population. This list, map, or directory is called the ***frame***, which can be school lists, trade association lists, or even lists sold by list brokers.

Ideally, a one-to-one correspondence exists between the frame units and the population units. In reality, the frame and the target population are often different. For example, suppose the target population is all families living in Detroit. A feasible frame would be the residential pages of the Detroit telephone books. How would the frame differ from the target population? Some families have no telephone. Other families have unlisted numbers. Still other families might have moved and/or changed numbers since the directory was printed. Some families even have multiple listings under different names.

Frames that have over registration contain the target population units plus some additional units. Frames that have under registration contain fewer units than does the target population. Sampling is done from the frame, not the target population. In theory, the target population and the frame are the same. In reality, a business researcher's goal is to minimise the differences between the frame and the target population.

2.4 CONCEPT OF SAMPLING

The U.S. Bureau of the census used it first in 1940. Prior to that recorded instances are relatively few in number. After 1920 sampling began to develop systematically and much of the growth was in agricultural field rather than in social research. In recent years, sampling has become an essential part of research procedure and every researcher required to be

familiar with its logic and some of its important techniques. Sampling is not typical of science only. In a way, we also practice crude versions of sampling in our day to day lines. The house wives, for example, press a few cods of boiled rice to be able to declare that it is ready to be served. Understandably, it is not feasible to examine each and every grain in the cooking pot, thus, instead of studying each and every unit, in sampling method, a small portion is selected which represents the whole population.

A sample is a subset of measurements selected from the population. Sampling from the population is often done randomly, such that every possible sample of n elements will have an equal chance of being selected. A sample selected in this way is called a simple random sample, or just a random sample. A random sample allows chance to determine its elements.

According to P.V. Young, "A statistical sample is a miniature picture or cross section of the entire group or aggregate from which the sample is taken. The entire group from which a sample is chosen is known as the 'population'; "Universe" or 'Supply'.

According to Goode and Hatt, "A sample as the name implies, is a smaller representation of a larger whole."

The idea of sampling is quite old, though the theory of sampling has developed in recent years. Very often a handful of grains of boiling rice is examined to ascertain whether it is cooked or not, and a doctor examines a few drops of blood to ascertain the blood type. Likewise this technique is employed in many other fields.

The main object of sampling technique is to draw conclusions about the whole by examining only a part of it.

2.4.1 Need of Sampling

Sampling is essential and is used in practice for a variety of reasons such as:

1. Sampling saves time, the data can be collected and summarised more quickly with a sample than a complete count of the whole population.
2. In case of infinite population, sampling is the only method for statistical analy-sis.
3. Sampling reduces the cost of experiment because only a few selected items are studied

in sampling.

4. Sampling remains the only choice when a test involves the destruction of the item under study.

5. Sampling usually enables to estimate the sampling errors and, thus, assists in obtaining information concerning some characteristic of the population.

2.4.2 Essentials of Sampling

The choice of a sample as representatives of the whole group is based upon following assumptions:

1) Underlying homogeneity amidst complexity: Although things, especially phenomena, appear to be very complex in nature, so that no two things appear alike, a keener study has disclosed that beneath this apparent diversity there is underlying fundamental unity. Apparently every student may appear to be different. There are differences of health, body, habits, personality etc. But fundamentally they are similar in many respects, so that a study of some of them will throw significant light upon the whole group. It is the possibility of such ideal types in the whole population that makes sampling possible. If no two students were alike in any respect the sampling would have been impossible.

2) Possibility of Representative Election: The second assumption is that it is possible to draw a representative sample. It has been proved that if a certain number of units are selected from a mass on purely random basis, every unit will have a change of being included and the sample so selected will contain all types of units, so that it may be representative of the whole group. This principle is popularly known as the law of statistical regularity and is the very basis of all sampling enquiries.

3) Absolute Accuracy not Essential: The third basic factor is the fact that absolute accuracy is not essential in case of mass study. In large scale studies we have to depend upon averages which are considered fairly significant in any type of enquiry, the result of sampling studies although not hundred percent accurate are nevertheless sufficiently accurate to permit valid generalisations.

4) Independent: All units of a sample must be independent of each other. In other words inclusion of one item in the sample should not be dependent upon the inclusion of other items of the universe.

5) Adequacy: The number of items in the sample should be fairly adequate so that some reliable conclusion can be drawn**.**

2.4.3 Principles of Sampling

1. **Principle of 'Statistical Regularity':** The principle of statistical regularity is derived from the theory of probability in mathematics. According to this principle, when a large number of items is selected at random from the universe, then it is likely to possess the same characteristics as that of the entire population. This principle asserts that the sample selection is random, i.e. every item has an equal and likely chance of being selected. It is believed that sample selected randomly and not deliberately acts as a true representative of the population. Thus, this principle is characterized by the large sample size and the random selection of a representative sample.

2. **Principle of 'Inertia of Large Numbers':** The principle of Inertia of large numbers states that the larger the size of the sample the more accurate the conclusion is likely to be. This principle is based on the notion, that large numbers are more stable in their characteristics than the small numbers, and the variation in the aggregate of large numbers is insignificant. It does not mean that there is no variation in the large numbers, there is, but is less than in the smaller numbers.

3. **Principle of Optimisation:** The principle of optimization takes into account the factors of (a) Efficiency and (b) cost.

   (a) **Efficiency:** Efficiency is measured by the inverse of sampling variance of the estimator. The principle of optimization ensures that a given level of efficiency will be reached with the minimum possible resources and minimum cost.

   (b) **Cost:** Cost is measured by expenditure incurred in terms of money or man powers. So, the term optimization means that, it is based on developing methods of sample selection and of estimation; these provide a given value of cost with the maximum possible efficiency.

4. **Principle of Validity:** By validity of a sample design, we mean that the sample should be so selected that the results could be interpreted objectively in terms of probability. According to this, sampling provides valid estimates about population parameters. This principle ensures that there is some definite and pre-assigned probability for each individual of the aggregate (population) to be included in the sample.

2.5 PROCESS OF SAMPLING

The main steps involved in the process of sampling are:

1. **Objectives:** The objective of the survey must be defined in clear and concrete terms. Generally, in survey a investigation team is not quite clear in mind as to what they want and how they are going to use the results. Some of the objectives may be immediate and some far-reaching. The investigator should take care of these objectives with the available resources in terms of money, manpower and the time limit required for the availability of the survey.

2. **Defining the Population:** The population from which sample is chosen should be defined in clear and unambiguous terms. The geographical, demographic and other boundaries of the population must be specified so that no ambiguity arises regarding the coverage of the survey.

3. **Sampling Frame and Sampling Units:** The sampling unit is the ultimate unit to be sampled for the purpose of the survey. The sampling units must cover the entire population and they must be distinct, unambiguous and non-overlapping in the sense that every element of the population belongs to one and only one sampling unit. In a Socio economic survey, whether a family or a member of a family is to be the ultimate sampling unit. Once the sampling units are defined, one must see whether a sampling frame which is a list of all the units in the population, is available. The construction of the frame is often one of the major practical problems since it is the frame which determines the structure of the sample survey. The list of units have to be carefully scrutinised and examined to ensures that it is free from duplicity or incompleteness and are up-to-date. A good frame is hard to come by and only good experience helps to construct a good frame.

4. **Selection of Proper Sampling Design:** This is the most important step in planning a sample survey. There is a group of sampling designs (to be discussed later) and selection of the proper one is an important task. The design should take into account the available resources and the time-limit, if any, besides the degree of accuracy desired. The cost and precision should also be considered before the final selection of sampling design.

5. **Method of Collection of Data:** For collection of data, either the interview method or the mail questionnaire method is to be adopted. Although the later method is less costly but there is a large scope of non-response in it. In the cases, where the information is to be collected by observation they must decide upon the method of measurement.

6. **Data to be collected:** Collection of data must be done in conformity with the objectives of the survey and the nature of the data. After it is decided upon, one must prepare a questionnaire or a schedule of enquiry. A schedule or a questionnaire contains a list of items of which information is sought, but the exact form of the questions to be asked is not standardized but left to the judgment of the investigators. A questionnaire should be in a specified order. The questions should be clear, brief, collaborative, non offending and unambiguous and to the point so that not much scope of speculation is left on the part of the respondent or interviewer.

7. **Field Work Organisation:** Field work, itself has several stages and so it is to be well organized. The different stages include training the field workers, supervising the field workers, etc. It is absolutely essential that the personnel should be thoroughly trained in locating the sample units, the methods of collection of required data before starting the field word. The success of a survey to a great extent depends upon the reliable field work. Inspection after field work by the adequate supervisors should also be performed.

8. **Summary and Analysis of Data:** This is the last step wherein inference is to be made on the basis of collected data. This step again consists of the following steps: a) The filled in questionnaires should be carefully scrutinized to find out whether the data furnished are plausible and consistent; b) Depending upon the quantity of data, a

hand-tabulation or machine tabulation is to be drawn; c) After the data has been properly scrutinized, edited and tabulated, a very careful statistical analysis is to be made; and d) Finally a report incorporating detailed statement of the different stages of the survey should be prepared. In the presentation of the result, it is advisable to report technical aspects of the design.

2.6 SUMMARY

In this lesson we studied the concept of census and sampling. Also, we studied in detail various principles as well as the process used for selecting an appropriate sample. Sampling is the process whereby some elements (individuals) in the population are selected for a research study. The population consists of all individuals with a particular characteristic that is of interest to the researchers. If data are obtained from all members of the population, then we have a census; if data are obtained from some members of the population, then we have a sample.

In this lesson we have also discussed about the conditions, principles of sample surveys as well as the process of sampling.

2.7 GLOSSARY

- **Population:** In statistics, a population is a set of similar items or events which is of interest for some question or experiment.
- **Sample:** In statistics and quantitative research methodology, a data sample is a set of data collected and the world selected from a statistical population by a defined procedure.
- **Parameter:** Parameters in statistics is an important component of any statistical analysis. In simple words, a parameter is any numerical quantity that characterises a given population or some aspect of it. This means the parameter tells us something about the whole population.
- **Statistic:** A statistic is a characteristic of a sample. Generally, a statistic is used to estimate the value of a population parameter.

- **Sampling Frame:** Is a Frame that could be used as a basis for sampling (allows determining Probability of selection) and normally is any list, material or device that delimits, identifies, and allows access to the elements of the Survey population.
- **Sampling Unit:** A sampling unit is one of the units into which an aggregate is divided for the purpose of sampling, each unit being regarded as individual and indivisible when the selection is made.

2.8 SELF ASSESSMENT QUESTIONS

(i) Fill in the Blanks:

1. A sample is a study of.....of the population.
2. A population is the .....of limits under study.
3. Random sample is also referred to as .....sampling.
4. Any numerical value calculated from sample data is called.
5. Standard deviation of sampling distribution of any statistic is called.

(ii) Indicate (✓) whether the following statements are True or False.

1. A sample is less expensive than a census. **T/F**
2. The results obtained in a census study are always more reliable than those obtained in a sample study. **T/F**
3. Judgement sampling is a type of probability sampling method. **T/F**

2.9 LESSON END EXERCISE

1. When would you prefer complete enumeration over sample survey?

2. List the principle steps of Sampling.

3. What do you mean by principle of optimisation?

4. Define a sample and describe the conditions for sample survey briefly**.**

5. Describe the census and situations where it is essential.

6. Describe the advantages of sampling over census.

2.10 SUGGESTED READING

- Levin, R.I. Robin, D.S. *Statistics for Management*, Prentice-Hall of India. New Delhi.
- Aczel, A. D. Sounderpandian, J. *Complete Business Statistics*, Mc Graw Hill Publishing. New Delhi. downloaded

- Anderson, S., W. *Statistics for Business and Economics*, Cengage Learning. New Delhi.
- Kazmeir L. J. *Business Statistics*, Tata Mc Graw Hill. New Delhi.
- Vohra, N. D. *Business Statistics*. Tata Mc Graw Hill. New Delhi.

PROBABILITY AND NON-PROBABILITY SAMPLING TECHNIQUES

STRUCTURE

3.1 INTRODUCTION

In practical problems the statistician is often with the necessity of discussing population where he cannot examine every member. For example, an inquirer into the heights of the population of a city cannot afford the time or expense required to measure the height of every individual; nor can a producer who wants to know what proportion of his product is defective examine every single product. In such cases an investigator can examine a limited number of individuals/items/units of the population and hope that they will tell him, with reasonable trust worthiness, as much as he wants to know about the population from which they come. We are thus led to the questions: what can be said about the population when we can examine only a limited number of its members? This specific question is the origin of the theory of sampling. Also, there are two main methods used in survey research for selecting sample, viz., probability sampling and non-probability sampling. The big difference is that in probability sampling all persons has a chance of being selected, and results are more likely to be accurately reflecting the entire population. We had already discussed about the concept and essentials of sampling in lesson 2. Therefore, the focus of this lesson is on the techniques of sampling i.e., probability and non-probability sampling techniques.

In sampling, there are few terminologies that a researcher should be familiar with. For

example, let us say you are working in a research project on computing implementation for elderly and disabled citizens for a smart home system. You are supposed to find out the average age of senior and disabled citizens involved in your study.

(a) The community, families living in the town with smart homes form the population or study population and are usually denoted by the letter N.

(b) The sample group of elderly people or senior citizens and disable people in the vicinity of the smart home community is called sample.

(c) The number of elderly people or senior citizens and disabled people you obtain information to find their average age is called the sample size and is usually denoted by letter n.

(d) The way you select senior citizens and disabled people is called the sampling design or strategy.

(e) Each citizen or disabled people that become the basis for selecting your sample is called the sampling unit or sampling element.

(f) A list identifying each respondent in the study population is called sampling frame. In case when all elements in a sampling population cannot be individually identified, you cannot have a sampling frame for the study population.

(g) Finally, the obtained findings based on the information of the respondents are called sample statistics.

3.2 OBJECTIVES

After studying this lesson, you will be able to:

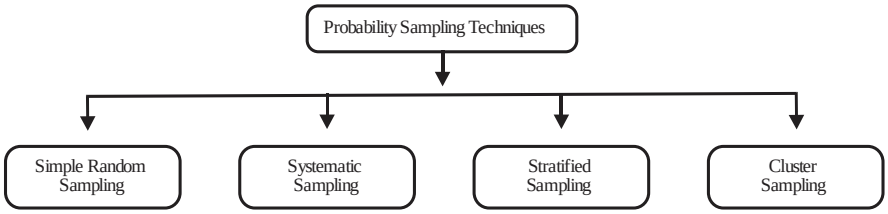
- Explain various terms associated with sampling.
- Describe various methods of probability sampling.
- Find out when to use the different methods of probability sampling.
- Understand why sampling is so common in business decisions.
- Understand the difference between probability and non-probability sampling.

- Differentiate among different methods of sampling, which include convenience, judgment, quota, and snowball.

3.3 PROBABILITY SAMPLING METHODS

The various sampling techniques can be grouped into two categories, i.e., probability sampling (also known as random sampling) and non-probability sampling (non-random sampling). Non-probability sampling techniques were already discussed in lesson 3. Therefore, here our focus is on probability sampling techniques. Probability sampling methods are those in which every item in the universe has a known chance, or probability, of being selected in the sample. This implies that the selection of sample items is independent of the person making the sample, that is, the sampling operation is controlled so objectively that the items will be chosen strictly at random. It may be noted that the term random sample is not used to describe the data in the sample but the process employed to select the sample. Randomness is thus a property of the sampling procedure instead of an individual sample. As such randomness can enter processed sampling in a number of ways and hence random samples may be many kinds.

PROBABILITY SAMPLING TECHNIQUES



3.3.1 SIMPLE RANDOM SAMPLING

Researchers use two major sampling techniques: probability sampling and non-probability sampling. With probability sampling, a researcher can specify the probability of an element's (participant's) being included in the sample. With non-probability sampling, there is no way of estimating the probability of an element's being included in a sample. If the researcher's interest is in generalising the findings derived from the sample to the general population, then probability sampling is far more useful and precise. Unfortunately, it is also much more difficult and expensive than non-probability sampling. Probability sampling

is also referred to as random sampling or representative sampling. The word random describes the procedure used to select elements (participants, cars, test items) from a population.

When random sampling is used, each element in the population has an equal chance of being selected (simple random sampling). The sample is referred to as representative because the characteristics of a properly drawn sample represent the parent population in all ways.

Step 1. ***Defining the Population***

Before a sample is taken, we must first define the population to which we want to generalise our results. The population of interest may differ for each study. It could be the population of professional football players in the United States or the registered voters in Bowling Green, Ohio. It could also be all college students at a given University, or all sophomores at that institution. It could be female students, or introductory psychology students, or 10-year-old children in a particular school, or members of the local senior citizens center. The point should be clear; the sample should be drawn from the population to which you want to generalise the population in which you are interested. It is unfortunate that many researchers fail to make explicit their population of interest. Many investigators use only college students in their samples, yet their interest is in the adult population of the United States. To a large extent, the generalisability of sample data depends on what is being studied and the inferences that are being made. For example, imagine a study that sampled college juniors at a specific University. Findings showed that a specific chemical compound produced pupil dilation. We would not have serious misgivings about generalising this finding to all college students, even tentatively to all adults, or perhaps even to some non-human organisms. The reason for this is that physiological systems are quite similar from one person to another, and often from one species to another. However, if we find that controlled exposure to unfamiliar political philosophies led to radicalisation of the experimental participants, we would be far more reluctant to extend this conclusion to the general population.

Step 2. ***Constructing a List***

Before a sample can be chosen randomly, it is necessary to have a complete list of the

population. In some cases, the logistics and expense of constructing a list of the entire population is simply too great, and an alternative procedure is forced upon the investigator. We could avoid this problem by restricting our population of interest-by defining it narrowly. However, doing so might increase the difficulty of finding or constructing a list from which to make our random selection.

For example, you would have no difficulty identifying female students at any given University and then constructing a list of their names from which to draw a random sample. It would be more difficult to identify female students coming from a three-child family, and even more difficult if you narrowed your interest to firstborn females in a three-child family. Moreover, defining a population narrowly also means generalising results narrowly.

Caution must be exercised in compiling a list or in using one already constructed. The population list from which you intend to sample, must be both recent and exhaustive. If not, problems can occur. By an exhaustive list, we mean that all members of the population must appear on the list. Voter registration lists, telephone directories, homeowner lists, and school directories are sometimes used, but these lists may have limitations. They must be up to date and complete if the samples chosen from them are to be truly representative of the population. In addition, such lists may provide very biased samples for some research questions we ask. For example, a list of homeowners would not be representative of all individuals in a given geographical region because it would exclude transients and renters. On the other hand, a ready-made list is often of better quality and less expensive to obtain than a newly constructed list would be.

Some lists are available from a variety of different sources. Professional organisations, such as the American Psychological Association, the American Medical Association, and the American Dental Association, have directory listings with mailing addresses of members. Keep in mind that these lists do not represent all psychologists, physicians, or dentists. Many individuals do not become members of their professional organisations. Therefore, a generalisation would have to be limited to those professionals listed in the directory. In universities and colleges, complete lists of students can be obtained from the registrar.

Let’s look at a classic example of poor sampling in the hours prior to a presidential election. Information derived from sampling procedures is often used to predict election outcomes.

Individuals in the sample are asked their candidate preferences before the election, and projections are then made regarding the likely winner. More often than not, the polls predict the outcome with considerable accuracy. However, there are notable exceptions, such as the 1936 *Literary Digest* magazine poll that predicted "Landon by a Landslide" over Roosevelt, and predictions in the U.S. presidential election of 1948 that Dewey would defeat Truman.

We have discussed the systematic error of the *Literary Digest* poll. Different reasons resulted in the wrong prediction in the 1948 presidential election between Dewey and Truman. Polls taken in 1948 revealed a large undecided vote. Based partly on this and early returns on the night of the election, the editors of the *Chicago Tribune* printed and distributed their newspaper before the election results were all in. The headline in bold letters indicated that Dewey defeated Truman. Unfortunately for them, they were wrong. Truman won, and the newspaper became a collector's item.

One analysis of why the polls predicted the wrong outcome emphasised the consolidation of opinion for many undecided voters. It was this undecided group that proved the prediction wrong. Pollsters did not anticipate that those who were undecided would vote in large numbers for Truman. Other factors generally operate to reduce the accuracy of political polls. One is that individuals do not always vote the way they say they are going to. Others may intend to do so but change their mind in the voting booth.

Also, the proportion of potential voters who actually cast ballot differs depending upon the political party and often upon the candidates who are running. Some political analysts believe (along with politicians) that even the position of the candidate's name on the ballot can affect the outcome (the debate regarding butterfly ballots in Florida during the 2000 presidential election comes to mind).

Step 3. *Drawing the Sample*

After a list of population members has been constructed, various random sampling options are available. Some common ones include tossing dice, flipping coins, spinning wheels, drawing names out of a rotating drum, using a table of random numbers, and using computer programs. Except for the last two methods, most of the techniques are slow and cumbersome. Tables of random numbers are easy to use, accessible, and truly random.

Let's look at the procedures for using the table. The first step is to assign a number to each individual on the list. If there were 1,000 people in the population, you would number them 0 to 999 and then enter the table of random numbers. Let us assume your sample size will be 100. Starting anywhere in the table, move in any direction you choose, preferably up and down. Since there are 1,000 people on your list (0 through 999) you must give each an equal chance of being selected. To do this, you use three columns of digits from the tables. If the first three-digit number in the table is 218, participant number 218 on the population list is chosen for the sample. If the next three-digit number is 007, the participant assigned number 007 (or 7) is selected. Continue until you have selected all 100 participants for the sample. If the same number comes up more than once, it is simply discarded. In the preceding fictional population list, the first digit (9) in the total population of 1,000 (0–999) was large. Sometimes the first digit in the population total is small, as with a list of 200 or 2,000. When this happens, many of the random numbers encountered in the table will not be usable and therefore must be passed up. This is very common and does not constitute a sampling problem. Also, tables of random numbers come in different column groupings. Some come in columns of two digits, some three, some four, and so on. These differences have no bearing on randomness. Finally, it is imperative that you not violate the random selection procedure. Once the list has been compiled and the process of selection has begun, the table of random numbers dictates who will be selected. The experimenter should not alter this procedure. A more recent method of random sampling uses the special functions of computer software. Many population lists are now available as software databases (such as Excel, Quattro Pro, Lotus123) or can be imported to such a database. Many of these database programs have a function for generating a series of random numbers and a function for selecting a random sample from a range of entries in the database.

Step 4. ***Contacting Members of a Sample***

Researchers using random sampling procedures must be prepared to encounter difficulties at several points. As we noted, the starting point is an accurate statement that identifies the population to which we want to generalise. Then we must obtain a listing of the population, accurate and up-to-date, from which to draw our sample. Further, we must decide on the random selection procedure that we wish to use. Finally, we must contact each of those selected for our sample and obtain the information needed. Failing

to contact all individuals in the sample can be a problem, and the representativeness of the sample can be lost at this point. To illustrate what we mean, assume that we are interested in the attitudes of college students at your University. We have a comprehensive list of students and randomly select 100 of them for our sample. We send a survey to the 100 students, but only 80 students return it. We are faced with a dilemma. Is the sample of 80 students who participated representative? Because 20% of our sample was not located, does our sample under represent some views? Does it over represent other views? In short, can we generalise from our sample to the college population? Ideally, all individuals in a sample should be contacted. As the number contacted decreases, the risk of bias and not being representative increases. Thus, in our illustration, to generalise to the college population would be to invite risk. Yet we do have data on 80% of our sample. Is it of any value? Other than simply dropping the project or starting a new one, we can consider an alternative that other researchers have used. In preparing our report, we would first clearly acknowledge that not all members of the sample participated and therefore the sample may not be random-that is, representative of the population. Then we would make available to the reader or listener of our report the number of participants initially selected and the final number contacted, the number of participants cooperating, and the number not co-operating. We would attempt to assess the reasons participants could not be contacted and whether differences existed between those for whom there were data and those for whom there were no data. If no obvious differences were found, we could feel a little better about the sample's being representative.

Differences on any characteristic between those who participated and those who did not should not automatically suggest that the information they might give would also differ. Individuals can share many common values and beliefs, even though they may differ on characteristics such as gender or education. In situations requiring judgments, such as the one described, the important thing is for the researcher to describe the strengths and weaknesses of the study (especially telling the reader that only 80 of the 100 surveys were returned), along with what might be expected as a result of them.

The problem just described may be especially troublesome when surveys or questionnaires deal with matters of a personal nature. Individuals are usually reluctant to provide information on personal matters, such as sexual practices, religious beliefs, or political philosophy. The

more personal the question, the fewer the number of people who will respond. With surveys or questionnaires of this nature, a large number of individuals may refuse to cooperate or refuse to provide certain information. Some of these surveys have had return rates as low as 20%. Even if we knew the population from which the sample was drawn and if the sample was randomly selected, a return rate as low as 20% is virtually useless in terms of generalising findings from the sample to the population. Those individuals responding to a survey (20% of the sample) could be radically different from the majority of individuals not responding (80% of the sample).

Let's apply these four steps of random sampling to our TV violence study. Our first step is to define the population. We might begin by considering the population as all children in the United States that are 5–15 years old. Our next step will be to obtain an exhaustive list of these children. Using U. S. Census data would be one approach, although the task would be challenging and the Census does miss many people. The third step is to select a random sample. As noted earlier in the chapter, the simplest technique would be to use a database of the population and instruct the database software to randomly select children from the population. The number to be selected is determined by the researcher and is typically based on the largest number that can be sampled given the logistical resources of the researcher. Of course, the larger the sample, the more accurately it will represent the population. In fact, formulas can be used to determine sample size based on the size of the population, the amount of variability in the population, the estimated size of the effect, and the amount of sampling error that the researcher decides is acceptable (refer to statistics books for specifics). After the sample is selected from the population, the final step is to contact the parents of these children to obtain consent to participate. You will need to make phone calls and send letters. Again, this will be a challenge; you expect that you will be unable to contact a certain percentage, and that a certain percentage will decline to participate. All this effort, and we have not even begun to talk about collecting data from these children.

From this example, it is clear that random sampling can require an incredible amount of financial resources. As noted earlier in the chapter, we have two options. We can define the population more narrowly (perhaps the 5- to 15-year-olds in a particular school district)

and conduct random sampling from this population, or we can turn to a sampling technique other than probability sampling.

3.3.1.1 Merits

1. Since it is a probability sampling, it eliminates the bias due to the personal judgement or discretion of the investigator. Accordingly, the sample selected is more representative of the population than in case of judgement sampling.

2. Because of its random character, it is possible to ascertain the efficient of the estimates by considering the standard errors of their sampling distributions.

3. The theory of random sampling is highly developed so that it enables us to obtain the most reliable and maximum information at the least cost, and results in savings in time, money and labour.

3.3.1.2 Demerits

1. Simple random sampling requires an up to date frame, i.e., a complete and up-to-date list of the population units to be sampled. In practice, since this is not readily available in many inquiries, it restricts the use of this sampling design.

2. In field surveys if the area of coverage is fairly large, then the units selected in the random sample are expected to be scattered widely geographically and thus it may be quite time consuming and costly to collect the requisite information or data.

3. If the sample is not sufficiently large, then it may not be representative of the population and thus may not reflect the true characteristics of the population.

4. The numbering of the population units and the preparation of the slips is quite time consuming and uneconomical particularly if the population is large. Accordingly, this method can't be used effectively to collect most of the data in social sciences.

3.3.2 SYSTEMATIC SAMPLING

Systematic sampling is another random sampling technique. Unlike stratified random sampling, systematic sampling is not done in an attempt to reduce sampling error. Rather, systematic sampling is used because of its convenience and relative ease of administration.

With systematic sampling, every kth item is selected to produce a sample of size n from a population of size N. The value of k, sometimes called the sampling cycle, can be determined by the following formula. If k is not an integer value, the whole-number value should be used

$$k = N/n$$

where,

n = sample size
N = population size
k = size of interval for selection

As an example of systematic sampling, a management information systems researcher wanted to sample the manufacturers in Texas. He had enough financial support to sample 1,000 companies (n). The Directory of Texas Manufacturers listed approximately 17,000 total manufacturers in Texas (N) in alphabetical order. The value of k was 17 (17,000/ 1,000) and the researcher selected every 17th company in the directory for his sample. Did the researcher begin with the first company listed or the 17th or one somewhere between? In selecting every kth value, a simple random number table should be used to select a value between 1 and k inclusive as a starting point. The second element for the sample is the starting point plus k. In the example, k = 17, so the researcher would have gone to a table of random numbers to determine a starting point between 1 and 17. Suppose he selected the number 5.He would have started with the 5th company, then selected the 22nd (5 + 17), and then the 39th, and so on.

Besides convenience, systematic sampling has other advantages. Because systematic sampling is evenly distributed across the frame, a knowledgeable person can easily determine whether a sampling plan has been followed in a study. However, a problem with systematic sampling can occur if the data are subject to any periodicity, and the sampling interval is in syncopation with it. In such a case, the sampling would be non-random. For example, if a list of 150 college students is actually a merged list of five classes with 30 students in each class and if each of the lists of the five classes has been ordered with the names of top students first and bottom students last, then systematic sampling of every 30th student could cause selection of all top students, all bottom students, or all mediocre students; that

is, the original list is subject to a cyclical or periodic organization. Systematic sampling methodology is based on the assumption that the source of population elements is random.

3.3.2.1 Merits

1. Systematic sampling is very easy to operate and checking can also be done quickly. Accordingly, it results in considerable saving in time and labour relative to simple random sampling or stratified sampling.

2. Systematic sampling may be more efficient than simple random sampling provided the frame is complete and up-to-date and the units are arranged serially in a random order like the names in a telephone directory where the units are arranged in alphabetical order. However, even in alphabetical arrangement, certain amount of non-random character may persist.

3.3.2.2 Demerits

1. Systematic sampling works well only if the complete and up-to-date frame is available and if the units are randomly arranged. However, these requirements are not generally fulfilled.

2. Systematic sampling gives biased results if there are periodic features in the frame and the sampling interval (k) is equal to or a multiple of the period.

3.3.3 STRATIFIED SAMPLING

This procedure known as stratified random sampling is also a form of probability sampling. To stratify means to classify or to separate people into groups according to some characteristics, such as position, rank, income, education, or ethnic background. These separate groupings are referred to as subsets or subgroups. For a stratified random sample, the population is divided into groups or strata. A random sample is selected from each stratum based upon the percentage that each subgroup represents in the population. Stratified random samples are generally more accurate in representing the population than are simple random samples. They also require more effort, and there is a practical limit to the number of strata used. Because participants are to be chosen randomly from each stratum, a complete list of the population within each stratum must be constructed. Stratified sampling

is generally used in two different ways. In one, primary interest is in the representativeness of the sample for purposes of commenting on the population. In the other, the focus of interest is comparison between and among the strata.

Let's look first at an example in which the population is of primary interest. Suppose we are interested in the attitudes and opinions of university faculty in a certain state toward faculty unionisation. Historically, this issue has been a very controversial one evoking strong emotions on both sides. Assume that there are eight universities in the state, each with a different faculty size (faculty size = 500 + 800 + 900 + 1,000 + 1,400 + 1,600 + 1,800 + 2,000 = 10,000). We could simply take a simple random sample of all 10,000 faculty and send those in the sample a carefully constructed attitude survey concerning unionisation. After considering this strategy, we decide against it. Our thought is that universities of different size may have marked differences in their attitudes, and we want to be sure that each university will be represented in the sample in proportion to its representation in the total university population. We know that, on occasion, a simple random sample will not do this. For example, if unionisation is a particularly "hot" issue on one campus, we may obtain a disproportionate number of replies from that faculty. Therefore, we would construct a list of the entire faculty for each university and then sample randomly within each university in proportion to its representation in the total faculty of 10,000. For example, the university with 500 faculty members would represent 5% of our sample; assuming a total sample size of 1,000, we would randomly select 50 faculty from this university. The university with 2,000 faculty would represent 20% of our sample; thus, 200 of its faculty would be randomly selected. We would continue until our sample was complete. It would be possible but more costly and time consuming to include other strata of interest-for example, full, associate, and assistant professors. In each case, the faculty in each stratum would be randomly selected.

As previously noted, stratified samples are sometimes used to optimize group comparisons. In this case, we are not concerned about representing the total population. Instead, our focus is on comparisons involving two or more strata. If the groups involved in our comparisons are equally represented in the population, a single random sample could be used. When this is not the case, a different procedure is necessary. For example, if we were interested in making comparisons between whites and blacks, a simple random

sample of 100 people might include about 85 to 90 whites and only 10 to 15 blacks. This is hardly a satisfactory sample for making comparisons. With a stratified random sample, we could randomly choose 50 whites and 50 blacks and thus optimize our comparison. Whenever strata rather than the population are our primary interest, we can sample in different proportions from each stratum.

Although random sampling is optimal from a methodological point of view, it is not always possible from a practical point of view.

Stratified random sampling can be either proportionate or disproportionate. ***Proportionate stratified random sampling*** occurs when the percentage of the sample taken from each stratum is proportionate to the percentage that each stratum is within the whole population. For example, suppose voters are being surveyed in Boston and the sample is being stratified by religion as Catholic, Protestant, Jewish, and others. If Boston's population is 90% Catholic and if a sample of 1,000 voters is taken, the sample would require inclusion of 900 Catholics to achieve proportionate stratification. Any other number of Catholics would be disproportionate stratification. The sample proportion of other religions would also have to follow population percentages. Or consider the city of El Paso, Texas, where the population is approximately 77% Hispanic. If a researcher is conducting a citywide poll in El Paso and if stratification is by ethnicity, a proportionate stratified random sample should contain 77% Hispanics. Hence, an ethnically proportionate stratified sample of 160 residents from El Paso's 600,000 residents should contain approximately 123 Hispanics. Whenever the proportions of the strata in the sample are different from the proportions of the strata in the population, ***disproportionate stratified random sampling occurs.***

3.3.3.1 Merits

1. More representatives: Since the population is first divided into various strata and then a sample is drawn from each stratum there is a little possibility of any essential group of the population being completely excluded. A more representative sample is thus secured, C.J. Grohmann has rightly pointed out that this type of sampling balances the uncertainty of random sampling against the bias of deliberate selection.
2. Greater accuracy: Stratified sampling ensures greater accuracy. The accuracy is

maximum if each stratum is so formed that it consists of uniform or homogeneous items.

3. Greater geographical concentration: As compared with random sample, stratified samples can be more concentrated geographically, i.e., the units form the different strata may be selected in such a way that all of them are localised in one geographical area. This would greatly reduce the time and expenses of interviewing.

3.3.3.2 Demerits

1. Disproportional stratified sampling requires the assignment of weights to different strata and if the weights assigned are faulty, the resulting sample will not be representative and might give biased results.

2. The items from each stratum should be selected at random. But this may be difficult to achieve in the absence of skilled sampling supervisors and a random selection within each stratum may not be ensured.

3. Because of the likelihood that a stratified sample will be more widely distributed geographically than a simple random sample cost per observation may be quite high.

3.3.4 CLUSTER SAMPLING

Cluster (or area) sampling is a fourth type of random sampling. Cluster (or area) sampling involves dividing the population into non-overlapping areas, or clusters. However, in contrast to stratified random sampling where strata are homogeneous within, cluster sampling identifies clusters that tend to be internally heterogeneous. In theory, each cluster contains a wide variety of elements, and the cluster is a miniature, or microcosm, of the population. Examples of clusters are towns, companies, homes, colleges, areas of a city, and geographic regions. Often clusters are naturally occurring groups of the population and are already identified, such as states or Standard Metropolitan Statistical Areas. Although area sampling usually refers to clusters that are areas of the population, such as geographic regions and cities, the terms cluster sampling and area sampling are used interchangeably in this text. After randomly selecting clusters from the population, the business researcher either selects all elements of the chosen clusters or randomly selects individual elements into the sample from the clusters. One example of business research that makes use of clustering is test

marketing of new products. Often in test marketing, the United States is divided into clusters of test market cities, and individual consumers within the test market cities are surveyed.

Sometimes the clusters are too large, and a second set of clusters is taken from each original cluster. This technique is called two-stage or multistage sampling. For example, a researcher could divide the United States into clusters of cities. He could then divide the cities into clusters of blocks and randomly select individual houses from the block clusters. The first stage is selecting the test cities and the second stage is selecting the blocks. Clusters are usually convenient to obtain, and the cost of sampling from the entire population is reduced because the scope of the study is reduced to the clusters. The cost per element is usually lower in cluster or area sampling than in stratified sampling because of lower element listing or locating costs. The time and cost of contacting elements of the population can be reduced, especially if travel is involved, because clustering reduces the distance to the sampled elements. In addition, administration of the sample survey can be simplified. Sometimes cluster or area sampling is the only feasible approach because the sampling frames of the individual elements of the population are unavailable and therefore other random sampling techniques cannot be used. If the elements of a cluster are similar, cluster sampling may be statistically less efficient than simple random sampling. Moreover, the costs and problems of statistical analysis are greater with cluster or area sampling than with simple random sampling.

3.3.4.1 Merits

Multistage sampling is more flexible as compared to other methods of sampling. It is simple to carry out and results in administrative convenience by permitting the field work to be concentrated and yet covering large area.

An important practical advantage of multistage sampling is that we need the second stage frame only for those units which are selected in the first sample and this leads to great saving of operating cost.

Consequently this technique is of great utility, particularly in surveys of under developed area where no up-to-date frame is available for subdivision of the material into reasonably small sampling units.

3.3.4.2 Demerits

Errors are likely to be larger in this method. The variability of the estimates under this method may be greater than that of estimates based on simple random sampling. This variability depends on the composition of the primary units. In general, a multistage sampling is usually less efficient than a suitable single stage sampling of the same size.

3.4 NON PROBABILITY SAMPLING TECHNIQUES

Sampling techniques used to select elements from the population by any mechanism that does not involve a random selection process are called non random sampling techniques. Because chance is not used to select items from the samples, these techniques are non-probability techniques and are not desirable for use in gathering data to be analyzed by the methods of inferential statistics. Sampling error cannot be determined objectively for these sampling techniques. Four non-random sampling techniques are discussed here: convenience sampling, judgment sampling, quota sampling, and snowball sampling.

3.4.1 Convenience Sampling

In convenience sampling, elements for the sample are selected for the convenience of the researcher. The researcher typically chooses elements that are readily available, nearby, or willing to participate. The sample tends to be less variable than the population because in many environments the extreme elements of the population are not readily available. The researcher will select more elements from the middle of the population. For example, a convenience sample of homes for door-to-door interviews might include houses where people are at home, houses with no dogs, houses near the street, first-floor apartments, and houses with friendly people. In contrast, a random sample would require the researcher to gather data only from houses and apartments that have been selected randomly, no matter how inconvenient or unfriendly the location. If a research firm is located in a mall, a convenience sample might be selected by interviewing only shoppers who pass the shop and look friendly.

3.4.2 Judgement Sampling

Judgment sampling occurs when elements selected for the sample are chosen by the judgment of the researcher. Researchers often believe they can obtain a representative

sample by using sound judgment, which will result in saving time and money. Sometimes ethical, professional researchers might believe they can select a more representative sample than the random process will provide. They might be right! However, some studies show that random sampling methods outperform judgment sampling in estimating the population mean even when the researcher who is administering the judgment sampling is trying to put together a representative sample. When sampling is done by judgment, calculating the probability that an element is going to be selected into the sample is not possible. The sampling error cannot be determined objectively because probabilities are based on non random selection.

Other problems are associated with judgment sampling. The researcher tends to make errors of judgment in one direction. These systematic errors lead to what are called biases. The researcher also is unlikely to include extreme elements. Judgment sampling provides no objective method for determining whether one person's judgment is better than another's.

3.4.3 Quota Sampling

A third non random sampling technique is quota sampling, which appears to be similar to stratified random sampling. Certain population subclasses, such as age group, gender, or geographic region, are used as strata. However, instead of randomly sampling from each stratum, the researcher uses a non random sampling method to gather data from one stratum until the desired quota of samples is filled. Quotas are described by quota controls, which set the sizes of the samples to be obtained from the subgroups. Generally, a quota is based on the proportions of the subclasses in the population. In this case, the quota concept is similar to that of proportional stratified sampling.

Quotas often are filled by using available, recent, or applicable elements. For example, instead of randomly interviewing people to obtain a quota of Italian Americans, the researcher would go to the Italian area of the city and interview there until enough responses are obtained to fill the quota. In quota sampling, an interviewer would begin by asking a few filter questions; if the respondent represents a subclass whose quota has been filled, the interviewer would terminate the interview.

Quota sampling can be useful if no frame is available for the population. For example, suppose a researcher wants to stratify the population into owners of different types of cars

but fails to find any lists of Toyota van owners. Through quota sampling, the researcher would proceed by interviewing all car owners and casting out non-Toyota van owners until the quota of Toyota van owners is filled.

Quota sampling is less expensive than most random sampling techniques because it essentially is a technique of convenience. However, cost may not be meaningful because the quality of non random and random sampling techniques cannot be compared. Another advantage of quota sampling is the speed of data gathering. The researcher does not have to call back or send out a second questionnaire if he does not receive a response; he just moves on to the next element. Also, preparatory work for quota sampling is minimal.

The main problem with quota sampling is that, when all is said and done, it still is only a non random sampling technique. Some researchers believe that if the quota is filled by randomly selecting elements and discarding those not from a stratum, quota sampling is essentially a version of stratified random sampling. However, most quota sampling is carried out by the researcher going where the quota can be filled quickly. The object is to gain the benefits of stratification without the high field costs of stratification. Ultimately, it remains a non probability sampling method.

3.4.4 Snowball Sampling

Another non random sampling technique is snowball sampling, in which survey subjects are selected based on referral from other survey respondents. The researcher identifies a person who fits the profile of subjects wanted for the study. The researcher then asks this person for the names and locations of others who would also fit the profile of subjects wanted for the study. Through these referrals, survey subjects can be identified cheaply and efficiently, which is particularly useful when survey subjects are difficult to locate. It is the main advantage of snowball sampling; its main disadvantage is that it is non random.

3.5 SUMMARY

In this lesson we studied in detail various probability and non-probability sampling techniques used for collecting information from population. With probability sampling, a researcher can specify the probability of an element's (participant's) being included in the sample. With non probability sampling, there is no way of estimating the probability of an element's

being included in a sample. Convenience sampling is quick and inexpensive because it involves selecting individuals who are readily available at the time of the study (such as introductory psychology students). The disadvantage is that convenience samples are generally less representative than random samples; therefore, results should be interpreted with caution.

When we conduct research, we are generally interested in drawing some conclusion about a population of individuals that have some common characteristic. However, populations are typically too large to allow observations on all individuals, and we resort to selecting a sample. In order to make inferences about the population, the sample must be representative. Thus, the manner in which the sample is drawn is critical. Probability sampling uses random sampling in which each element in the population (or a subgroup of the population with stratified random sampling) has an equal chance of being selected for the sample. When probability sampling is not possible, non-probability sampling must be used. Larger samples are more likely to accurately represent characteristics of the population, and smaller samples are less likely to accurately represent characteristics of the population. Therefore, researchers strive for samples that are large enough to reduce sampling error to an acceptable level. Even when samples are large enough, it is important to evaluate the specific method by which the sample was drawn. We are increasingly exposed to information obtained from self-selected samples that represent only a very narrow subgroup of individuals. Much of such information is meaningless because the subgroup is difficult to identify.

Therefore, in this lesson we have studied the concept and uses of various probability as well as non-probability sampling techniques with their merits and demerits.

3.6 GLOSSARY

- **Probability Sampling:** A probability sampling method is any method of sampling that utilizes some form of random selection.
- **Non-probability Sampling:** Non-probability sampling is a sampling technique where the samples are gathered in a process that does not give all the individuals in the population equal chances of being selected.
- **Stratified Sampling:** Stratified sampling is a probability sampling technique wherein

the researcher divides the entire population into different subgroups or strata, then randomly selects the final subjects proportionally from the different strata.

- **Sample Size:** The number (n) of observations taken from a population through which statistical inferences for the whole population are made.
- **Cluster Sampling:** With cluster sampling, the researcher divides the population into separate groups, called clusters. Then, a simple random sample of clustersis selected from the population. The researcher conducts his analysis on data from the sampled clusters.
- **Simple Random Sampling:** Simple random sampling is a sampling technique where every item in the population has an even chance and likelihood of being selected in the sample.
- **Systematic Sampling:** Systematic sampling is a type of probability *sa*mpling method in which sample members from a larger population are selected according to a random starting point and a fixed, periodic interval. This interval, called the sampling interval, is calculated by dividing the population size by the desired sample size.

3.7 SELF ASSESSMENT QUESTIONS

(i) Indicate (✓) whether the following statements are true or false.

1. With probability samples the chance, or probability, of each case being selected from the population is unknown. **T/F**
2. The sampling frame for any probability sample is a complete list of all the cases in the population from which your sample will be drawn. **T/F**
3. Choice of sampling technique or techniques is dependent on your research question(s) and objectives and the feasibility of gaining access to the data. **T/F**
4. Generalisations about populations from data collected using any probability sample is based on intuition**.** **T/F**

3.8 LESSON END EXERCISE

1. What type of sampling is best for qualitative research?

2. Is convenience sampling qualitative or quantitative?

3. Distinguish between strata and cluster.

4. Is there any relationship between probability sample and simple random sampling?

5. Describe the importance of sampling. Critically examine the merits of probability sampling methods.

6. Specify and explain the factors that make sampling preferable to a complete census in statistical investigation.

7. How would you determine the sample size for stratified sampling? Explain with the help of a suitable example.

3.9 SUGGESTED READING

- Levin, R.I. Robin, D.S. *Statistics for Management*, Prentice-Hall of India. New Delhi.
- Aczel, A. D. Sounderpandian, J. *Complete Business Statistics*, Mc Graw Hill Publishing. New Delhi. downloaded
- Anderson, S., W. *Statistics for Business and Economics*, Cengage Learning. New Delhi.
- Freund, J. E., Williams, F.M. *Elementary Business Statistics*-The Modern Approach. Prentice Hall of India Private Ltd., New Delhi.

SAMPLING VS NON-SAMPLING ERRORS, EFFECTIVENESS OF SAMPLING AND DETERMINATION OF SAMPLE SIZE

STRUCTURE

4.1 INTRODUCTION

The aim of sampling is usually to estimate one or more population values (parameters) from a sample. The next chapter deals in depth with this issue of estimation, but we mention here that estimates such as sample means or proportions are random quantities. If we were to repeat the sampling process, the estimate would vary and this sample-to-sample variability can be described by a distribution (e.g. the distribution of the sample mean or sample proportion). The estimate is not guaranteed to be the same as the value that we are estimating, so we call the difference the error in the estimate. There are different kinds of error.

4.2 OBJECTIVES

After going through this lesson, you will be able to understand:

- Understand the concept of errors in statistics.
- Know about sampling and non-sampling errors.
- How sampling and non-sampling errors will be minimised.
- The process used for determining the sample size.

4.3 SAMPLING ERRORS

Sampling error is the error that arises in a data collection process as a result of taking a sample from a population rather than using the whole population.

Sampling error is one of two reasons for the difference between an estimate of a population

parameter and the true, but unknown, value of the population parameter. The other reason is non-sampling error.

Even if a sampling process has no non-sampling errors then estimates from different random samples (of the same size) will vary from sample to sample, and each estimate is likely to be different from the true value of the population parameter.

The sampling error for a given sample is unknown but when the sampling is random, for some estimates (for example, sample mean, sample proportion) theoretical methods may be used to measure the extent of the variation caused by sampling error.

4.3.1 Types of Sampling Errors

When you survey a sample, your interest usually goes beyond just the people in the sample. Rather, you are trying to get information to project onto a larger population. For this reason, it is important to understand common sampling errors so you can avoid them. Five common types of sampling errors:

- **Population Specification Error:** This error occurs when the researcher does not understand who they should survey. For example, imagine a survey about breakfast cereal consumption. Who to survey? It might be the entire family, the mother, or the children. The mother might make the purchase decision, but the children influence her choice.

- **Sample Frame Error:** A frame error occurs when the wrong sub-population is used to select a sample. A classic frame error occurred in the 1936 presidential election between Roosevelt and Landon. The sample frame was from car registrations and telephone directories. In 1936, many Americans did not own cars or telephones, and those who did were largely Republicans. The results wrongly predicted a Republican victory.

- **Selection Error:** This occurs when respondents self-select their participation in the study-only those that are interested respond. Selection error can be controlled by going extra lengths to get participation. A typical survey process includes initiating pre-survey contact requesting cooperation, actual surveying, and post-survey follow-up. If a response is not received, a second survey request follows, and perhaps interviews using alternate modes such as telephone or person-to-person.

- **Non-Response:** Non-response errors occur when respondents are different than

those who do not respond. This may occur because either the potential respondent was not contacted or they refused to respond. The extent of this non-response error can be checked through follow-up surveys using alternate modes.

- **Sampling Errors:** These errors occur because of variation in the number or representativeness of the sample that responds. Sampling errors can be controlled by (1) careful sample designs, (2) large samples, and (3) multiple contacts to assure representative response.

4.3.2 Sources of Sampling Errors

A sampling error is a problem in the way that members of a population are selected for research or data collection, which impacts the validity of results. Numerically, a sampling error expresses the difference between results for the sample and estimated results for the population.

Subjects are selected through several different methods, broadly categorised as probability-based or non-probability-based. Probability-based methods are considered to yield the most valid results because each member of a population has an equal chance of selection; as long as a sufficiently large sample is selected, the group should be representative of the population.

No sampling method is infallible. In simple random sampling, considered to be the most foolproof method, subjects for the sample are randomly selected from the entire population to create a subset. Even in this case, however, sample size is an issue. In general, a larger group of subjects will be more representative of the population. Imagine, for example, a study in which thirty subjects are selected from a population of a thousand-random selection could not ensure that the sample would represent the population. Other sampling errors include:

1. **Non-response:** Subjects may fail to respond, and those who respond may differ from those who don't in significant ways.
2. **Self-selection:** If subjects volunteer, that may indicate that they have a particular bias related to the study, which can skew results.
3. **Sample frame error:** A non-representative subgroup may be selected as a sample.

4. **Population specification error:** The researcher fails to identify the population of interest with enough precision.

5. A sufficiently large sample size, randomised selection and attention to study design can all help to improve the validity of data.

4.3.3 Measurement and Control of Sampling Errors

Of the two types of errors, sampling error is easier to identify. Although sampling error is unavoidable when collecting a random sample, we can take measures to estimate and reduce sampling error. The margin of error that you commonly see with survey results is in fact an estimate of sampling error. Because it is just an estimate, there is a small chance (typically five percent or less) that the margin of error is actually larger than stated in a report.

The techniques used for measuring and reducing sampling error are:

1. Increase the Sample Size

A larger sample size leads to a more precise result because the study gets closer to the actual population size.

2. Divide the Population into Groups

Instead of a random sample, test groups according to their size in the population. For example, if people of a certain demographic make up 35% of the population, make sure 35% of the study is made up of this variable.

3. Know your Population

The error of population specification is when a research team selects an inappropriate population to obtain data. Know who buys your product, uses it, works with you, and so forth. With basic socio-economic information, it is possible to reach a consistent sample of the population. In cases like marketing research, studies often relate to one specific population like Facebook users, Baby Boomers, or even homeowners.

4. *Reducing Sampling Error by Increasing Sample Size*

One way to reduce sampling error is to increase the size of your sample by selecting more subjects to observe. Sampling error and sample size have an inversely correlated relationship,

meaning that as sample size grows, sampling error decreases. However, it's important to note that increasing sample size usually results in an increase in cost. The more people that you want to survey in your study, the more expensive your study will be, as there are costs associated with identifying respondents or participants. There will also be an increased cost from a time usage perspective.

We've found that after bringing sample size to 1,000 participants, researchers generally start to get fewer bangs for their buck. This is due to the relationship between sample size and margin of error. Once you have a sample size of 1,000, even if you more than double your sample size to 2,500 you are only decreasing your margin of error by one percent.

5. Reducing Sampling Error with Solid Sample Design

Sampling error can also be reduced by ensuring that you have a solid sample design. For example, if your target population is made up of defined subpopulations, then you could reduce margin of error by sampling each subpopulation independently.The tactics above only reduce sampling error -they do not eliminate it. The only true way to eliminate sampling error entirely is to examine each and every individual member of your target population. This is often impractical, and in many cases impossible. But that's not to say that generating random samples is an ineffective means of investigating a population. It's still a convenient and effective way to examine a large-scale, complex population.

Example

A population specification error means that XYZ does not understand the specific types of consumers who should be included in the sample. If, for example, XYZ creates a population of people between the ages of 15 and 25 years old, many of those consumers do not make the purchasing decision about a video streaming service because they do not work full-time. On the other hand, if XYZ put together a sample of working adults who make purchase decisions, the consumers in this group may not watch 10 hours of video programming each week.

Selection error also causes distortions in the results of a sample, and a common example is a survey that only relies on a small portion of people who immediately respond. If XYZ makes an effort to follow up with consumers who don't initially respond, the results of the survey may change. Furthermore, if XYZ excludes consumers who don't respond right away, the sample results may not reflect the preferences of the entire population.

4.4 NON SAMPLING ERRORS

It is a general assumption in sampling theory that the true value of each unit in the population can be obtained and tabulated without any errors. In practice, this assumption may be violated due to several reasons and practical constraints. This results in errors in observations as well as in tabulation. Such errors which are due to factors other than sampling are called non-sampling errors. The non-sampling errors are unavoidable in census and surveys. The data collected by complete enumeration in census is free from sampling error but would not remain free from non-sampling errors. The data collected through sample surveys can have both-sampling errors as well as non-sampling errors. Non-sampling errors arise because of the factors other than the inductive process of inferring about the population from a sample. In general, the sampling errors decrease as the sample size increases whereas non-sampling error increases as the sample size increases. In some situations, the non-sampling errors may be large and deserve greater attention than the sampling error.

In any survey, it is assumed that the value of the characteristic to be measured has been defined precisely for every population unit. Such a value exists and is unique. This is called the true value of the characteristic for the population value. In practical applications, data collected on the selected units are called survey values and differ from the true values. Such difference between the true and observed values is termed as observational error or response error. Such an error arises mainly from the lack of precision in measurement techniques and variability in the performance of the investigators. Therefore, non-sampling error is the error that arises in a data collection process as a result of factors other than taking a sample.

Non-sampling errors have the potential to cause bias in polls, surveys or samples.

There are many different types of non-sampling errors and the names used to describe them are not consistent. Examples of non-sampling errors are generally more useful than using names to describe them.

4.4.1 Types of sampling Errors

Non-sampling errors may be broadly classified into three categories.

(a) Specification Errors: These errors occur at planning stage due to various reasons, e.g., inadequate and inconsistent specification of data with respect to the objectives of surveys/census, omission or duplication of units due to imprecise definitions, faulty method of enumeration/interview/ambiguous schedules etc.

(b) Ascertainment Errors: These errors occur at field stage due to various reasons e.g., lack of trained and experienced investigations, recall errors and other type of errors in data collection, lack of adequate inspection and lack of supervision of primary staff etc.

Ascertainment errors may be further sub-divided into

i. Coverage errors owing to over-enumeration or under-enumeration of the population or sample, resulting from duplication or omission of units and from non-response.

ii. Content errors relating to wrong entries due to errors on the part of investigators and respondents.

(c) Tabulation Errors: These errors occur at tabulation stage due to various reasons, e.g., inadequate scrutiny of data, errors in processing the data, errors in publishing the tabulated results, graphs etc.

Same division can be made in the case of tabulation error also. There is a possibility of missing data or repetition of data at tabulation stage which gives rise to coverage errors and also of errors in coding, calculations etc. which gives rise to content errors.

4.4.2 Sources of Non-sampling Errors

Non sampling errors can occur at every stage of planning and execution of survey or census. It occurs at planning stage, field work stage as well as at tabulation and computation stage. The main sources of non-sampling errors are lack of proper specification of the domain of study and scope of investigation, incomplete coverage of the population or sample, faulty definition, defective methods of data collection and tabulation errors.

More specifically, one or more of the following reasons may give rise to non-sampling errors or indicate its presence:

a) The data specification may be inadequate and inconsistent with the objectives of the survey or census.

b) Due to imprecise definition of the boundaries of area units, incomplete or wrong identification of units, faulty methods of enumeration etc, data may be duplicated or may be omitted.

c) The methods of interview and observation collection may be inaccurate or inappropriate.

d) The questionnaire, definitions and instructions may be ambiguous.

e) The investigators may be inexperienced or not trained properly.

f) The recall errors may pose difficulty in reporting the true data.

g) The scrutiny of data is not adequate.

h) The coding, tabulation etc. of the data may be erroneous.

i) There can be errors in presenting and printing the tabulated results, graphs etc.

j) In a sample survey, the non-sampling errors arise due to defective frames and faulty selection of sampling units.

k) These sources are not exhaustive but surely indicate the possible source of errors

The non-response error may occur due to refusal by respondents to give information or the sampling units may be inaccessible. This error arises because the set of units getting excluded may have characteristic so different from the set of units actually surveyed as to make the results biased. This error is termed as non-response error since it arises from the exclusion of some of the anticipated units in the sample or population. One way of dealing with the problem of non-response is to make all efforts to collect information from a sub-sample of the units not responding in the first attempt.

4.4.3 Measurement and Control of Sampling Errors

Some suitable methods and adequate procedures for control can be adopted before initiating the main census or sample survey. Some separate programmes for estimating the different types of non-sampling errors are also required. Some such procedures are as follows:

1. Consistency check: Certain items in the questionnaires can be added which may serve as a check on the quality of collected data. To locate the doubtful observations, the

data can be arranged in increasing order of some basic variable. Then they can be plotted against each sample unit. Such graph is expected to follow a certain pattern and any deviation from this pattern would help in spotting the discrepant values.

2. Sample Check: An independent duplicate census or sample survey can be conducted on a comparatively smaller group by trained and experienced staff. If the sample is properly designed and if the checking operation is efficiently carried out, it is possible to detect the presence of non-sampling errors and to get an idea of their magnitude. Such procedure is termed as method of sample check.

3. Post-census and post-survey checks: It is a type of sample check in which a sample (or subsample) is selected of the units covered in the census (or survey) and re-enumerate or re-survey it by using better trained and more experienced survey staff than those involved in the main investigation. This procedure is called as post-survey check or post-census. The effectiveness of such check surveys can be increased by re-enumerating or re-surveying immediately after the main census to avoid recall error taking steps to minimise the conditioning effect that the main survey may have on the work of the check-survey.

4. External record check: Take a sample of relevant units from a different source, if available, and to check whether all the units have been enumerated in the main investigation and whether there are discrepancies between the values when matched. The list from which the check-sample is drawn for this purpose, need not be a complete one.

5. Quality control techniques: The use of tools of statistical quality control like control chart and acceptance sampling techniques can be used in assessing the quality of data and in improving the reliability of final results in large scale surveys and census.

6. Study or recall error: Response errors arise due to various factors like the attitude of respondents towards the survey, method of interview, skill of the investigators and recall errors. Recall error depends on the length of the reporting period and on the interval between the reporting period and data of survey. One way of studying recall error is to collect and analyze data related to more than one reporting period in a sample (or sub-sample) of units covered in the census or survey.

7. Interpenetrating sub-samples: The use of interpenetrating sub-sample technique

helps in providing an appraisal of the quality of information as the interpenetrating sub-samples can be used to secure information on non-sampling errors such as differences arising from differential interviewer bias, different methods of eliciting information etc. After the sub-samples have been surveyed by different groups of investigators and processed by different team of workers at the tabulation stage, a comparison of the final estimates based on the sub-samples provides a broad check on the quality of the survey results.

4.5 EFFECTIVENESS OF SAMPLING

The effectiveness of sampling depends upon various benefits to. Some of the benefits determine the effectiveness of sampling are given below:

1. Sampling saves time to a great extent by reducing the volume of data. You do not go through each of the individual items.

2. Sampling Avoids monotony in works. You do not have to repeat the query again and again to all the individual data.

3. When you have limited time, survey without using sampling becomes impossible. It allows us to get near-accurate results in much lesser time

4. When you use proper methods, you are likely to achieve higher level of accuracy by using sampling than without using sampling in some cases due to reduction in monotony, data handling issues etc.

5. By using sampling, you can get detailed information on the data even by employing small amount of resources.

Also, the effective sample size is an estimate of the sample size required to achieve the same level of precision if that sample was a simple random sample. Mathematically, it is defined as n/D, where n is the sample size and D is the design effect. It is used as a way of summarising the amount of information in data. It has three main areas of application: survey analysis, time series analysis, and Bayesian statistic. The main application of effective sample size calculations is for qualitative assessments of the sample size. The sample size measures the number of individual samples measured or observations used in a survey or experiment. It is believed that a sample size of 30 is required for an analysis to be valid, then the effective sample size-rather than the actual sample size- is used in such an assessment.

Sometimes effective sample sizes are used as an input into statistical calculations in place of the actual sample size. In survey analysis, the way that a survey is designed affects the precision of survey estimates (i.e., the standard error of statistics). Stratification, clustering, and weighting all usually increase the standard errors of estimates in real-world surveys.

Most commonly, the effective sample size is used as a way of quantifying the effect of weighting a survey. For example, if a survey of 1,000 people has an effective sample size for a statistic of 500, it means that the amount of sampling error is equivalent to that which would have been obtained by a study of 500 people that did not need to be weighted.

A common misunderstanding in survey analysis is that a survey has an effective sample size. This is rarely the case. Most statistics that are calculated have their own effective sample size. For example, if you compute the effective sample size for the average of one variable, it will typically be different from the effective sample size computed for another variable. Where all the statistics have the same effective sample size, it means that an approximation of some kind has been used.

When autocorrelation exists in a time series, this also reduces the effective sample size. For example, if the first-order autocorrelation is 0.5, then the effective sample size of 100 observations is only 33 observations.

In Bayesian statistics, it is common to use the posterior draws from Markov chain Monte Carlo (MCMC) for statistical inference. The MCMC process causes the draws to be correlated. This means that the effective sample size is generally lower than the number of draws. For this reason, the effective sample size -rather than the actual sample size- is typically used when determining if an MCMC model has converged.

4.6 DETERMINATION OF SAMPLE SIZE

Sample size refers to the number of elements to be included in the study. Determining the sample size involves several qualitative and quantitative considerations. Important qualitative factors to be considered in deter mining the sample size include: the importance of the decision, the nature of the research, the number of variables, the nature of the analysis, sample sizes used in similar studies, incidence rates, completion rates, and resource constraints.

What should be the size of the sample or how large or small should be 'n'? If the sample size ('n') is too small, it may not serve to achieve the ob may incur huge cost and waste resources. As a general rule, one can say that the sample must be of an optimum size i.e., it should neither be excessively large nor too small. Technically, the sample size should be large enough to give a confidence interval of desired width and as such the size of the sample must be chosen by some logical process before sample is taken from the universe. Size of the sample should be determined by a researcher keeping in view the following points:

1. Nature of universe: Universe may be either homogenous or heterogeneous in nature. If a small sample can serve the purpose. But if the items are heterogeneous, a large sample would be required. Technically, this can be termed as the dispersion factor.

2. Number of classes proposed: If many class-groups (groups and sub-groups) are to because a small sample might not be able to give a reasonable number of items in each class-group.

3. Nature of study: If items are to be intensively and continuously studied, the sample should be small. For a general survey the size of the sample should be large, but a small sample is

4. Type of sampling: Sampling technique plays an important part in determining the size of the sample. A small random sample is apt to be much superior to a larger but badly selected sample.

5. Standard of accuracy and acceptable confidence level: If the standard of acuracy or the level of precision is to be kept high, we shall require relatively larger sample. For doubling the accuracy for a fixed significance level, the sample size has to be increased fourfold.

6. Availability of finance: In practice, size of the sample depends upon the amount of money available for the study purposes. This factor should be kept in view while determining the size of sample for large samples result in increasing the cost of sampling estimates.

7. Other considerations: Nature of units, size of the population, size of questionnaire, availability of trained investigators, the conditions under which the sample is being conducted, the time available for completion of the study are a few other considerations to which a

researcher must pay attention while selecting the size of the sample.

The statistically determined sample size is the net or final sample size: the sample remaining after eliminating potential respondents who do not qualify or who do not complete the interview. Depending on incidence and completion rates, the size of the initial sample may have to be much larger. The statistical approach to determining sample size that we consider is based on traditional statistical inference. In this approach the precision level is specified in advance. The confidence interval approach to sample size determination is based on the construction of confidence intervals around the sample means or proportions using the standard error formula.

As an example, suppose that a researcher has taken a simple random sample of 300 households to estimate the monthly amount invested in savings schemes and found that the mean household monthly investment for the s sample is €182. Past studies indicate that the population standard deviation can be assumed to be €55.

We want to find an interval within which a fixed proportion of the sample means would fall. Suppose that we want to determine an interval around the population mean that will include 95% of the sample means, based on samples of 300 households. The 95% could be divided into two equal parts, half below and half above the mean, as shown in Figure 4.1.

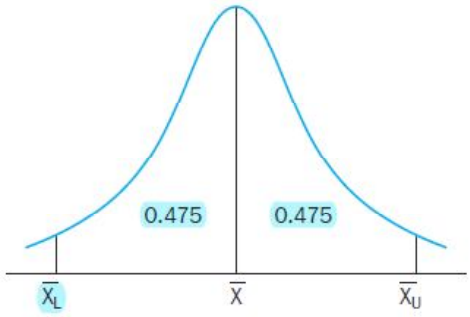


Figure 4.1: The 95% Confidence Interval

Calculation of the confidence interval involves determining a distance below (X_L) and above (X_U) the population mean (X), which contains a specified area of the normal curve.

The z values corresponding to (X_L) and (X_U) may be calculated as:

$$z_L = \frac{\bar{X}_L - \mu}{\sigma_{\bar{X}}}$$

$$z_U = \frac{\bar{X}_U - \mu}{\sigma_{\bar{X}}}$$

Where Z_L = -Z and Zu = + z. Therefore, the lower value of X is

$$X_L = \mu - z\sigma_X$$

and the upper value of $X = \mu + z\sigma_X$

Note that μ is estimated by X the confidence interval is given by:

$$X + z\sigma_X$$

We can now set a 95% confidence interval around the sample mean of €182. As a first step, we compute the standard error of the mean:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{55}{\sqrt{300}} = 3.18$$

According to statistical table the central 95% normal distribution lies within +1.96 z values. The 95%confidence interval is given by:

$$\begin{aligned} &\bar{X} \pm 1.96\sigma_{\bar{X}} \\ &= 182.00 \pm 1.96\ (3.18) \\ &= 182.00 \pm 6.23 \end{aligned}$$

Thus, the 95% confidence interval ranges from €175.77 to €188.23. The probability of finding the true population mean to be within €175.77 and €188.23 is 95%.

4.6.1 Sample Size Determination: Mean

The approach used here to construct a confidence interval can be adapted to determine the sample size that will result in a desired confidence interval.3 Suppose that the researcher wants to estimate the monthly household savings investment more pre ciselyso that the

estimate will be within ± 5.00 of the true population value. What € should be the size of the sample? The following steps will lead to an answer.

1. **Specify the level of precision.** This is the maximum permissible difference (D) between the sample mean and the population mean. In our example, D = ± €5.00.

2. **Specify the level of confidence.** Suppose that a 95% confidence level is desired.

3. **Determine the z value** associated with the confidence level using Table 2 in the Appendix of Statistical Tables. For a 95% confidence level, the probability that the population mean will fall outside one end of the interval is 0.025 (0.05/2). The associated z value is 1.96.

4. **Determine the standard deviation of the population.** This may be known from secondary sources. If not, it might be estimated by conducting a pilot study. Alternatively, it might be estimated on the basis of the researcher's judgement. For example, the range of a normally distributed variable is approximately equal to ± 3 standard deviations, and one can thus estimate the standard deviation by dividing the range by 6. The researcher can often estimate the range based on knowledge of the phenomenon.

5. **Determine the sample size** using the formula for the standard error of the mean.

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

or

$$\sigma_{\bar{X}} = \frac{D}{z}$$

or

$$\frac{\sigma}{\sqrt{n}} = \frac{D}{z}$$

or

$$n = \frac{\sigma^2 \times z^2}{D^2}$$

In our example,

$$n = \frac{55^2(1.96)^2}{5^2}$$

$$= 464.83$$
$$= 465 \text{ (rounded to the next highest integer)}$$

It can be seen from the formula for sample size that sample size increases with an increase in the population variability, the degree of confidence, and the precision level required of the estimate.

6. If the resulting sample size represents 10% or more of the population, the finite population correction (fpc) should be applied. The required sample size should then be calculated from the formula.

$$n_c = \frac{nN}{N + n - 1}$$

where n = sample size without fpc
n_c = sample size with fpc

7. If the population standard deviation is unknown and an estimate is used, it should be re-estimated once the sample has been drawn. The sample standard deviation, s, is used as an estimate of A revised confidence interval should then be calculated to determine the precision level actually obtained.

Suppose that the value of 55.00 used for was an estimate because the true value was unknown. A sample of n = 465 is drawn, and these observations generate a mean X of 180.00 and a sample standard deviation s of 50.00. The revised confidence interval is then:

$$\bar{X} \pm zs_{\bar{X}} = 180.00 \pm 1.96 \times \frac{50.0}{\sqrt{465}}$$
$$= 180.00 \pm 4.55$$

or $175.45 \leq \mu \leq 184.55$. Note that the confidence interval obtained is narrower than planned, because the population standard deviation was overestimated, as judged by the sample standard deviation.

8. In some cases, precision is specified in relative rather than absolute terms. In other words, it may be specified that the estimate be within plus or minus R percentage points of the mean. Symbolically,

$$D=R\mu \quad \sigma/\mu$$

in these case, the sample size may be determined by

$$n = \frac{\sigma^2 \times z^2}{D^2}$$

$$= \frac{C^2 \times z^2}{R^2}$$

where the cofficient of variation $C = \sigma/\mu$ would have to be estimated.

The population size, N, does not directly affect the size of the sample, except when the finite population correction factor has to be applied. Although this may be counter intuitive, upon reflection it makes sense. For example, if all the population elements are identical on the characteristics of interest, then a sample size of one will be sufficient to estimate the mean perfectly. This is true whether there are 50, 500, 5,000 or 50,000 elements in the population. What directly affects the sample size is the variability of the characteristic in the population. This variability enters into the sample size calculation by way of population variance σ^2 or sample variance s^2.

4.6.2 Sample Size Determination: Proportions

If the statistic of interest is a proportion rather than a mean, the approach to sample size determination is similar. Suppose that the researcher is interested in estimating the pro portion of households possessing a debit card. The following steps should be followed:

1. **Specify the level of precision.** Suppose that the desired precision is such that the p allowable interval is set as $D = p - \pi = \pm 0.05$.

2. **Specify the level of confidence.** Suppose that a 95% confidence level is desired.

3. Determine the z value associated with the confidence level. As explained in the case of estimating the mean, this will be $z = 1.96$.

4. Estimate the population proportion As explained earlier, the population proportion may be estimated from secondary sources, or from a pilot study, or may be based on the judgement of the researcher. Suppose that based on secondary data the researcher estimates that 64% of the households in the target population pos psesa debit card. Hence, $\pi = 0.64$.

5. Determine the sample size using the formula for the standard error of the proportion.

$$\sigma_p = \frac{p - \pi}{z}$$

$$= \frac{D}{z}$$

$$= \sqrt{\frac{\pi(1-\pi)}{n}}$$

or

$$n = \frac{\pi(1-\pi)z^2}{D^2}$$

In our example,

$$n = \frac{0.64(1-0.64)(1.96)^2}{(0.05)^2}$$

$$= 354.04$$

$$= 355 \text{ (rounded to the next highest integer)}$$

6. If the resulting sample size represents 10% or more of the population, the finite population correction (fpc) should be applied. The required sample size should then be calculated from the formula.

$$n_c = \frac{nN}{N + n - 1}$$

Where n – sample size without fpc

n_c = sample size with fpc

7. If the estimate of turns out to be poor, the confidence interval will be more or less precise than desired. Suppose that after the sample has been taken, the proportion p is calculated to have a value of 0.55. The confidence interval is then re-estimated by employing s_p to estimate the unknown σ_p as:

$$p \pm zs_p$$

where

$$S_p = \sqrt{\frac{p(1-p)}{n}}$$

In our example

$$S_p = \sqrt{\frac{0.55(1-0.55)}{355}}$$

$$= 0.0264$$

The confidence interval, then, is

$$0.55 \pm 1.96(0.0264) = 0.55 \pm 0.052$$

which is wider than that specified. This is because the sample standard deviation based on p = 0.55 was larger than the estimate of the population standard deviation based on

$\pi = 0.64$.

If a wider interval than specified is unacceptable, the sample size can be determined to reflect the maximum possible variation in the population. This occurs when the product is the greatest, which happens when is set at 0.5. This result can also be seen intuitively. Since one half of the population has one value of the characteristic and the other half the other value, more evidence would be required to obtain a valid inference than if the situation was more clear cut and the majority had one particular value. In our example, this leads to a sample size of:

$$n = \frac{0.5(0.5)(1.96)^2}{(0.05)^2}$$

$$= 384.16$$

$$= 385 \text{ (rounded to the next higher integer)}$$

8. Sometimes, precision is specified in relative rather than absolute terms. In other words, it may be specified that the estimate be within plus or minus R percentage points of the

population proportion. Symbolically,

$$D=R\pi$$

In such a case, the sample size may be determined by

$$n=\frac{z^2(1-\pi)}{R^2\pi}$$

4.7 SUMMARY

The objective of this lesson is to know about sampling and non-sampling errors as well as to discuss about the effectiveness of sampling and determination of sample size. Sampling error includes systematic error and random error. Systematic error occurs when the sample is not properly drawn (an error of the researcher). Random error is the degree to which the sample is not perfectly representative of the population. Even with the best sampling techniques, some degree of random error is expected. We have studied different methods to sample from population, viz., simple random sample, stratified sample, cluster sample etc. Each of these involves randomness in the sample-selection process, so the estimated mean or proportion is unlikely to be exactly the same as the underlying population parameter that is being estimated. When sampling books from a library or sacks of rice from the output of a factory, sampling error is the main or only type of error. Further, when sampling from some types of population- especially human populations - problems often arise when conducting one of the above sampling schemes. For example, some sampled people are likely to refuse to participate in your study. Such difficulties also result in errors and these are called non-sampling errors. Non-sampling errors can be much higher than sampling errors and are much more serious. Unlike sampling errors, the size of non-sampling errors cannot be estimated from a single sample-it is extremely difficult to assess their likely size.

Non-sampling errors often distort estimates by pulling them in one direction. It is therefore important to design a survey to minimise the risk of non-sampling errors. In the end this discussion, it is true to say that sampling error is one which is completely related to the sampling design and can be avoided, by expanding the sample size. Conversely, non-sampling error is a basket that covers all the errors other than the sampling error and so, it is unavoidable by nature as it is not possible to completely remove it. Also, in this lesson we have discussed about two approaches for determining the sample size i.e., the number

of individuals to form a sample on the basis of which inferences have been drawn about the population.

Summary of Sample Size Determination For Means and Proportions

Steps	Means	Proportions
1 Specify the level of precision.	$D = \pm €5.00$	$D = p - \pi = \pm 0.05$
2 Specify the confidence level (CL).	CL = 95%	CL = 95%
3 Determine the z value associated with the CL.	z value is 1.96	z value is 1.96
4 Determine the standard deviation of the population.	Estimate σ: $\sigma = 55$	Estimate π: $\pi = 0.64$
5 Determine the sample size using the formula for the standard error.	$n = \frac{\sigma^2 z^2}{D^2}$ $n = \frac{55^2(1.96)^2}{5^2}$ $= 465$	$n = \frac{\pi(1-\pi)z^2}{D^2}$ $n = \frac{0.64(1-0.64)(1.96)^2}{(0.05)^2}$ $= 355$
6 If the sample size represents 10% of the population, apply the finite population correction (fpc).	$n_c = \frac{nN}{N+n-1}$ $= \bar{X} + zs_{\bar{x}}$	$n_c = \frac{nN}{N+n-1}$ $p \pm zs_p$
7 If necessary, re-estimate the confidence interval by employing s to estimate σ.	$D = R\mu$	$D = R\pi$
8 If precision is specified in relative rather than absolute terms, determine the sample size by substituting for D.	$n = \frac{C^2 z^2}{R^2}$	$n = \frac{z^2(1-\pi)}{R^2\pi}$

4.8 GLOSSARY

- **Confidence Interval:** The confidence interval is a range of values, above and below a finding, in which the actual value is likely to fall. The confidence interval represents the accuracy or precision of an estimate

- **Random Sampling Error:** Sampling error is the error caused by observing a sample instead of the whole population. The sampling error is the difference between a sample statistic used to estimate a population parameter and the actual but unknown value of the parameter.

- **Non-Sampling Error:** Non-sampling error is a catch-all term for the deviations of estimates from their true values that are not a function of the sample chosen, including various systematic errors and random errors that are not due to sampling.

- **Sample Size:** The sample size is a term used in market research for defining the number of subjects included in a sample size.

4.9 SELF ASSESSMENT QUESTIONS

(i) Please Tick (✓)the correct answer:-

1. Which of the following is not an example of non-sampling risk?

   a) Failing to evaluate results properly
   b) Use of an audit procedure inappropriate to achieve a given audit objective
   c) Obtaining an unrepresentative sample
   d) Failure to recognise an error

2. Why do sampling errors occur?

   a) Differences between sample and population
   b) Differences among samples themselves
   c) Choice of elements of sampling
   d) all of the above

3. Any calculation on the sampling data is called:

   (a) Parameter (b) Static
   (c) Mean (d) Error.

4. Probability distribution of a statistics is called:

   (a) Sampling (b) Parameter
   (c) Data (d) Sampling distribution

5. In probability sampling, probability of selecting an item from the population is known and is:

   (a) Equal to zero (b) Non zero
   (c) Equal to one (d) All of the above

6. Standard deviation of sample mean without replacement_____ standard deviation of sample mean with replacement:

(a) Less than (b) More than

(c) 2 times (d) Equal to

4.10 LESSON END EXERCISE

1. Discuss how sampling errors vary with the size of the sample.

2. List and discuss the various forms of non-sampling errors.

3. What suggestions can you provide to minimise non-sampling errors ?

4. Explain the process used for determining the sample size.

5. Define:
   i. Statistical Errors
   ii. Biased and Unbiased Errors
   iii. Absolute and Relative Errors

6. What are statistical errors? What are the sources of errors? Explain the methods of measuring them.

4.11 SUGGESTED READING

- Levin, R.I. Robin, D.S. *Statistics for Management*, Prentice-Hall of India. New Delhi.
- Aczel, A. D. Sounderpandian, J. *Complete Business Statistics*, Mc Graw Hill Publishing. New Delhi. downloaded
- Anderson, S., W. *Statistics for Business and Economics*, Cengage Learning. New Delhi.
- Kazmeir L. J. *Business Statistics*, Tata Mc Graw Hill. New Delhi.
- Vohra, N. D. *Business Statistics*. Tata Mc Graw Hill. New Delhi.
- Freund, J. E., Williams, F.M. *Elementary Business Statistics*-The Modern Approach. Prentice Hall of India Private Ltd., New Delhi.

SAMPLING DISTRIBUTION AND STANDARD ERROR, PILOT AND FINAL SURVEY, PRECAUTIONS IN DATA COLLECTION

STRUCTURE

5.1 INTRODUCTION

Having discussed the various methods available for picking up a sample from a population, we would naturally be interested in drawing statistical inferences-making generalisations about the population on the basis of a sample drawn from it. The generalizations to be made about the population are usually either by way of; estimating the unknown population parameters, or testing appropriate hypotheses stated in relation to population parameters in the light of sample data. These generalizations, together with the measurement of their reliability, are made in terms of the relationship between the values of any sample statistic and those of the corresponding population parameters. Population parameter is any number computed (or estimated) for the entire population viz. population mean, population median, population proportion, population variance and so on. Population parameter is unknown but fixed, whose value is to be estimated from the sample statistic that is known but random. Sample Statistic is any numbers computed from our sample data viz. sample mean, sample median, sample proportion, sample variance and so on. It may be appreciated that no single value of the sample statistic is likely to be equal to the corresponding population parameter. This owes to the fact that the sample statistic being random, assumes different values in different samples of the same size drawn from the same population. As we all know that any sample statistics is a random variable and therefore, has a probability distribution better known as the Sampling Distribution of the statistic.

Also, one of the most important areas of research tools in the field of applied social science is the 'survey research'. It is one of the most relevant techniques basically used for collecting data and involves any measurement procedures that prominently include asking questions from respondents or the subjects selected for the research study. The term "survey" can be defined as a process which may involve an investigation/ examination or assessment in the form of a short paper and-pencil feedback form to an intensive one-on-one in-depth interview. With the help of the questionnaire or other statistical tools, the method tries to gather data about people, their thoughts and behaviours. This lesson tries to focus on the concept of sampling distribution and standard error, pilot and final survey as well as the precautions need to be taken in data collection.

5.2 OBJECTIVES

After reading this lesson, you will be able to:

- Understand the concept of standard distribution and standard error.
- Know about the basics of pilot and final survey.
- Understand the various precautions to be taken while collecting data.

5.3 SAMPLING DISTRIBUTION AND STANDARD ERROR

The sampling distribution of a statistic is the probability distribution of all possible values the statistic may take when computed from random samples of the same size drawn from a specified population. In reality, of course we do not have all possible samples and all possible values of the statistic. We have only one sample and one value of the statistic. This value is interpreted with respect to all other outcomes that might have happened, as represented by the sampling distribution of the statistic. In this lesson, we will refer to the sampling distributions of only the commonly used sample statistics like sample mean, sample proportion, etc., which have a role in making inferences about the population.

Sample statistics form the basis of all inferences drawn about populations. Thus, sampling distributions are of great value in inferential statistics. The sampling distribution of a sample statistic possess well-defined properties which help lay down rules for making generalizations about a population on the basis of a single sample drawn from it. The variations in the value of sample statistic not only determine the shape of its sampling distribution, but also account for the element of error in statistical inference. If we know the probability distribution of the sample statistic, then we can calculate risks (error due to chance) involved in making generalizations about the population. With the help of the properties of sampling distribution of a sample statistic, we can calculate the probability that the sample statistic assumes a particular value (if it is a discrete random variable) or has a value in a given interval. This ability to calculate the probability that the sample statistic lies in a particular interval is the most important factor in all statistical inferences. We will demonstrate this by an example.

Suppose we know that 40% of the population of all users of hair oil prefers our brand to

the next competing brand. A "new improved" version of our brand has been developed and given to a random sample of 100 users for use. If 55 of these prefer our "new improved" version to the next competing brand, what should we conclude? For an answer, we would like to know the probability that the sample proportion in a sample of size 100 is as large as 55% or higher when the true population proportion is only 40%, i.e. assuming that the new version is no better than the old. If this probability is quite large, say 0.5, we might conclude that the high sample proportion viz. 55% is perhaps because of sampling errors and the new version is not really superior to the old. On the other hand, if this probability works out to a very small figure, say 0.001, then rather than concluding that we have observed a rare event we might conclude that the true population proportion is higher than 40%, i.e. the new version is actually superior to the old one as perceived by members of the population. To calculate this probability, we need to know the probability distribution of sample proportion i.e. the sampling distribution of the proportion.

5.3.1 Sampling Distribution of the Mean

Suppose we have a simple random sample of size n, picked up from a population of size N. We take measurements on each sample member in the characteristic of our interest and denote the observation as x_1, x_2,..... x_n respectively. The sample mean for this sample is defined as:

$$X = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

If we pick up another sample of size n from the same population, we might end up with a totally different set of sample values and so a different sample mean. Therefore, there are many (perhaps infinite) possible values of the sample mean and the particular value that we obtain, if we pick up only one sample, is determined only by chance. In other words, the sample mean is a random variable. The possible values of this random variable depends on the possible values of the elements in the random sample from which sample mean is to be computed. The random sample, in turn, depends on the distribution of the population from which it is drawn. As a random variable, X has a probability distribution. This probability distribution is the sampling distribution of X, i.e. the mean. Therefore, the sampling distribution of X is the probability distribution of all possible values the random variable X may take when a sample of size n is taken from a specified population. To observe the

distribution of X empirically, we have to take many samples of size n and determine the value of X for each sample. Then, looking at the various observed values of X, it might be possible to get an idea of the nature of the distribution.

5.3.2 Standard Error

The standard deviation of X is also called the standard error of the mean. It indicates the extent to which the observed value of sample mean can be away from the true value, due to sampling errors. For example, if the standard error of the mean is small, we may be reasonably confident that whatever sample mean value we have observed cannot be very far away from the true value. However, standard deviation (SD) measures the amount of variability, or dispersion, from the individual data values to the mean, while the standard error of the mean measures how far the sample mean (average) of the data is likely to be from the true population mean. The standard error of mean is always smaller than the Standard deviation.

Standard deviation and standard error are both used in all types of statistical studies, including those in finance, medicine, biology, engineering, psychology, etc. In these studies, the standard deviation (SD) and the estimated standard error of the mean (SEM) are used to present the characteristics of sample data and to explain statistical analysis results. However, some researchers occasionally confuse the SD and SEM. Such researchers should remember that the calculations for SD and SEM include different statistical inferences, each of them with its own meaning. SD is the dispersion of individual data values. In other words, SD indicates how accurately the mean represents sample data. However, the meaning of SEM includes statistical inference based on the sampling distribution. Standard error of the mean is the SD of the theoretical distribution of the sample means (the sampling distribution).

Calculating of Standard Deviation (SD) :

$$\text{standard deviation } \sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

$$\text{variance} = \sigma^2$$

$$\text{standard error } (\sigma_{\bar{x}}) = \frac{\sigma}{\sqrt{n}}$$

where:

$\bar{x}$ = the sample's mean

n = the sample size

The formula for the SD requires a few steps:

1. First, take the square of the difference between each data point and the sample mean, finding the sum of those values.

2. Then, divide that sum by the sample size minus one, which is the variance.

3. Finally, take the square root of the variance to get the SD.

Standard error of mean is calculated by taking the standard deviation and dividing it by the square root of the sample size.

Standard Error of Mean = $SD/\sqrt{n}$

Where, SD = Standard Deviation and n = sample size

Standard error gives the accuracy of a sample mean by measuring the sample-to-sample variability of the sample means. The SEM (standard error of mean) describes how precise the mean of the sample is as an estimate of the true mean of the population. As the size of the sample data grows larger, the SEM decreases versus the SD; hence, as the sample size increases, the sample mean estimates the true mean of the population with greater precision. In contrast, increasing the sample size does not make the SD necessarily larger or smaller, it just becomes a more accurate estimate of the population SD.

5.4 PILOT AND FINAL SURVEY

The method of survey research is a non-experimental (that is, it does not involves any observation under controlled conditions), descriptive research method which is one of the

quantitative method used for studying of large sample. In a survey research, the researcher collects data with the help of standardised questionnaires or interviews which is administered on a sample of respondents from a population (population is sometimes referred to as the universe of a study which can be defined as a collection of people or object which possesses at least one common characteristic). The method of survey research is one of the techniques applied for social research which can be helpful in collection of data both through direct (such as a direct face to face interview) and indirect observation (such as opinions on library services of an institute).

Therefore, in research once the target population is defined by the researcher, he or she needs to design a survey research. On the basis of the framed design, the research decides to conduct a survey, selects instrument for survey (for example telephonic interview) with the help of which data will be collected. After the selection of the instrument, the researcher conducts a pilot study (a small survey taken in advance of a final investigation or final survey). The pilot study helps the researcher to analyse the significance and relevance of the instruments selected by the researcher for the present research.

5.4.1 Pilot Survey

A pilot study is very important in the success of the research. It is often recommended to address variety of issues such as the validity and reliability of the instrument (questionnaire or schedule) to be used in the study. The reliability and validity of the instrument are vital for the future result of the research. If the instrument is not reliable and not valid, then the result conducted is void. According to the Association for Qualitative Research, a pilot study is a small study conducted in advance of a final survey, specifically to test aspects of the research design and to allow necessary adjustment before final commitment to the design. Similarly, pilot studies are defined as a mini version of a full-scale study which is also known as feasibility studies. It is also defined as the process of pre-testing of the research instrument such as questionnaire, tests, and interview schedule. Conducting pilot study has several advantages; it gives advance warning about where the main project fail, which research procedure not followed, or whether the proposed methods or the questionnaire or instrument are inappropriate or too complicated to the respondents. The reasons why pilot studies should be conducted are:

1. Developing and testing adequacy of research instruments.
2. Assessing the feasibility of a (full-scale) study/survey.
3. Designing a research protocol.
4. Assessing whether the research protocol is realistic and workable.
5. Establishing whether the sampling frame and technique are effective.
6. Assessing the likely success of proposed recruitment approaches.
7. Identifying logistical problems which might occur using proposed methods
8. Estimating variability in outcomes to help to determine sample size.
9. Collecting preliminary data.
10. Determining what resources (finance, staff) are needed for a planned study
11. Assessing the proposed data analysis techniques to uncover potential problems
12. Developing a research question and research plan.
13. Training a researcher in as many elements of the research process as possible.
14. Convincing funding bodies that the research team is competent and knowledgeable.
15. Convincing funding bodies that the main study is feasible and worth funding.
16. Convincing other stakeholders that the main study is worth supporting.

A Pilot study is crucial to the success of the research. It should be done correctly so that it will give a correct description of the instrument to be used. Here are the steps used to pre-test a questionnaire on a small group of volunteers, who are as similar as possible to the target population:

- Administer the questionnaire to pilot subjects in exactly the same way as it will be administered in the main study.
- Ask the subjects for feedback to identify ambiguities and difficult questions.
- Record the time taken to complete the questionnaire and decide whether it is reasonable.

It's important to test your survey questionnaire before using it to collect final data. Pretesting and piloting can help you identify questions that don't make sense to participants, or problems with the questionnaire that might lead to biased answers.

1. Developing new measurement instruments for use in research (e.g. psychological instruments for measuring concepts such as confidence, motivation, etc.).

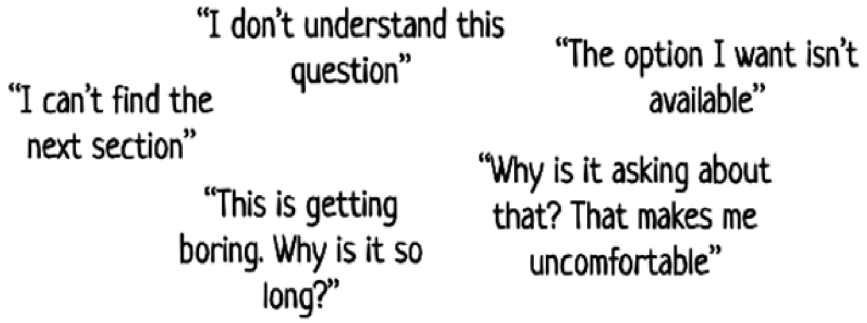
2. Qualitative focus groups or interviews.

3. Any testing is better than no testing: People often think that testing a survey takes a long time. They think they don't have the time or resources for it, and so they end up just running the survey without any testing. This is a big mistake. Even testing with one person is better than no testing at all. So if you don't have the time or resources to do everything in this guide, just do as much as you can with what you have available. As a general rule, you should aim to pre-test all your surveys and forms with at least 5 people. Even with this small number of people you'll be surprised how many improvements you can make. Piloting is only really needed for large or complex surveys, and it takes significantly more time and effort.

4. Pretesting: Once you've finished designing your survey questionnaire, find 5-10 people from your target group to pre-test it. If you can't get people from your exact target group then find people who are as close as possible. For example, if your target group is young people aged 15-25, try to include some who are younger, some who are older, boys and girls with different socioeconomic backgrounds.

   Although 5-10 people might not sound like many, you will usually find that most of them have the same problems with the survey. So even with this small number of people you should be able to identify most of the major issues. Adding more people might identify some additional smaller issues, but it also makes pretesting more time consuming and costly. Ask them to complete the survey while thinking out loud Once you've found your testers, ask them to complete the survey one at a time (they shouldn't be able to watch each other complete it). The testers should complete the survey the same way that it will be completed in the actual project. So if it's an online survey they should complete it online, if it's a verbal survey you should have a trained interviewer ask them the questions. While they are completing the survey ask them

to think out loud. Each time they read and answer a question they should tell you exactly what comes into their mind. Take notes on everything they say.



5. Observe how they complete the survey: You should also observe them completing the survey. Look for places where they hesitate or make mistakes, such as the example below. This is an indication that the survey questions and layout are not clear enough and need to be improved. Keep notes on what you observe.

6. Make improvements based on the results: Once all the testers have completed the survey review your notes from each session. At this point it's normally clear what the major problems are so you can go about improving the survey to address those problems. Normally this is all that's needed. However, if major changes are needed to the questions or structure it might be necessary to repeat the pretesting exercise with different people before starting the survey.

5.4.2 Advantages of Pilot Survey

1. It helps you to identify and address any issues that can affect the actual survey. For instance, a pilot survey can help uncover inappropriate questions that can negatively impact your survey results.

2. A pilot survey is a cost-effective method of data collection in the long run. Pilot surveys save you a lot of time and money because you can identify and correct any mistakes that would have ruined the overall data-gathering process.

3. With a pilot survey, you can measure the validity of the research process. Results from the survey can be used to determine the effectiveness of the research method; that is, how useful it is in retrieving information from the research subjects.

4. A pilot survey is an opportunity to test alternative methods and choose the ones that produce the most valid results for the actual research.

5. It is an opportunity to evaluate the different steps in your data collection and research process. From the pilot survey results, you can make any required changes to your data collection method to help you improve your research.

6. It often provides the researcher with ideas, approaches, and clues you may not have explored before conducting the pilot study.

7. A pilot survey is an opportunity to get valid feedback from members of the convenience sample to help you improve your data collection process.

5.4.3 Final Survey

The pre-test/pilot survey results should be used to improve the survey design and implementation plans. For example, in the pre-test of a telephone survey of four physician specialty groups, one of the authors obtained response rates in the range of 40% to 50%. In subsequent focus groups with physicians, the discussion concentrated on what could be done to increase the response rates. One suggestion that was followed was to give the physicians a choice of how to respond: they could either complete a mail questionnaire or be interviewed by telephone. The rationale was to let each physician choose the method that was best for the physician's schedule. The suggestion was a good one because the overall response rate for the main study increased to 67%. Pretesting may also help to decide how much time to allot between follow up contacts in a mail survey or whether the final contact in a mail survey should be by mail or by telephone. There are very few hard and fast rules in survey research. Researchers need to be attuned to the factors that can affect response rates and data quality and be prepared to make adjustments.

During this stage, final changes should be made in the sampling plan, the questionnaire, interviewer-training procedures and materials, data-coding plans, and plans for analyzing the data. For example, in sampling we may learn that we need to select more phone numbers to get the desired number of completed interviews because we found more nonworking phone numbers than expected in the pre-tests. For the questionnaire, we may find that changing the order of some questions improves the flow of the interview, or we

may be able to develop closed-ended response choices for some questions that were initially asked as open-ended questions. Another common occurrence is that we find the answers of a particular subgroup in the population, such as those older than age 55 years or black respondents, to be different from the responses of other subgroups or different from the responses expected; or we may find the number of completed interviews with these subgroups to be less than anticipated. Then we need to decide whether the sample sizes for these subgroups will be adequate or whether we need to oversample certain groups. After this stage, the researcher begins to monitor the results of the sampling and data collection activities and begin coding and data file preparation.

5.5 PRECAUTIONS IN DATA COLLECTION

We can understand that there are a lot of published and unpublished sources where researcher can get secondary data. However, the researcher must be cautious in using this type of data. The reason is that such type of data may be full of errors because of bias, inadequate size of the sample, errors of definitions etc. Bowley expressed that it is never safe to take published or unpublished statistics at their face value without knowing their meaning and limitations. Hence, before using secondary data, researchers must use the following precautions while collecting data.

1. **Suitability of Secondary Data:** Before using secondary data, you must ensure that the data are suitable for the purpose of your enquiry. For this, you should compare the objectives, nature and scope of the given enquiry with the original investigation. For example, if the objective of our enquiry is to study the salary pattern of a firm including perks and allowances of employees. But, secondary data is available only on basic pay. Such type of data is not suitable for the purpose of the study.

2. **Reliability of the Data:** For the reliability of secondary data, these can be tested: i) unbiasness of the collecting person, ii) proper check on the accuracy of field work, iii) the editing, tabulating and analysis done carefully, iv) the reliability of the source of information, v) the methods used for the collection and analysis of the data. If the data collecting organisations are government, semi-government and international, the secondary data are more reliable corresponding to data collected by individual and private organisations.

3. **Adequacy:** Data Adequacy of secondary data is to be judged in the light of the objectives of the research. For example, our objective is to study the growth of industrial production in India. But the published report provides information on only few states, and then the data would not serve the purpose. Adequacy of the data may also be considered in the light of duration of time for which the data is available. For example, for studying the trends of per capita income of a country, we need data for the last 10 years, but the information available for the last 5 years only, which would not serve our objective. Hence, we should use secondary data if it is reliable, suitable and adequate.

4. **Definition of Units:** The investigator must ensure that the definitions of units which are used by him are the same as in the earlier investigation.

5. **Degree of Accuracy:** The investigator should keep in mind the degree accuracy maintained by each investigator.

6. **Time and Condition of Collection of Facts:** It should be ascertained before making use of available data to which period and conditions, the data was collected.

7. **Comparison:** Investigator should keep in mind whether the secondary data' reasonable, consistent and comparable.

8. **Test Checking:** The use of the secondary data must do test checking and see that totals and rates have been correctly calculated.

9. **Homogeneous Conditions:** It is not safe to take published statistics at their face value without knowing their means, values and limitations.

5.6 SUMMARY

The objective of this lesson is to know about concept of sampling distribution and standard error of mean. Sampling distribution is the probability distribution of means of different samples drawn from a same population whereas standard error is the standard deviation of the sampling distribution of mean or proportion. Thereafter, the concept of pilot survey and final survey is discussed. A Pilot study is very much important in the success of the research. It identifies variety of issues about the instrument and the research in general.

Therefore, it should be taken with almost consideration. Pilot study should be carefully planned and analysed especially if the researcher uses researcher-made instrument. This instrument is subjected to determine its validity and reliability. No research-made questionnaire should be used to conduct research without it undergone pilot study. Adapted, modified or enhanced instrument should also be pilot tested especially if it is adapted from foreign authors. The diversity of the respondents, the culture, language, the age, the economic status and many more are factors that made group of individuals differ. As an example a questionnaire that asks about physical activity and uses skiing as an example may not be relevant in settings where there is no snow. The lesson is also discussed about the various precautions which should be kept in mind by the researcher while collecting data both from primary as well as secondary sources.

5.7 GLOSSARY

- **Sample Distribution:** The distribution of the values of a sample statistic computed for each possible sample that could be drawn from the target population under a specified sampling plan.
- **Standard Error:** The standard deviation of the sampling distribution of the mean or proportion.
- **Pre-testing:** Testing the questionnaire on a small sample of respondents for the purpose of improving the questionnaire by identifying and eliminating potential problems.
- **Secondary Data:** Data collected for some purpose other than the problem at hand.
- **Primary Data:** Data originated by the researcher specifically to address the research problem.

5.8 SELF ASSESSMENT QUESTIONS

(i) Please Tick (✓)the correct answer:-

1. Probability distribution of a statistics is called:

   (a) Sampling (b) Parameter
   (c) Data (d) Sampling distribution

2. In probability sampling, probability of selecting an item from the population is known and is:

(a) Equal to zero (b) Non zero
(c) Equal to one (d) All of the above

3. Standard deviation of sample mean without replacement_____ standard deviation of sample mean with replacement:

(a) Less than (b) More than
(c) 2 times (d) Equal to

4. _____ must be ensured to reduce the impact of businesses that may lead to incorrect conclusions.

(a) Adequacy of data (b) Accuracy of data
(c) Both (a) and (b) (d) None of the above

5. Which of the following are the precautions that should be taken before using secondary data?

a) The data must be from a reliable agency.
b) The data must be suitable for the purpose of enquiry.
c) The data must be adequate.
d) All of the above

6. The _____ method of collecting data may be biased, depending upon the mode of selection of samples.

(a) Census (b) Sampling
(c) Both (a) and (b) (d) None of the above

5.9 LESSON END EXERCISE

1. Distinguish between primary and secondary data.

_____
_____
_____

2. Explain the concept of sampling distribution.

3. What precautions a researchers must take while collecting data?

4. Explain the concept of standard error.

5. Define:

   i. Pre-testing
   ii. Questionnaire
   iii. Schedule

6. What is the need for pre-testing the drafted questionnaire?

5.10 SUGGESTED READING

- Levin, R.I. Robin, D.S. *Statistics for Management*, Prentice-Hall of India. New Delhi.

- Aczel, A. D. Sounderpandian, J. *Complete Business Statistics*, Mc Graw Hill Publishing. New Delhi. downloaded
- Anderson, S., W. *Statistics for Business and Economics*, Cengage Learning. New Delhi.
- Freund, J. E., Williams, F.M. *Elementary Business Statistics*-The Modern Approach. Prentice Hall of India Private Ltd., New Delhi.
- Clave, B. S. *Statistics for Business and Economics* - Prentice Hall Publication, New Delhi

PROBABILITY AND ANALYSIS OF VARIANCE

CONCEPT AND ROLE OF PROBABILITY; APPROACHES OF PROBABILTIY: CASSICAL, RELATIVE FREQUENCY, SUBJECTIVE AND AXIOMATIC

STRUCTURE

6.1 INTRODUCTION

The word 'probability' or 'chance' is very commonly used in day-to-day conversation and generally people have a vague idea about its meaning. For example, we come across statements like "probably it may rain tomorrow", "It is likely that Mr. X may not come for taking his class today", "The chances of teams A and B winning a certain match are equal", "Probably you are right", "It is possible that I may not able to join you at the tea party". All these terms - possible, probably, likely, etc., convey the same sense, i.e. the event is not certain to take place or, in other words, there is uncertainty about happening of the event in question. In a layman language terminology the word "probability" thus connotes that there is uncertainty about the happening of event. However, in mathematics and statistics we try to present conditions under which we can make sensible numerical statements about uncertainty and apply certain methods of calculating numerical values of probabilities and expectations. The theory of probability has its origin in the games of chance related to gambling such as throwing a die, tossing a coin, drawing cards from a pack of cards, etc. Jerame Cardon (1501-76), an Italian mathematician, was the first man to write a book on the subject entitled Book on Games of Chance (Liberde LudoAleae) which was published after his death in 1663.

6.2 OBJECTIVES

After reading this lesson, you will be able to:

- Appreciate the relevance of probability theory in decision making.
- Understand the different approaches to probability.
- Calculate probabilities in different situations.
- Understand different terms used in probability theory.

6.3 CONCEPT OF PROBABILITY

Galileo (1564-1642), an Italian mathematician, was the first man to attempt quantitative measure of probability while dealing with some problems related to the theory of dice in gambling. However, systematic and scientific foundation of the mathematical theory of probability was laid in mid-seventeenth century by two French Mathematicians B. Pascal

(1623-62) and Pierre de Fermat (1601-65). Swiss mathematician James Bernoulli (1654-1705) was another stalwart who made extensive study of the subject for over two decades and his treatise on probability (Arts Conjectandi) published in 1713 is a major contribution to the theory of probability. De Moivre' (1667-1754) contributed a lot to the subject and his work published in 1718 is the famous book, The Doctrine of Chances. Thomas Bayes (1702-61), introduced the concept of inverse probability. Pierre Simon de Laplace (1749-1827) after extensive research published his, monumental work in 1812 "Theone Analytique des probabilityes (Theory of Analytical Probability) which constitutes the classical theory of probability. R.A. Fisher and Von Mises introduced empirical approach to probability through the notion of sample space. The modern theory of probability was developed by eminent Russian mathematicians like Chebychev (1821- 94), A. Markov (1856-1922) and A. N. Kolmogorov. Kolmogorov axiomed the theory of probability and his book 'Foundations of Probability' published in 1933 introduced probability as a set function and is regarded as a classic. Starting with games of chance, probability today has become one of the important tools of statistics. In fact, statistics and probability are so fundamentally interrelated that it is difficult to discuss statistics without an understanding of the meaning of probability. A knowledge of probability theory makes it possible to interpret statistical results, since many statistical procedures involve conclusions based on samples which are always affected by random variation, and it is by means of probability theory that we can express numerically the inevitable uncertainties in the resulting conclusions. Probability theory is being applied in the solution of social, economic, political and business problems. The insurance industry, which emerged in the 19th century, required precise knowledge about the risk of loss in order to calculate premium. Within a few decades many learning centers were studying probability as a tool for understanding social phenomena.

Today the concept of probability has assumed great importance and the mathematical theory of probability has become the basis for statistical applications in both social and decision-making research. In fact probability has become a part of our everyday life. In personal and Management decisions, we face uncertainty and use probability theory, whether or not we admit the use of something so sophisticated. To quote Levin we live in a world in which we are unable to forecast the future with complete certainty. Our need to cope with uncertainty leads us to the study and use of Probability theory. In many instances we, as concerned citizens will have some knowledge about the possible outcomes of a

decision. By organizing this information and considering it systematically, we will be able to recognise our assumptions, communicate our reasoning to others and make a sound decision than we could by using a shot-in-the-dark approach. Probability theory, in fact, is the foundation of statistical inference.

Definition of Probability

The probability of a given event is an expression of likelihood or chance of occurrence of an event. A probability is a number which ranges from 0 (zero) to 1 (one), 0 for an event which cannot occur and 1 for an event certain to occur. How the number is assigned would depend on the interpretation of the term 'probability'.

6.3.1 Role of Probability

Since its humble beginning at the gambling tables in seventeenth century, probability theory has been developed and employed to treat and solve many weighty problems. It is the foundation of the classical decision procedures of estimation and testing. Probability models can be very useful for making predictions. It is concerned with the construction of econometric models with managerial decisions on planning and control, with the occurrence of accidents of all kinds and with random disturbances in an electrical mechanism. It is involved in the observation of the life span of a radioactive atom, of the phenotypes of the off-spring, the crossing of two species of plants, the discussion about sex of an unborn baby, etc.

In fact, it has become an indispensable tool for all types of formal studies that involve uncertainty. It should be noted that the concept of probability is employed not only for various type of scientific investigations but also for many problems in everyday life. It will not be an exaggeration to say that probability has become a part of our everyday life whether or not we admit or are conscious of the use of something so sophisticated. It is still a dream to forecast the future with 100 percent certainty in any decision problem. The probability theory provides a media of coping up with uncertainty. Further, the concept of probability is of great importance in everyday life. Statistical analysis is based on this valuable concept. In fact the role played by probability in modern science is that of a substitute for certainty.

1. The probability theory is very much helpful for making prediction. Estimates and predictions form an important part of research investigation. With the help of statistical methods, we make estimates for the further analysis. Thus, statistical methods are largely dependent on the theory of probability.
2. It has also immense importance in decision making.
3. It is concerned with the planning and controlling and with the occurrence of accidents of all kinds.
4. It is one of the inseparable tools for all types of formal studies that involve uncertainty.
5. The concept of probability is not only applied in business and commercial lines, rather than it is also applied to all scientific investigation and everyday life.
6. Before knowing statistical decision procedures one must have to know about the theory of probability.
7. The characteristics of the Normal Probability. Curve is based upon the theory of probability.

6.3.2 Basic Terms of Probability

Before discussing the procedure for calculating probability it is necessary to define certain terms as given below:

1. Experiment and Events: - The term experiment refers to an act which can be repeated under some given conditions. Random experiments are those experiments whose results depend on chance such as tossing of a coin, throwing of dice. The results of a random experiment are called outcomes. If in an experiment all the possible outcomes are known in advance and none of the outcomes can be predicted with certainty, then such an experiment is called a random experiment and the outcomes as events or chance events. Events are generally denoted by capital letters A, B, C, … etc. An event whose occurrence is inevitable when a certain random experiment is performed is called a certain or sure event. An event which can never occur when a certain random experiment is performed is called an impossible event. For example, in a toss of a dice the occurrence of anyone of the numbers 1, 2, 3, 4, 5, 6 is a sure event, while occurrence of 8 is an impossible event.

An event which mayor may not occur while performing a certain random experiment is known as a random event. Occurrence of 2 is a random event in the above experiment of tossing of a dice.

2. Mutually Exclusive Events: Two events are said to be mutually exclusive or incompatible when both cannot happen simultaneously in a single trial or in other words, the occurrence of anyone of them precludes the occurrence of the other. For example, if a single coin is tossed either head can be up or tail can be up, both cannot be up at the same time. Similarly, a person may be either alive or dead at a point of time-he cannot be both alive as well as dead at the same time. To take another example, if we toss a dice and observe 3, we cannot expect 5 also in the same toss of dice. Symbolically, if A and B are mutually exclusive events P (AB) = 0.

3. Independent and Dependent Events: Two or more events are said to be independent when the outcome of one does not affect and is not affected by the other. For example, if a coin is tossed twice, the result of the second throw would in no way be affected by the result of the first throw. Similarly, the results obtained by throwing a dice are independent of the results obtained by drawing an ace from a pack of cards. To consider two events that in not independent. Let A stand for a firm's spending large amount of money on advertisement and B for its showing an increase in sales. Of course, advertising does not guarantee higher sales, but the probability that the firm will show an increase in sales will be higher if A has taken place. Dependent events are those in which the occurrence or non-occurrence of one event in anyone trial affects the probability of other events in other trials. For example, if a card is drawn from a pack of playing cards and is not replaced, this will alter the probability that the second card drawn is, say an ace. Similarly, the probability of drawing a queen from a pack of 52 cards is 4/52 or 1/13. But if the card drawn (queen) is not replaced in the pack, the probability of drawing again a queen is 3/51 (since the pack now contains only 51 cards out of which there are 3 queens).

4. Equally Likely Events: Events are said to be equally likely when one does not occur more often than the others. For example. If an unbiased coin or dice is thrown, each face may be expected to be observed approximately the same number of times in the long run. Similarly, the cards of a pack of playing cards are so closely alike that we expect each

card to appear equally often when a large number of drawings are made with replacement. However, if the coin or the dice is biased we should not expect each face to appear exactly the same number of times.

5. Simple and Compound Events: In case of simple events we consider the probability of the happening or not happening of single events. For example, we might be interested in finding out the probability of drawing a red ball from a bag containing 10 white and 6 red balls. On the other hand, in case of compound events we consider the joint occurrence of two or more events. For example, if a bag contains 10 white and 6 red balls and if two successive draws of 3 balls are made, we shall be finding out the probability of getting 3 white balls in the first draw and 3 black balls in the second draw-we are thus dealing with a compound event.

6. Exhaustive Events: Events are said to be exhaustive when their totality includes all the possible outcomes of a random experiment. For example, while tossing a dice, the possible outcomes are 1, 2, 3, 4, 5, 6 and hence the exhaustive number of cases is 6. If two dice are thrown once, the possible outcomes are:

Table 1

(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)

The sample space of the experiment i.e., 36 ordered pairs (62). Similarly for a throw of 3 dice exhaustive number of cases will be 216 (i.e. 63) and for n dice they will be 6n. Similarly, black and red cards are examples of collectively exhaustive events in a draw from a pack of cards.

7. Complementary Events: Let there are two events A and B. A is called the complementary event of B if A and B are mutually exclusive and exhaustive. For example,

when a dice is thrown, occurrence of an even number (2, 4, 6) and odd number (1, 3, 5) are complementary events. Simultaneous occurrence of two events A and B is generally written as AB.

6.4 APPROACHES OF PROBABILITY

There are four different schools of thought on the concept of probability.

6.4.1 Classical Approach

The classical approach to probability is the oldest and simplest. It originated in eighteenth century in problems pertaining to games of chance, such as throwing of coins, dice or deck of cards, etc. The basic assumption underlying the classical theory is that the outcomes of a random experiment are 'equally likely'. The 'event' whose probability is sought consists of one or more possible outcomes of the given activity such as when a die is rolled once, anyone of the six possible outcomes i.e., 1, 2, 3, 4, 5, 6 can occur. These activities are referred to in modern terminology as "experiment which is a term that refers to processes which result in different possible outcomes or observations. The term "equally likely", though undefined, conveys the notion that each outcome of an experiment has the same chance of appearing as any other. Thus in a throw of a dice occurrence of 1, 2, 3, 4, 5, 6 are equally likely events.

The definition of probability given by French Mathematician, Laplace and generally adopted by disciples of the classical school runs follows:

Probability it is said, is the ratio of the number of "favorable" cases to the total number of equally likely cases. If probability of occurrence of A is denoted by p(A), then by this definition we have:

$$P(A) = \frac{\text{No. of favorable cases/chances}}{\text{Total Number of Cases/chances}}$$

For calculating probability we have to find the following things:

1. Number of favourable cases.
2. Total number of equally likely cases.

For example, if a coin is tossed, there are two equally likely results, a head or a tail,

hence the probability of a head is 1/2. Similarly, if a dice is thrown, the probability of obtaining an even number is 3/6 or 1/2 since three of the six equally possible results are even numbers. Symbolically, if an event A can happen in "a" ways out of a total of 'n' equally likely and mutually exclusive ways then the probability of occurrence of the event (called its success) is denoted by: P = P (A) = $\frac{a}{n}$ and the probability of non occurrence of the event (called its failures) is given by:

q = P (not A) = 1- = n-a/n

Since the sum of the successful and unsuccessful outcomes is equal to the total number of events, we have

a + b = n

Dividing by n

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

Probability, therefore, may be written as a ratio. The numerator of the fraction corresponding to this ratio represents the number of successful (or unsuccessful) outcomes, while the denominator represents the total number of possible outcomes. The scale of probability extends from zero to unity. When p = 0, it denotes impossibility of the event taking place, i.e., the event cannot take place. However, this is true only when the number of possible outcomes is finite. For example, the probability of throwing seven with a single dice is zero. On the other hand, when p=1 it denotes certainty, i.e., the event is bounded to take place. In practical life the probability lies between these two extremes values 0 and 1. Classical probability is often called a priori probability because if we keep using orderly examples of unbiased dice, fair coin, etc. we can state the answer in advance (a priori) without rolling a dice, tossing a coin, etc.

Example 1. From a bag containing 10 black and 20 white balls, a ball is drawn at random. What is the probability that it is black?

Solution : Total number of balls in the bag = 10 + 20 = 30.

Number of black balls = 10

Probability of getting a black ball =
p(A) = Number of favourable cases = 10/ 30 = 1/3= P

Total Number of cases

Probability of not getting a black ball = q = P (not A) = 20/30 = 2/3

Thus, p + q = 1/3 + 2/3 = 1

Limitations of Classical Approach

The classical definition of probability given above suffers from certain limitations.

1. The definition cannot be applied whenever it is not possible to make a simple enumeration of cases which can be considered equally likely. For example, how does it apply to probability of rain? What are the possible cases? We might think that there are two possibilities 'rain' or 'no rain'. But at any given time it will not usually be agreed that they are equally likely. Similarly, if a person jumps from the top of Qutab Minar the probability of his survival will not be 50% since survival and death i.e., the two mutually exclusive and exhaustive outcomes, are not equally likely.

2. The classical approach also fails to answer questions like "What is the probability that a male will die before the age of 60?" "What is the probability that a bulb will burn less than 2,000 hours?" etc. All these are legitimate questions unlikely which we want to bring into the real m of probability theory. Real life situations are disorderly as they often are make it difficult and at times impossible to apply classical concept.

6.4.2 Relative Frequency Approach

In the 1800's "British statisticians, interested in a theoretical foundation for calculating risk of losses in life insurance and commercial insurance began defining probabilities from statistical data collected on births and deaths. Today this approach is called relative frequency of occurrence. This classical definition is difficult or impossible to apply as soon as we deviate from the fields of coins, dice, cards and other simple games of chance. Secondly, the classical approach may not explain actual results in certain cases. For example, if a coin is tossed 10 times we may get 6 heads and 4 tails. The probability of a head is thus 0.6 and that of a tail 0.4. However, if the experiment is carried out a large number of times we should expect approximately equal number of heads and tails. As n increases, i.e., approaches? (infinity), we find that the probability of getting a head or tail approaches 0.5. The probability of an

event can thus be defined as the relative frequency with which it occurs in an indefinitely large number of trials. If an event occurs a times out of n, its relative frequency is $\frac{a}{n}$;

the value which is approached by when n becomes infinity is called the limit of the relative frequency.

Symbolically, P (A) = $\log_{n\to\infty} \frac{a}{n}$

Theoretically, we can never obtain the probability of an event as given by the above limit. However, in practice we can only try to have a close estimate of P (A) based on a large number of observations. i.e., n.

In the relative frequency definition the fact that the probability is the value which is approached by when n becomes infinity, emphasizes a very important point i.e., probability involves a long-term concept. This means that if we toss a coin only 10 times, we may not get exactly 5 heads and 5 tails. However, as the experiment is carried out larger and larger number of times, say, coin is thrown 10,000 times we can expect heads and tails very close to 50 %. The two approaches, classical and empirical thought seemingly same. Differ widely in the former P (A) and a/n were practically equal when n was large whereas in the latter we say that P(A) is the limit a/n as n tends to infinity. In the second approach thus, the probability itself is the limit of the relative frequency as the number of observations increases indefinitely.

The relative frequency approach though useful in practice has difficulties from a mathematical point of view, since an actual limiting number may or may not really exist. Quite often people use this approach without evaluating a sufficient number of outcomes. For example, Mr. Kohli pointed out that his father and mother (aged 75 and 70) both had a serious heart problem in Jan-Feb 1999 and hence in winter people above 70 have a high probability of heart attack. His friends took it seriously and started giving special attention to their parents. On a deep thinking we may find that there is not enough evidence of establishing a relative frequency of occurrence probability. It may be observed that the empirical probability P(A) can never be obtained and one can only attempt at a close estimate of P(A) by making n sufficiently large.

For this reason, modern probability theory has been developed axiomatically in which probability is an undefined concept much the same as point and line are undefined in geometry. In this text however, we shall confine ourselves only to the first approach. It may be pointed out at the very outset that probability should not be understood in the sense of certainty. For example, when we say that the probability of getting head or tail when an unbiased coin is tossed is half, it does not mean that if the coin is tossed 16 times we must get 8 heads and 8 tails. What it means is that when an unbiased coin is tossed a large number of times and as n increases we will usually get close to 50 % heads and 50 % tails. The probability obtained by following relative frequency definition is call a posteriori or empirical probability as distinguished from a priori probably obtained by following the classical approach. A clear distinction between a priori and empirical probability is quit important for proper understanding of the concept of probability. First a priori probability is normally encountered problems in dealing with games of chance for example, dice and card. Normally, we think of a priori probability as being deductive in nature i.e. from cause to effect and based on theory instead of evidence of experience or experimentation. On the other hand, probability derived from past experience is called empirical probability and used in many practical problems. Priori probability may be determined rather easily whenever complete information is available on the various event that may occur. For example, the probability of drawing a king from a deck of 52 cards is 4/52 or 1/13. Similarly, the probability of getting 3 in a single throw of a dice is 1/6.

6.4.3 Subjective Approach

The subjective approach to assigning probabilities was introduced in the year 1926 by Frank Ramsey in his book, "The Foundation of Mathematics and other Logical Essays. The concept was further developed by Bernal Koopman, Richard Good and Leonard Savage. The subjective probability defined as the probability assigned to an event by an individual based on whatever evidence is available. Hence, such probabilities are based on the beliefs of person making the probability statement. For example, if a teacher wants to find out the probability of Mr. X topping in M.Com examination in Delhi University this year, he may assign a value between zero and one according to his degree of belief for possible occurrence. He may take into account

consideration factors such as the past academic performance, the views of his colleagues, the attendance record, performance in periodic tests etc. The application of personalistic concept to statistical problems has occurred virtually entirely in the post-World War II period, particularly in connection with statistical decision theory. This concept emphasizes the fact that since probability of an event is the degree of belief or degree of confidence placed in the occurrence of an event by a particular individual based on the evidence available to him, different individuals may differ in their degrees of confidence even when offered the same evidence.

This evidence may consist of relative frequency of occurrence data and any other quantitative or non quantitative information. Persons might arrive at different probability assignments because of differences in values, experience and attitudes etc. If an individual believe that it is unlikely that an event will occur, he will assign a probability close to zero to its occurrence. On the other hand, if he believes that it is very likely that the event will occur, he will assign a probability close to one. The personalistic approach is very broad and highly flexible. It permits probability assignment to events for which there may be no objective data, or for which there may be a combination of subjective and objective data. However, one has to be very careful and consistent in the assignment of these probabilities otherwise the decisions make may be misleading. Used with care the concept is extremely useful in the context of situations in business decision making.

6.4.4 Axiomatic Approach

The axiomatic approach to probability was introduced by the Russian mathematician, A.N. Kolmogorov in the year 1933. Kolmogorov axiomised the theory of probability in his book Foundations of Probability, published in 1933 introduces probability as a set function and is considered as a classic. When this approach is followed, no precise definition of probability is given, rather we give certain axioms or postulates on which probability calculations are based. The whole field of probability theory for finite sample spaces is based upon the following three axioms: The probability of an event ranges from zero to one. If the event cannot take place its probability shall be zero and if it is certain, i.e., bound to occur its probability shall be one.

The probability of the entire sample space is 1, i.e, P(S) =1.

If A and B are mutually exclusive (or disjoint) events then the probability of occurrence of either A or B denoted by P(A or B) shall be given by: P(A U B) = P(A)+P(B)

It may be pointed out that out of the four interpretations of the concept of probability, each has its own merits and one may use whichever approach is convenient and appropriate for the problem under consideration.

Experts disagree about which approach is the proper one to use.

6.5 SUMMARY

In this lesson we have become familiar with the concept of Probability, after study its various definitions, approaches, terms and importance in decision making process.

6.6 GLOSSARY

- **Probability:** The probability of a given event is an expression of likelihood or chance of occurrence of an event.
- **Experiment and Events:** Experiment refers to describe an act which can be repeated under some given conditions. Random experiments are those experiments whose results depend on chance such as tossing of a coin, throwing of dice. The results of a random experiment are called outcomes. If in an experiment all the possible outcomes are known in advance and none of the outcomes can be predicted with certainty, then such an experiment is called a random experiment and the outcomes as events or chance events.
- **Mutually Exclusive Events:** Two events are said to be mutually exclusive or incompatible when both cannot happen simultaneously in a single trial.
- **Independent and Dependent Events:** Two or more events are said to be independent when the outcome of one does not affect and is not affected by the other.
- **Equally Likely Events:** Events are said to be equally likely when one does not

occur more often than the others.

- **Simple and Compound Events:** In case of simple events we consider the probability of the happening or not happening of single events.
- **Exhaustive Events:** Events are said to be exhaustive when their totality includes all the possible outcomes of a random experiment.

6.7 SELF ASSESSMENT QUESTIONS

A. Fill in the blanks:

1. Two events are said to be equally likely if _____.
2. A set of events is said to be independent if _____.
3. The probability of getting a multiple of 2 in a throw of a dice is 1/2 and of getting a multiple of 3 is 1/3. Hence the probability of getting a multiple of 2 or 3 is_____.
4. Let A and B be independent events and suppose the event C has probability 0 or1.
5. Then A, B and C are _____ events.
6. If A, B, C are pair wise independent and A is independent of B and C, then A, B, C are_____ independent.

6.8 LESSON END EXERCISE

1. Define the concept of probability.

   _____
   _____
   _____

2. Give a collectively exhaustive list of the possible outcomes of tossing two dice.

   _____
   _____
   _____

3. Give the probability for each of the following totals in the rolling of two dice: 1, 2, 5, 6, 7, 10 and 11.

4. Explain the various approaches to probability. Are these approaches contradictory?

5. When are two events said to be independent in the probability sense? Give examples of dependent and independent events.

6. Explain the terms 'mutually exclusive' and 'equally likely'.

6.9 SUGGESTED READING

- Levin, R.I. Robin, D.S. *Statistics for Management,* Prentice-Hall of India. New Delhi.
- Aczel, A. D. Sounderpandian, J. *Complete Business Statistics,* Mc Graw Hill Publishing. New Delhi. downloaded
- Anderson, S., W. *Statistics for Business and Economics,* Cengage Learning. New Delhi.
- Kazmeir L. J. *Business Statistics,* Tata Mc Graw Hill. New Delhi.
- Vohra, N. D. *Business Statistics*. Tata Mc Graw Hill. New Delhi.

ADDITION THEOREM, MULTIPLICATION THEOREM AND MATHEMATICAL EXPECTATION

STRUCTURE

7.1 INTRODUCTION

In probability theory, the law (or formula) of total probability is a fundamental rule relating marginal probabilities to conditional probabilities. It expresses the total probability of an outcome which can be realised via several distinct events-hence the name. The term law of total probability is sometimes taken to mean the law of alternatives, which is a special case of the law of total probability applying to discrete random variables. One author uses the terminology of the "Rule of Average Conditional Probabilities", while another refers to it as the "continuous law of alternatives" in the continuous case. This result is given by Grimmett and Welsh as the partition theorem, a name that they also give to the related law of total expectation. In this lesson, we will be discussing about the basic probability theorems, viz., the addition theorem and the multiplication theorem as well as the concept and the application of mathematical expectations.

7.2 OBJECTIVES

After reading this lesson, you will be able to:

- Understand the addition theorem of probability.
- Explain the multiplication theorem in case of independent and dependent events.
- Know the concept of mathematical expectations.
- Apply all these theorems for solving many business problems.

7.3 ADDITION THEOREM

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

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

The set S of all possible outcomes (or elementary events) of a given experiment is called the sample space of the experiment.

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

P (AUB) read as "A union B" denotes the union of the events A and B. Knowledge of permutation and combination is extremely useful in calculating probabilities. For the sake of convenience in understanding the concept of permutation and combination an appendix is given at the end of the text.

Proof of the Theorem : If an event A can happen in a_1 ways and B in a_2 ways then the number of ways in which either event can happen is $a_1 + a_2$. If the total number of possibilities is n, then by definition the probability of either the first or the second event happening is:

$$\frac{a_1 + a_2}{n} = \frac{a_1}{n} + \frac{a_2}{n} \text{ but } \frac{a_1}{n} = P(A) \text{ and } \frac{a_2}{n} = P(B)$$

Hence P(A or B)=P(A)+P(B). (When events are mutually exclusive, i.e. they cannot occur simultaneously).

The theorem can be extended to three or more mutually exclusive events. Thus P(A or B or C)=P(A)+P(B)+P(C).

7.3.1 Mutually Exclusive Events

When we are interested to find out the probability that one thing or another will occur. If these two events are mutually exclusive, we can express this probability using the addition rule for mutually exclusive events. Symbolically, this rule is expressed as:

P (A or B) = probability of either A or B happening and is calculated as follows:

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

7.3.2 Events are not Mutually Exclusive

When two events are not mutually exclusive or in other words, it is possible for both events to occur, the addition rule must be modified. For example, what is the probability of drawing either a king or a heart from a standard pack of cards? It is obvious that the events king and heart can occur together as we can draw a king of hearts (since king and heart are not mutually exclusive events). We must deduce from the probability of drawing either a king or a heart, the chance that we can draw both of them together. Hence for finding the probability of one or more of two events that are not mutually exclusive we use

the modified form of addition theorem.

P(A or B)=P(A)+P(B)-P(A and B).

P(A or B)= Probability of A or B happening when A or B are not mutually exclusive.

P(A)=Probability of A happening

P(B)=Probability of B happening

P(AB)=Probability of A and B happening together.

In the example taken the probability of drawing a king or a heart shall be:

P(king or heart)=P(king)+P(heart)-P(king and heart)

$$= 4/52 + 13/52 - 1/52 = \frac{4}{13}$$

7.3.3 Practical Applications

Example 1: One card is drawn from a standard pack of 52. What is the probability that it is either a king or a queen?

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

the probability that the card drawn is a king = 4/52 and the probability that the card drawn is a queen = 4/52.

Since the events are mutually exclusive, the probability that the card drawn is either a king or a queen = $\frac{4}{52} + \frac{4}{52} = \frac{2}{13}$

Example 2: The Managing Committee of Vaishalli Welfare Association formed a subcommittee of 5 persons to look into electricity problem. Profiles of the 5 person are:
Male age 40

Male age 43

Female age 38

Female age 27

Male age 65

If a chairperson has to be selected from this, what is the probability that he would be either female or over 30 years?

Solution: P (female or over 30) = P(female)+P(over 30)-P(female and over 30)

$$= \frac{2}{5} + \frac{4}{5} - \frac{1}{5} = \frac{5}{5} = 1$$

Example 3. A Person is known to hit the target in 3 out of 4 shots, whereas another person is known to hit the target in 2 out of 3 shots. Find the probability of the target being hit at all when they both try.

Solution:

The probability that the first person hits the target = ¾

The events are not mutually exclusive because both of them may hit the target,

P(AB) = P(A).P(B), since A and B are independent events.

The require probability = $\left(\frac{3}{4} + \frac{2}{3}\right) - \left(\frac{3}{4} \times \frac{4}{5}\right) = \frac{17}{12} - \frac{6}{12} = \frac{11}{12}$

Here we have applied the theorem

P (A or B) = P(A)+P(B)-P(A and B).

7.4 MULTIPLICATION THEOREM

This theorem states that if two events A and B are independent, the probability that they both will occur is equal to the product of their individual probability. Symbolically, if A and B are independent, then P (A and B) = P (A) × P (B)

The theorem can be extended to three or more independent events. Thus P (A, B and C) = P (A) × P (B) × P (C)

Proof of the theorem: If an event A can happen in n_1 ways of which a_1 are successful and the event B can happen in n_2 ways of which a_2 are successful, we can combine each successful event in the first with each successful event in the second case. Thus, the total

number of successful happenings in both cases is $a_1 \times a_2$.

Similarly, the total number of possible cases is $n_1 \times n_2$. Then by definition of probability of the occurrence of both events is

$$\frac{a_1 + a_2}{n} = \frac{a_1}{n} + \frac{a_2}{n} = \text{but } \frac{a_1}{n} = P(A) \text{ and } \frac{a_2}{n} = P(B)$$

P (A and B) = P (A) × P (B).

In the similar way the theorem can be extended to three or more events.

7.4.1 Independent Events

Two events are said to be independent of each other if the occurrence or non-occurrence of one event in any trial does not affect the occurrence of the other event in any trial. Events A and B are independent of each other if and only if the following three conditions hold:

Conditions for the independence of two events A and B:

P (A /B) =P (A)
P (B /A) = P (B)
And P (A ∩ B) = P (A). P (B)

The first two equations have a clear, intuitive appeal. The top equation says that when A and B are independent of each other, then the probability of A stays the same even when we know that B has occurred - it is a simple way of saying that knowledge of B tells us nothing about A when the two events are independent. Similarly, when A and B are independent, then knowledge that A has occurred gives us absolutely no information about B and its likelihood of occurring.

The third equation, however, is the most useful in applications. It tells us that when A and B are independent (and only when they are independent), we can obtain the probability of the joint occurrence of A and B (i.e. the probability of their intersection) simply by multiplying the two separate probabilities. This rule is thus called the Product rule for Independent Events.

As an example of independent events, consider the following: Suppose I roll a single die.

What is the probability that the number 5 will turn up? The answer is 1/6. Now suppose that I told you that I just tossed a coin and it turned up heads. What is now the probability that the die will show the number 5? The answer is unchanged, 1/6, because events of the die and the coin are independent of each other. We see that P (6 / H) = P (6) , which is the first rule above.

The rules for intersection of two independent events can be extended to sequences of more than two events.

7.4.2 Dependent Events/Conditional Probability

The multiplication theorem explained above is not applicable in case of dependent events. Two events A and B are said to be dependent when B can occur only when A is

known to have occurred (or vice versa). The probability attached to such an event is called the conditional probability and is denoted by P(A/B) or, in other words, probability of A given that B has occurred.

If two events A and B are dependent, then the conditional probability of B given A is

$$P(B/A) = \frac{P(AB)}{P(A)} \text{ or } \frac{P(A\cap B)}{P(A)} \text{ or } \frac{P(A) + P(B) - P(A\cup B)}{P(A)}$$

Proof: Suppose a_1 is the number of cases for the simultaneous happening of A and B

a_1 out of a_1+a_2 cases in which A can happen with or without happening of B.

$P(B/A) = \frac{a1}{a1+a2} = \frac{a1/n}{(a1+a2)/n}$ = P (AB)/ P(A)

Similarly it can be shown that $P(A/B) = \frac{P(AB)}{P(A)}$

The general rule of multiplication in its modified form in terms of conditional probability becomes:

P(A and B)=P(B)×P(A/B)
or
P(A and B)=P(A)×P(B/A)
For three events A, B and C, we have

P(ABC)=P(A)×P(B/A)×P(C/AB)

i.e., the probability of occurrence of A, B and C is equal to the probability of A, times the probability of B given that A has occurred, times the probability of given that both A and B have occurred.

As a measure of uncertainty, probability depends on information. We often face situations where the probability of an event A is influenced by the information that another event B has occurred. Thus, the probability we would give the event "Xerox stock price will go up tomorrow" depends on what we know about the company and its performance; the probability is conditional upon our information set. If we know much about the company, we may assign a different probability to the event than if we know little about the company. We may define the probability of event A conditional upon the occurrence of event B. In this example, event A may be the event that the stock will go up tomorrow, and event B may be a favourable quarterly report.

7.4.3 Practical Applications

Example 4. A bag contains 5 white and 3 black balls. Two balls are drawn at random one after the other without replacement. Find the probability that both balls drawn are black**.**

Solution:

Probability of drawing a black ball in the first attempt is P(A) =

Probability of drawing the second black ball given that the first ball drawn is black P(AB) = 3/8

the probability that both balls drawn are black is given by

P(AB) = P(A) × P $(\frac{A}{B})$= $\frac{3}{8} \times \frac{2}{7}$ = 6/56

Example 5. Find the probability of drawing a queen, a king and a knave in that order from a pack of cards in three consecutive draws, the cards drawn not being replaced**.**

Solution.

Probability of drawing a queen = 4/52

Probability of drawing a king after a queen has been drawn = 4/50

Probability of drawing a king after a queen have been drawn =

Is $\frac{4}{52} \times \frac{4}{51} \times \frac{4}{50} = \frac{\mathbf{64}}{\mathbf{132600}} = 0.00048$

Since they are dependent events, the required probability of the compound event

Is = 0.00048

Example 6. A man wants to marry a girl having qualities: white complexion-the probability of getting such a girl is one in twenty; handsome dowry-the probability of getting this is one in fifty; westernized manners and etiquettes-the probability here is one in hundred. Find out the probability of his getting married to such a girl when 'the possession of these attributes is independent'.

Solution:

Probability of a girl with white complexion=1/20=0.05

Probability of a girl with handsome dowry=1/50=0.02

Probability of a girl with westernized manners=1/100=0.01

Since the events are independent, the probability of simultaneous occurrence

qualities = 0.05× 0.02×0.01 = 0.00001.

If we are given n independent events A_1, A_2, A3,…., An with respective

occurrence as P_1, P_2, P_3,….,Pn, then the probability of occurrence of at least one A_1, A_2, A3,…., An can be determined as follows:

P(happening of at least one of the events) = 1-P(happening of none of the events).

The following example shall illustrate the application of the above principle.

Example 7: A problem in statistics is given to five students A, B, C, D and E. Their

chances of solving it are 1/2, 1/3, 1/4, 1/5, and 1/6. What is the probability that the problem will be solved?

Solution:

Probability that A fails to solve the problem is 1- 1/2 = 1/2.

Probability that B fails to solve the problem is 1- 1/3 = 2/3

Probability that C fails to solve the problem is 1- 1/4 = 3/4

Probability that D fails to solve the problem is 1- 1/5 = 4/5

Probability that E fails to solve the problem is 1- 1/6 = 5/6

Since the events are independent, the probability that all the five students fail to solve the problem is $\frac{1}{2} \times \frac{2}{3} \times \frac{3}{4} \times \frac{4}{5} \times \frac{5}{6} = \frac{1}{6}$

The problem will be solved if anyone of them is able to solve it.

the probability that the problem will be solved $1-.\frac{1}{6} = \frac{5}{6}$

7.5 MATHEMATICAL EXPECTATION

Mathematical expectation, also known as the expected value**,** which is the summation of all possible values from a random variable. It is also known as the product of the probability of an event occurring, denoted by P(x), and the value corresponding with the actually observed occurrence of the event. For a random variable expected value is a useful property. E(X) is the expected value and can be computed by the summation of the overall distinct values that is the random variable. The mathematical expectation is denoted by the formula:

$$E(X) = \Sigma\ (x_1p_1, x_2p_2, \ldots, x_np_n)$$

where, x is a random variable with the probability function, *f*(x), p is the probability of the occurrence, and n is the number of all possible values.

The mathematical expectation of an indicator variable can be 0 if there is no occurrence of an event A, and the mathematical expectation of an indicator variable can be 1 if there is an occurrence of an event A. For example, a dice is thrown, the set of possible outcomes is {1, 2, 3, 4, 5, 6} and each of this outcome has the same probability 1/6. Thus, the expected value of the experiment will be 1/6*(1+2+3+4+5+6) = 21/6 = 3.5. It is important to

know that "expected value" is not the same as "most probable value" and, it is not necessary that it will be one of the probable values.

Properties of Expectation

1. If X and Y are the two variables, then the mathematical expectation of the sum of the two variables is equal to the sum of the mathematical expectation of X and the mathematical expectation of Y. Or $E(X+Y)=E(X)+E(Y)$

2. The mathematical expectation of the product of the two random variables will be the product of the mathematical expectation of those two variables, but the condition is that the two variables are independent in nature. In other words, the mathematical expectation of the product of the *n* number of independent random variables is equal to the product of the mathematical expectation of the *n* independent random variables Or $E(XY)=E(X)E(Y)$

3. The mathematical expectation of the sum of a constant and the function of a random variable is equal to the sum of the constant and the mathematical expectation of the function of that random variable. Or, $E(a+f(X))=a+E(f(X))$, where, a is a constant and $f(X)$ is the function.

4. The mathematical expectation of the sum of product between a constant and function of a random variable and the other constant is equal to the sum of the product of the constant and the mathematical expectation of the function of that random variable and the other constant. Or, $E(aX+b) = aE(X)+b$, where, a and b are constants.

5. The mathematical expectation of a linear combination of the random variables and constant is equal to the sum of the product of 'n' constant and the mathematical expectation of the 'n' number of variables. Or $E(\Sigma a_i X_i)=\Sigma\ a_i E(X_i)$ Where, a_i, $(i=1\ldots n)$ are constants.

Example 8: What is the expected number of coin flips for getting two consecutive heads?

Solution: Let the expected number of coin flips be x. If the first flip is a tail then the probability of the event is 1/2. Thus, the total number of flips required is x+1. If the first flip is a head and the second flip is a tail, then the probability of the event is 1/4 and the total number of flips we require is x+2. If the first flip is a head and the second flip is also heads,

then the probability of the event is 1/4 and the total number of flips we require is 2.

By Adding, the equations we get

$x = (1/2)(x+1) + (1/4)(x+2) + (1/4)2$

By solving the equation, we get $x = 6$.

So, the expected number of coin flips for getting two consecutive heads is 6.

7.6 SUMMARY

In this lesson we have discussed about theorems/rules of probability, viz., the multiplication rule and the addition rule are used for computing the probability of A and B, as well as the probability of A or B for two given events A, B defined on the sample space. In sampling with replacement each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered to be not independent. The events A and B are mutually exclusive events when they do not have any outcomes in common.

7.7 GLOSSARY

- **Addition Theorem:** If A and B are any two events then the probability of happening of at least one of the events is defined as P (AUB) = P (A) + P (B).

- **Multiplication Theorem:** The probability of simultaneous occurrence of two events A and B is equal to the product of the probability of the other, given that the first one has occurred. This is called the Multiplication Theorem of probability.

- **Theorem:** a formula, proposition, or statement in mathematics or logic deduced or to be deduced from other formulas or propositions.

- **Mathematical Expectation:** Mathematical expectation, also known as the expected value, is the summation or integration of a possible values from a random variable. It is also known as the product of the probability of an event occurring, denoted P(x), and the value corresponding with the actual observed occurrence of the event.

7.8 SELF ASSESSMENT QUESTIONS

A. Multiple Choice Questions:

1. The probability of drawing any one spade card from a pack of cards is:

   a. $\frac{1}{52}$ b. $\frac{1}{13}$ c. $\frac{4}{13}$ d. $\frac{1}{4}$

2. Which of these represents the multiplication theorem of probability?

   a. $P(A \cap B) = P(B)\, P(B/A)$ b. $P(A \cap B) = P(A)\, P(B/A)$
   c. $P(A \cap B) = P(A)\, P(B/B)$ d. $P(A \cap B) = P(A)\, P(A/A)$

3. A bag contains 9 identical balls, of which are 4 are blue and 6 are green. Three balls are taken out randomly from the bag after one another. Find the probability that all three balls are blue?

   a. 5/8 b. 6/19
   c. 5/21 d. 4/7

7.9 LESSON END EXERCISE

1. Explain addition theorem of exhaustive events.

2. Differentiate between mutually exclusive events and exhaustive events.

3. Discuss modified version of both addition as well as multiplication theorem.

4. A proof reader is interested in finding the probability that the number of mistakes in a page will be less than 10. From his past experience he finds that out of 3600 pages he has proofed, 200 pages contained no errors, 1200 pages contained 5 errors, and 2200 pages contained 11 or more errors. Can you help him in finding the required probability?

5. Calculate the probability of drawing an ace form a deck of 52 cards.

6. What is the probability that a leap year selected at random will contain either 53 Thursdays or 53 Fridays?

7. The probability that A can solve a problem is 0.7 and the probability that B can solve that problem is 0.6. Considering that these two events are independent, find the probability that the problem gets solved by either of them.

7.10 SUGGESTED READING

- Levin, R.I. Robin, D.S. *Statistics for Management*, Prentice-Hall of India. New Delhi.
- Aczel, A. D. Sounderpandian, J. *Complete Business Statistics*, Mc Graw Hill Publishing. New Delhi. downloaded

- Anderson, S., W. *Statistics for Business and Economics*, Cengage Learning. New Delhi.
- Kazmeir L. J. *Business Statistics*, Tata Mc Graw Hill. New Delhi.
- Vohra, N. D. *Business Statistics*. Tata Mc Graw Hill. New Delhi.

NORMAL DISTRIBUTION

STRUCTURE

8.1 INTRODUCTION

The probability distribution of a random variable may be, theoretical listing of outcomes

and Probabilities, an empirical listing of outcomes and there observed relative frequencies, a subjective listing of outcomes associated, with their subjective or contrived probabilities representing the degree of and 'Conviction of the decision making as to the likelihood of the possible outcomes.

The observed frequency distributions are based on observation and experimentation and are obtained by grouping of data. They help in understanding properly the nature of data. As different from this type of distribution which is based on actual observation, it is possible to derive mathematically what the frequency distributions of certain populations should be. Such distributions as are expected on the basis of previous experience or theoretical considerations are known as, 'Theoretical frequency distributions' or 'Probability distributions'.

Among theoretical or expected frequency distributions, the binominal, poisson, and normal find much more wider application in practice than other. Hence we shall discuss these three in our subsequent lessons in detail.

8.2 OBJECTIVES

After reading this lesson, you will be able to:

- Identify the situations where normal distribution is applied.
- Appreciate the relevance of normal distribution theory in decision making.

8.3 CONCEPT OF NORMAL DISTRIBUTION

In order to have mathematical distribution suitable for dealing with quantities whose magnitude is continuously variable, a continuous distribution is needed. The normal distribution, also called the normal probability distribution, happens to be most useful theoretical distribution for continuous variables. Many statistical data concerning business and economic problems are displayed in the form of normal distribution. In fact normal distribution is the cornerstone of the modern statistics. The normal distribution was first described by Abaraham Demovre (1667-1754) as the limiting form of the binomial model in 1733.

Normal distribution was re-discovered by Gauss in 1809 and by Laplace in 1812. Both Gauss and Laplace were led to the distribution by their work on the theory of errors of

observations arising in physical measuring processes particularly in astronomy. Throughout the 18th and 19th centuries, various efforts were made to establish the normal model as the underlying law ruling all continuous random variables-thus the name normal. These efforts failed because of the false premises. The normal model has, nevertheless, become the most important probability model in statistical analysis. The normal distribution is an approximation to binomial distribution. Whether or not p is equal to q, the binomial distribution tends to the form of the continuous curve and when n becomes large at least for the material part of the range. As a matter of fact, the correspondence between the binomial and the curve is surprisingly close even for comparatively low values of n, provided that p and q are fairly near equality. The limiting frequency curve obtained as n becomes large is called the normal frequency curve or simply the normal curve.

The normal curve is represented in several forms. The following is the basic form relating to the curve with mean μ and variance σ 2.

The Normal Distribution

$$P(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$X = Values\ of\ the\ continuous\ random\ variable$

$\mu = Mean\ of\ the\ normal\ random\ variable$

$e = Mathematical\ constant\ approximated\ by\ 2.7183$

$\pi = Mathematical\ constant\ approximated\ by\ 3.1416$

$(\sqrt{2\pi} = 2.5066)$

When we Say that the curve has unit area, we mean that the total frequency N is equated to i for convenience in representation and calculation. To obtain ordinates for a particular distribution the ordinates given by the above formula are multiplied by N. The equation to a normal curve corresponding to a particular distribution is thus given by:

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

The graph of the normal distribution is characterized by two parameters: the mean, or average, which is the maximum of the graph and about which the graph is always symmetric; and the standard deviation, which determines the amount of dispersion away from the mean.

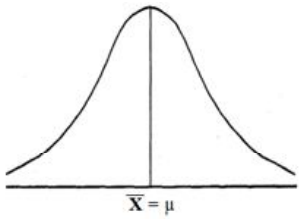


Figure 1: Graph of the Normal distribution

- The normal distribution can have different shapes depending on different values of μ and σ but there is one and only one normal distribution for any given pair of values of μ and σ .
- Normal distribution is a limiting case of binomial distribution when i) $n \to \infty$ and ii) neither p nor q is very small.
- Normal distribution is a limiting case of poisson distribution when its mean 'm' is large.
- The mean of a normally distributed population lies at the centre of its normal curve.
- The two tails of the normal probability distribution extend indefinitely and never touch the horizontal axis (which implies a positive probability for finding values of the random variable within any range from $-\infty$ *to* $+\infty$.

8.4 IMPORTANCE OF NORMAL DISTRIBUTION

The normal distribution has been long occupied a central place in the theory of statistics. Its importance will be clear from the following points:

1. The normal distribution has the remarkable property stated in the so called Central Limit Theorem. According to this theorem as the sample size n increases the distribution of mean, $\overline{X}$ of a random sample taken from practically any population approaches a normal distribution (with mean μ and standard deviation $\sigma/\sqrt{n}$). Thus, if samples of large size are drawn from a population that is not normally distributed nevertheless the successive sample means will form themselves a distribution that is approximately normal. As the size of the sample increased the sample means will tend to be normally

distributed. The central limit theorem applies to the distribution of most other statistics as well, such as the median and standard deviation (but not range). The central limit theorem gives the normal distribution its central place in the theory of sampling, since many important problems can be solved by the single pattern of sampling variability. As a result the work on statistical inferences is made simpler. This characteristic makes it possible to determine the minimum and maximum limits within which the population values lie. For example, within range of population mean $\pm 3\sigma$, 99.73% or almost all the items are covered.

2. As 'n' becomes large the normal distribution serves as a good approximation of many discrete distributions (such as the Binomial or the Poisson model) whenever the exact discrete probability is laborious to obtain or impossible to calculate accurately.

3. In theoretical statistics many problems can be solved only under the assumption of a normal population. In applied work as well as we often find that methods developed under the normal probability law yield satisfactory results even when the assumption of a normal population is not fully met, despite the fact that the problem can have a formal solution only if such a premise is hypothesised.

4. The normal distribution has numerous mathematical properties which make it popular and comparatively easy to manipulate. For example, the moments of the normal distribution are expressed in a simple form. The normal curve is reasonably close to many distributions of the humped type. If, therefore we are ignorant of the exact nature of a humped distribution of know the form but find it mathematically intractable. We may assume as a first approximation that the distribution is normal and see where this assumption leads us to.

5. The normal distribution is used extensively in statistical quality control in industry in setting up of control limits.

8.4.1 Properties of Normal Distribution

The following are the important properties of the normal curve and the normal distribution.

1. The normal curve is "bell-shaped" and symmetrical in its appearance. If the curves were folded along its vertical axis the two lines halves would coincide. The number of

cases below the mean in a normal distribution is equal to the number of cases above the mean which make the mean and median coincide. The height of the curve for a positive deviation of 3 units is the same as the height of the curve for negative deviation of 3 units.

2. The height of the normal curve is at its maximum at the mean. Hence the mean and the mode of the normal distribution coincide. Thus for a normal distribution mean, median and mode are all equal.

3. There is one maximum point of the normal curve which occurs at the mean. The height of the curve approaches nearer and nearer to the base but it never touches it, i.e., the curve is asymptotic to the base on either side. Hence its range is unlimited or infinite in both directions.

4. Since there is only one maximum point, the normal curve is unimodal. i.e., it has only one mode.

5. The ponits of inflexion, i.e., the points where the change incurvature occurs are X± .

6. As distinguished from Binomial and Poisson distributions where the variable is discrete, the variable distributed according to the normal curve is a continuous one.

7. The first and the third quartiles are equidistant from the median.

8. The mean deviation is 4th or more precisely 0.7979 of the standard deviation.

9. The area under σr the normal curve distributed as follows:

Mean ±1 covers 68.27% area; 34.135% area will lie on either side of the mean.σ

Mean ±2σ covers 95.45% area.

Mean ±3 covers 99.73% area.

8.4.2 Constants of Normal Distribution

The two main parameters (constants) of a (normal) distribution are the mean and standard deviation. The parameters determine the shape and probabilities of the distribution. The shape of the distribution changes as the parameter values change.

1. Mean (μ)

The mean is used by researchers as a measure of central tendency. It can be used to describe the distribution of variables measured as ratios or intervals. In a normal distribution graph, the mean defines the location of the peak, and most of the data points are clustered around the mean. Any changes made to the value of the mean move the curve either to the left or right along the X-axis.

2. Standard Deviation (σ)

The standard deviation measures the dispersion of the data points relative to the mean. It determines how far away from the mean the data points are positioned and represents the distance between the mean and the observations.

On the graph, the standard deviation determines the width of the curve, and it tightens or expands the width of the distribution along the x-axis. Typically, a small standard deviation relative to the mean produces a steep curve, while a large standard deviation relative to the mean produces a flatter curve.

8.4.3 Conditions for Normality

The following four conditions must prevail among the factors affecting the individual events that make up a given population, if the distribution of observations is to be normal.

1. The causal forces must be numerous and of approximately equal weight.
2. These forces must be the same over the universe from which the observations are drawn (although their incidence will vary from event to event). This is the condition of homogeneity.
3. The forces affecting events must be independent of one another.
4. The operation of the causal forces must be such that deviation above the population mean is balanced as to magnitude and number of deviations below the mean. This is the condition of symmetry.

8.4.4 Significance of Normal Distribution

The normal distribution is mostly used for the following purposes:

1. To approximate of 'fit' a distribution of measurement under certain conditions.
2. To approximate the binomial distribution and other discrete or continuous probability distributions under suitable conditions.
3. To approximate the distribution of means and certain other quantities calculated from samples, especially large samples.

8.5 RELATIONSHIP BETWEEN NORMAL, BINOMIAL AND POISSON DISTRIBUTION

The three distributions, namely, Binomial, Poisson and Normal, are very closely related to each other. As explained earlier when N is large while the probability P of the occurrence of an event is close to zero so that q = (1-p) the binomial distribution is very closely approximated by the Poisson distribution with m = np. Since there is a relation between the binomial and normal distributions, it follows that there is also relation between the Poisson and normal distributions.

8.6 SUMMARY

The normal distribution described above is the most useful theoretical distribution for continuous variables. In order to have mathematical distribution suitable for dealing with quantities whose magnitude is continuously variable, a continuous distribution is needed. The normal distribution, also called the normal probability distribution, happens to be most useful theoretical distribution for continuous variables. Many statistical data concerning business and economic problems are displayed in the form of normal distribution. In fact normal Distribution is the cornerstone of the modern statistics.

8.7 GLOSSARY

- **Normal Distribution:-** It is a function that represents the distribution of many random variables as symmetrical bell shaped curve.
- **Normal Deviate:** A Random variable with a Gaussian distribution is said to be normally distributed and is called a normal deviate.
- **Bell Curve:** The normal Distribution is sometimes informally called the bell-curve.

8.8 SELF ASSESSMENT QUESTIONS

A. Fill in the blanks:

1. The normal distribution is an approximation to_____.
2. In case of normal distribution X± 2 covers_____of the distribution.
3. In a normal distribution, the points of inflexian occur at_____
4. A normal curve is completely defined by the _____

8.9 LESSON END EXERCISE

1. What is normal probability distribution?

_____
_____
_____

2. Explain the characteristics features of a normal distribution.

_____
_____
_____

3. What are the properties of a normal curve?

_____
_____
_____

4. Explain the circumstances when the normal probability distribution is used in business?

_____
_____
_____

8.10 SUGGESTED READING

- Levin, R.I. Robin, D.S. *Statistics for Management*, Prentice-Hall of India. New Delhi.

- Aczel, A. D. Sounderpandian, J. *Complete Business Statistics*, Mc Graw Hill Publishing. New Delhi. downloaded
- Anderson, S., W. *Statistics for Business and Economics*, Cengage Learning. New Delhi.
- Kazmeir L. J. *Business Statistics*, Tata Mc Graw Hill. New Delhi.
- Vohra, N. D. *Business Statistics*. Tata Mc Graw Hill. New Delhi.

BINOMIAL AND POISSON DISTRIBUTION

STRUCTURE

9.9 Suggested Reading

9.1 INTRODUCTION

In the real world, we rarely come across experiments with single outcomes like heads or tails. Generally, we do the experiment as a set of events and carry it for n number of times which give us a collection of outcomes which we can represent in the form of theoretical distribution. By theoretical distribution, we take mean of a frequency distribution, which we obtain in relation to a random variable by some mathematical model. A random exponent is assumed as a model for theoretical distribution, and the probabilities are given by a function of the random variable is called probability function. For example, if we toss a fair coin, the probability of getting a head is 12. If we toss it for 50 times, the probability of getting a head is 25. We call this as the theoretical or expected frequency of the heads. But actually, by tossing a coin, we may get 25, 30 or 35 heads which we call as the observed frequency.

Thus, the observed frequency and the expected frequency may equal or may differ from each other due to fluctuation in the experiment. Generally, there are three types of theoretical distributions, viz; binomial distribution, poisson distribution and normal distribution. We have already discussed Normal distribution in lesson 8 which is applicable in case of continuous variables. But when the variable follows discrete distribution, we generally apply binomial and poisson distributions. Therefore, binomial and poisson distributions are also known as discrete probability distributions. Thus, in this lesson we are going to discuss about the binomial and poisson distributions.

9.2 OBJECTIVES

After reading this lesson, you will be able to:

- Understand the concept of Binomial Distribution.
- Identify situations where Binomial Distributions can be applied.
- Appreciate the relevance of Binomial Distribution.
- Appreciate the relevance of Poisson Distribution theory in decision making.

- Identify the situations where Poisson Distribution can be applied.

9.3 BINOMIAL DISTRIBUTION

The binomial distribution also known as 'Bernoulli Distribution' is associated with the name of a Swiss mathematician James Bernoulli also known as Jacques or Jacob (1654-1705). Binomial distribution is a probability distribution expressing the probability of one set of dichotomous alternatives i.e., success or failure. This distribution has been used to describe a wide variety of processes in business and the social sciences as well as other areas. The type of process which gives rise to this distribution is usually referred to as Bernoulli trial or as a Bernoulli process.

9.3.1 Meaning of Binomial Distribution

The prefix 'Bi' means two or twice. A binomial distribution can be understood as the probability of a trail with two and only two outcomes. It is a type of distribution that has two different outcomes namely, 'success' and 'failure'. Also, it is applicable to discrete random variables only. Thus, the binomial distribution summarised the number of trials, survey or experiment conducted. It is very useful when each outcome has an equal chance of attaining a particular value. The binomial distribution has some assumptions which show that there is only one outcome and this outcome has an equal chance of occurrence.

The three different criteria of binomial distributions are:

1. The number of the trial or the experiment must be fixed.

2. Every trial is independent. None of your trials should affect the possibility of the next trial.

3. The probability always stays the same and equal. The probability of success may be equal for more than one trial.

9.3.2 Relevance of Binomial Distribution

The binomial probability distribution is a discrete probability distribution that is useful in describing an enormous variety of real life events. For example, a quality control inspector wants to know the probability of defective light bulbs in a random sample of 10 bulbs if

10% of the bulbs are defective. He can quickly obtain the answer from tables of the binomial probability distribution.

The binomial distribution can be used when:

1. The outcome or results of each trial in the process are characterized as one of two types of possible outcomes. In other words, they are attributes.

2. The possibility of outcome of any trial does not change and is independent of the results of previous trials.

9.3.3 Properties of Binomial Distribution

1. The shape and location of binomial distribution changes a p changes for a given n or as n changes for a given p. As p increases for a fixed n, the binomial distribution shifts to the right.

2. The mode of the binomial distribution is equal to the value of x which has the largest probability. For example, if n=6 and p=0.3, the mode is equal to While for n=6 and p=0.9 the mode is equal to 6. The mean and mode are equal if np is an integer. For example, when n=6 and p=0.50, the mean and mode are both equal to 3. For fixed n, both mean and mode increase as p increases.

3. As n increases for a fixed p, the binomial distribution moves to the right, flattens and spreads out. The mean of the binomial distribution, np, obviously increases as n increases with p held constant. For larger n there are more possible outcomes of a binomial experiment and the probability associated with any particular outcome becomes smaller.

4. If n is large and if neither p nor q is too close to zero, the binomial distribution can be closely approximated by a normal distribution with standardized variable = $\frac{X-np}{\sqrt{npq}}$ given by the approximation becomes better with increasing n.

9.3.4 Constants of Binomial Distribution

The mean of the binomial distribution is np and the standard deviation npq

Proof: If p is the probability of success and q the probability of failure in one trial then in n independent trials the probabilities of 0,1,2,3,4,….,n successes are given by the 1st, 2nd,

3rd,….,$(n+1)^{th}$ term of the binomial expansion $(q+p)^n$. Thus we have:

X	P(x)	xxP(x)
0	q^n	$0 \times q^n$
1	$\binom{n}{1} q^{n-1}p$	$1 \times nq^{n-1}p$
2	$\binom{n}{2} q^{n-2}p^2$	$\frac{2n(n-1)}{2} q^{n-2}p^2$
-	-	-
-	-	-
-	-	-
n	p^n	$n \times pn$
	$\sum p(x) = 1$	

The arithmetic mean by definition $= \frac{\sum x\, p(x)}{\sum p(x)}$

$$\sum x.\, p(x) = 0.q^n = nq^{n-1}p + \frac{2n(n-1)}{2} q^{n-2}p^2 + \ldots + npn$$

$$= nq^{n-1}p + n(n-1)\, q^{n-2}p^2 + \ldots + np^n$$

Taking 'np' common

$$= np[q^{n-1} + n(n-1)\, q^{n-2}p + \ldots + p^{n-1}]$$

$np(q + p)^{n-1}$ [since the expansion in brackets is the expansion of the bionimal $(q + p)^{n-1}$]

$= np(1)^{n-1} = np$ since $q + p = 1$

Thus, $\sum x.\, p(x) = np$

Thus, the mean of binomial distribution is np.

The standard deviation of binomlal distribution in $\sqrt{npq}$

Constants of Binomial Distribution

Mean $= np$

Standard Deviation $= \sqrt{npq}$

First Moment $= \mu_1 = 0$

Second Moment $= \mu_2 = npq$

Third Moment $= \mu_4 = 3n^2p^2q^2 + npq(1-6pq)$

$$= \beta_1 = \frac{(p-q)^2}{npq}$$

$$= \beta_1 = 3 + \frac{(1-6pq)}{npq}$$

9.3.5 Applications

The following examples will illustrate the applications of binomial distribution.

Example 1: A coin is tossed six times. What is the probability of obtaining four or more heads?

Solution: When a coin is tossed the probabilities of head and tail in case of an unbiased coin are equal i.e., p = q = 1/2.

The various possibilities for all the events are the terms of the expansion $(q+p)^6$ are:

$$(q + p)^6 = q^6 + 6q^5p + 15q^4p^2 + 20q^3p^3 + 15q^2p^4 + 6pq^5 + p^6$$

The probability of obtaining 4 heads is

$$= 15q^2p^4 = 15x\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^4 = 0.234$$

The probability of obtaining 5 heads is

The probability of obtaining 6 heads is

$$= p^6 = \left(\frac{1}{2}\right)^6 = 0.016$$

$\therefore$ the probability of obtaining 4 or more heads

$= 0.234 + 0.094 + 0.016 = 0.344.$

Example 2: The basketball team is playing a series of 5 games against opponent. The winner is those who wins more games (out of 5). Let us assume that the basketball team is much more skilled and has 75% chances of winning. It means there is a 25% chance of losing. What is the probability of your team get 3 wins?

Solution:

In this example:

n = 5, p=0.75, q=0.25, x=3

Let's replace in the formula to get the answer:

$$P(x = 3) = \frac{5!}{3!\,(5-3)!}\ 0.75^3\ (1 - 0.75)^{5-3} = 0.264$$

Therefore, the probability that the team win 3 games is 0.264.

Example 3: A box of candies has many different colours in it. There is a 15% chance of getting a pink candy. What is the probability that exactly 4 candies in a box are pink out of 10?

Solution:

We have that:

n = 10, p=0.15, q=0.85, x=4

When we replace in the formula:

$$P(x = 4) = \frac{10!}{4!\,(10-4)!}\ 0.15^{4}\,(1 - 0.15)^{10-4} = 0.04$$

9.3 POISSON DISTRIBUTION

Poisson distribution is a discrete probability distribution and is very widely used in statistical work. It was developed by a French mathematician, Simeon Denis Poisson (1781-1840). Poisson distribution may be expected in cases where the chance of any individual event being a success is small. The distribution is used to describe the behaviour of rare events such as the number of accidents on road, number of printing mistakes in a book, etc., and has been called "the law of improbable events".

9.4.1 Meaning

The Poisson distribution is a theoretical discrete probability distribution that is very useful in situations where the events occur in a continuous manner. Poisson distribution is utilized to determine the probability of exactly x_0 number of successes taking place in unit time. Let us now discuss the Poisson Model.

At first, we divide the time into n number of small intervals, such that $n \rightarrow \infty$ and p denote the probability of success, as we have already divided the time into infinitely small intervals so $p \rightarrow 0$. So the result must be that in that condition is $n \times p = \lambda$ *(a finite constant)*. In recent years the statisticians have had a renewed interest in the occurrence of comparatively rare events, such as serious floods, accidental release of radiation from a nuclear reactor, and the like. The Poisson Distribution is defined as :

$$P(r) = \frac{e^{-m} m^{r}}{r} \quad \textit{where } r = 0,1,2,3,4,.....$$

e=2.7183 (the base of natural Logarithms)

m=the mean of the Poisson distribution, i.e, np or the average number of occurrences of an event

The Poisson Distribution is a discrete distribution with a single parameter m. As m increases,

the distribution shifts to the right. All poisson probability distributions are skewed to the right. This is the reason why poisson probability distribution has been called the probability distribution of rare events (the probabilities tend to the high for small numbers of occurrences). The Poisson probability distribution is concerned with certain processes that can be described by a discrete random variable. The probabilities of 0, 1, 2,….. successes are given by the successive terms of the expansion.

This can be written in a tabular form as follows:

Table 1

No. of successes (x)	Probability P(x)	No. of successes (x)	Probability P(x)
0	e^{-m}	4	$\frac{m^4 e^{-m}}{4!}$
1	me^{-m}	-	-
2	$\frac{m^2 e^{-m}}{2!}$	r	$\frac{m^r e^{-m}}{r!}$
3	$\frac{m^3 e^{-m}}{3!}$	-	-

The above table gives probabilities. If we want to know the expected number of occurrences for different successes, we have to multiply each term by N, i.e., the total number of observations.

9.4.2 Constants of Poisson Distribution

Since p is very small in case of poisson distribution, the value of q is almost equal to 1. The constants of the Poisson distribution can thus be easily obtained by putting 1 in place of q in the constants of binomial distribution. The mean of the Poisson distribution =m, and

the standard deviation is $\sqrt{m} = or\ \mu 2$ = m

Proof : The Poisson distribution is given as

Table 2

No. of successes (x)	0	1	2	3	4	---------
Probability P(x)	e^{-m}	me^{-m}	$\frac{m^2e^{-m}}{2!}$	$\frac{m^3e^{-m}}{3!}$	$\frac{m^4e^{-m}}{4!}$	----------

Find the mean and variance.

Calculation of Mean and Variance

Table 3

No. of successes (x)	Probability P(x)	No. of successes x.P(x)
0	e^{-m}	0
1	me^{-m}	me^{-m}
2	$\frac{m^2e^{-m}}{2!}$	$2 \times \frac{m^2e^{-m}}{2!}$
3	$\frac{m^3e^{-m}}{3!}$	$\frac{3 \times m^3e^{-m}}{3!}$
4	$\frac{m^4e^{-m}}{4!}$	$4 \times \frac{m^4e^{-m}}{4!}$
-	-	-
r	$\frac{m^re^{-m}}{r!}$	$r \times \frac{m^re^{-m}}{r!}$
-	-	-

The mean of poisson distribution = m, and

$$\text{Mean} = \sum x.p(x) = 0 + me^{-m} + m^2e^{-m} + \frac{3 \times m^3e^{-m}}{3!} + 4 \times \frac{m^4e^{-m}}{4!} + r \times \frac{m^re^{-m}}{r!} + \cdots$$

$$= me^{-m}\left(1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + \cdots..\right)$$

$$= me^{-m}.e^m = m \qquad \text{since } e^{-m} = \left(1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + \cdots..\right)$$

Hence the mean of the Poisson distribution is m.

Variance = $\mu 2$ = $v_2 - v_1^2$ (Where v1 and v2denote moment about origin.

$$v_2 = \sum \{x^2.p(x)\}$$

Table 4

No. of successes (x)	Probability P(x)	No. of successes x^2.P(x)
0	e^{-m}	0
1	me^{-m}	me^{-m}
2	$\frac{m^2e^{-m}}{2!}$	$4 \times \frac{m^2e^{-m}}{2!}$
3	$\frac{m^3e^{-m}}{3!}$	$\frac{9 \times m^3e^{-m}}{3!}$
4	$\frac{m^4e^{-m}}{4!}$	$16 \times \frac{m^4e^{-m}}{4!}$
-	-	-
r	$\frac{m^re^{-m}}{r!}$	$r^2 \times \frac{m^re^{-m}}{r!}$
-	-	-

$$\sum x^2.p(x) = 0 + me^{-m} + 2m^2e^{-m} + \frac{9 \times m^3e^{-m}}{2!} + 16 \times \frac{m^4e^{-m}}{3!} + \cdots$$

$$= me^{-m}\left(1 + 2m + 3\frac{m^2}{2!} + 4\frac{m^3}{3!} + \cdots..\right)$$

Breaking each term within the brackets into two parts each, we have

$$\sum x^2.p(x) = me^{-m}\left\{\left(1 + \frac{m^2}{2!} + \frac{m^3}{3!} + \cdots..\right) + \left(m + 2\frac{m^2}{2!} + 3\frac{m^3}{3!} + \cdots..\right)\right\}$$

$$= me^{-m}\left\{e^m + m\left(1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + \cdots..\right)\right\}$$

$$= me^{-m}(e^m + me^m) = me^{-m}.e^m(1 + m)$$

$$= m(1 + m)$$

$$= m + m^2$$

$$\mu_2 = v_2 - v_1^2$$

$$= m + m^2 - (m)^2 = m \quad [since\ v_1 = m\ ]$$

Thus $\sigma^2 = m$ *and* $\sigma = \sqrt{m}$

Therefore, the constants of poisson distribution are:

$$\mu_1 = 0, \quad \mu_2 = m, \quad \mu_3 = m, \quad and \quad \mu_4 = m + 3m^2$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{m^2}{m^3} = \frac{1}{m}.$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{m + 3m^2}{m^2} = 3 + \frac{1}{m}.$$

One of the great advantage of poisson distribution is that we need only the value of the mean in order to compute the values of various constants. This shall be clear from the illustration below.

9.4.3 Applications

Example 4: The mean of the poisson distribution is 2.25. Find the other constants of the distribution.

Solution: We are given mean or m = 2.25

$$\sigma = \sqrt{m} = 1.5$$

$$\mu_1 = 0, \quad \mu_2 = m = 2.25, \quad \mu_3 = m = 2.25,$$

$$and \quad \mu_4 = m + 3m^2 = 2.25 + (2.25)^2 = 17.44$$

$$\beta_1 = \frac{1}{m} = \frac{1}{2.25} = 0.444.$$

$$\beta_2 = 3 + \frac{1}{m} = 3 + 0.444. = 3.444$$

Example 5: In the manufacture of glassware, bubbles can occur in the glass which reduces the status of the glassware to that of a 'second'. If, on average, one in every 1000 items produced has a bubble, calculate the probability that exactly six items in a batch of three thousand are seconds.

Solution:

Suppose that X= number of items with bubbles, then $X \sim$ B (3000, 0.001)

Since n = 3000 > 100 and p = 0.001 < 0.005 we can use the Possion distribution with λ = np = 3000 x 0.001 = 3. The calculatio is :

$$P(X = 6) = e^{-3}\frac{3^6}{6!} \approx 0.0498 \times 1.0125 \approx 0.05$$

The result means that we have about a 5% chance of finding exactly six seconds in a batch of three thousand items of glassware.

Example 6: A manufacturer produces light-bulbs that are packed into boxes of 100. If quality control studies indicate that 0.5% of the light-bulbs produced are defective, what percentage of the boxes will contain: (a) no defective? (b) 2 or more defectives?

Solution:

As n is large and p, the P(defective bulb), is small, use the Poisson approximation to the binomial probability distribution. If X = number of defective bulbs in a box, then

$X \sim \text{P}(\mu)$ where $\mu = n \times p = 100 \times 0.005 = 0.5$

(a) $\text{P}(X = 0) = \dfrac{e^{-0.5}(0.5)^0}{0!} = \dfrac{e^{-0.5}(1)}{1} = 0.6065 \approx 61\%$

(b) $\text{P}(X = 2 \text{ or more}) = \text{P}(X = 2) + \text{P}(X = 3) + \text{P}(X = 4) + \ldots$ but it is easier to consider:

$\text{P}(X \geq 2) = 1 - [\text{P}(X = 0) + \text{P}(X = 1)]$

$\text{P}(X = 1) = \dfrac{e^{-0.5}(0.5)^1}{1!} = \dfrac{e^{-0.5}(0.5)}{1} = 0.3033$

i.e. $\text{P}(X \geq 2) = 1 - [0.6065 + 0.3033] = 0.0902 \approx 9\%$

9.5 SUMMARY

In this lesson we have discussed in detail the theoretical discrete distributions i.e., Binomial and Poisson Distributions. Also we have discussed the importance and use of these distributions in the decision theory. Binomial distribution describes the distribution of binary data from a finite sample. Thus it gives the probability of getting r events out of n trials. Poisson distribution describes the distribution of binary data from an infinite sample. Thus it gives the probability of getting r events in a population.

9.6 GLOSSARY

- **Poisson Distribution:** The Poisson Distribution is defined as: $e^{-m}\ m^r$

  $P(r) = e^{-m}\, m^r/r!$
  e=2.7183 (the base of natural Logarithms)
  m=the mean of the Poisson distribution, i.e, np or
  the average number of occurrences of an event

- **Random Variable:** Random variable is a variable which takes up possible values whose outcomes are numerical from a random phenomenon.

- **Mean of Poisson Distribution:** The mean of a poisson distribution is denoted with 'm' and is calculated as n×p
- **Binomial Distribution:** A random variable X is said to follow binomial distribution if it assumes only non-negative values and its probability mass function is given by:

$$P(X = x) = p(x) = \begin{cases} \binom{n}{x} p^x q^{n-x} : x = 0,1,2,3,\ldots,n; q = 1-p \\ 0, \qquad \text{otherwise} \end{cases}$$

- **Continuous Variable:** A continuous variable is the opposite of a discrete variable that has an infinite number of possible values.
- **Discrete Variable**: A discrete variable is a variable that can only take on a contain number of values. If you can count a set of items, then it's a discrete variable. The opposite of a discrete variable is a continuous variable.

9.7 SELF ASSESSMENT QUESTIONS

A. Fill in the Blanks:

1. Binominal distribution is associated with the name of_____ mathematician_____.
2. The mean of binominal distribution is_____.
3. The variance of binominal distribution is _____.
4. If mean of Poisson distribution is 8, μ shall be _____.
5. In a poisson distribution as 'n' increases, the distribution shifts to the _____.
6. If in poisson distribution N(P1) = 932, and m=.02, then N(P2) shall be_____.

9.8 LESSON END EXERCISE

1. A sample of 3 items is selected at random from a box containing 12 items of which 3 are defective. Find the possible number of defective combinations of the said 3 selected items along with probability of a defective combination.

2. A student obtained the following answer to a certain problem given to him. Mean = 2.4, Variance = 3.2 for a binomial distribution. Comment on the result.

3. Twelve dice were thrown 4096 times. Each 4,5 or 6 spot appearing was considered to be a success while a 1,2 or 3 spot was a failure. Calculate the theoretical frequencies for 0, 1, 2,….,12 successes.

4. What is Binomial distribution? Give a real life example where such a distribution is appropriate?

5. What are the main characteristics of a Binomial distribution ?

6. Discuss the role of business statistics.

7. Show that the Poisson distribution is the limiting form of Binomial Distribution ?

8. Derive the mean and variance of poisson distribution.

9. In a town 10 accidents took place in a span of 50 days. Assuming that the number of accidents per day follows the poisson distribution, find the probability that there will be three or more accidents in a day.

9.9 SUGGESTED READING

- Levin, R.I. Robin, D.S. Statistics for Management, Prentice-Hall of India. New Delhi.
- Aczel, A. D. Sounderpandian, J. Complete Business Statistics, Mc Graw Hill Publishing. New Delhi. downloaded
- Anderson, S., W. Statistics for Business and Economics, Cengage Learning. New Delhi.
- Kazmeir L. J. Business Statistics, Tata Mc Graw Hill. New Delhi.
- Vohra, N. D. Business Statistics. Tata Mc Graw Hill. New Delhi.

ANALYSIS OF VARIANCE

STRUCTURE

10.1 INTRODUCTION

The analysis of variance frequently referred to by the contraction ANOVA is a statistical technique specially designed to test whether the means of more than two quantitative populations are equal. The t-test is an adequate procedure for testing the null hypothesis

when we have means of only two samples to consider. However, in a situation where we have three or more samples to consider at a time an alternative procedure is needed for testing the hypothesis that all samples could likely be drawn from the same population. For example, five fertilizers are applied to four plots. We may be interested in finding out whether the effect of these fertilizers on the yield is significantly different. The answer to this problem is providing by the technique of analysis of variance.

Today, procedure of this analysis finds application in nearly every type of experimentaldesign, in natural science as well as social science. In fact it has come to acquire a place of great prominence in statistical analysis. That is because of the fact that the analysis of variance is amazingly versatile: it can be readily adopted to furnish, with broad limits, a

proper evaluation of data obtained from a large body of experiments which involve several continuous random variables. It can give us answers as to whether different sample data classified in terms of a single variable area meaningful. It can also provide us with meaningful comparisons of sample data which are classified according to two or more variables. Thus, in this lesson we are trying to find out the significant difference between the means of more than two samples by using one way as well as two way classifications.

10.2 OBJECTIVES

After reading this lesson, you will be able to:

- Understand the assumptions and applications of F-test
- Distinguish between one way and two way classification of ANOVA
- Study the applications of ANOVA

10.3 CONCEPT OF ANALYSIS OF VARIANCE

The analysis of variance technique, developed by R.A. Fisher in 1920's, is capable of fruitful application to a diversity of practical problems. Basically, it consists of classifying and cross classifying statistical results and testing whether the means of a specified classification differ significantly. In this way it is determined whether the given classification is important in affecting the results. For example, the output of a given process might be cross-classified by machines and operators (each operator having worked on each

machine). From this cross- classification, it could be determined whether the mean qualities of the outputs of the various machines differed significantly. Also, it could independently be determined whether the mean qualities of the outputs of the various machine differed significantly. Such a study would determine, for example, whether uniformity in quality of outputs could be increased by standardising the procedures of the operators (say, through special training) and similarly whether it could be increased by standardising the machines (say, through resetting). Analysis of variance thus enables us to analyse the total variation of our data into components which may be attributed to various "sources" or "causes" of variation.

10.3.1 Assumptions of Analysis of Variance

1. Normality: It may be noted that whenever any of these assumptions is not met, the analysis of variance technique cannot be employed to yield valid inferences. It in indeed fortunate that many economic and business experiments do conform, at least approximately to these premises. In some cases in experimental work, departure from these assumptions also exists. In such situations the analysis of variance can still be applied after the transformation of the data.

2. Homogeneity: In practice it has been observed that one or more of these assumptions can be "bent" without appreciable loss in the adequacy of the F-test. The researcher strives to meet the assumptions of the F-test, but he usually finds that if the data are reasonably close to meeting the assumptions, his conclusion based on the F-test are not markedly affected. If the underlying distributions are bimodal or very skewed the F-test results may not be valid.

3. Independence of Error: Conspicuously greater the variance around the sample means, the samples must be, widely dispersed around the grand mean, very likely not representing random samples from the same population. However, if the sample means are very narrowly dispersed around the grand mean, compared with dispersion around their sample means, the samples are likely to be random samples from a common population.

10.4 ONE WAY CLASSIFICATION

In one-way classification, the data are classified according to only one criterion or factor.

The null hypothesis is:

H_O: $\mu_1=\mu_2=\mu_3=\ldots\ldots\ldots\ldots=\mu_K$

H_1:$\mu 1 \neq \mu 2 \mu 3 \ldots\ldots\ldots\ldots\ldots..\mu_K$

10.4.1 Computations

All the means are not equal that is the arithmetic means of populations from which the K samples were randomly drawn are equal to one another. The steps in carrying out the analysis when data are classified according to one factor only are:

1. **Calculate variance between the samples:** The variance between samples (groups) measured the differences between the sample mean of each group and the overall mean weighted by the number of observation of each group. The variance between samples taken into account the random variations from observation to observation. It also measures difference from one group to another. The sum of squares between samples is denoted by SSC. For calculating variance between the samples we take the total of the square of the deviations of the means of various samples from the grand average and divide this total by the degree of freedom. Thus the steps in calculating variance between samples will be:

a) Calculated the mean of each sample, i.e., $_{1, 2, ...etc}$

b) Calculate the grand average pronounced "X double bar". Its value is obtained as follows:

Grand Average: (mean of X_1 + mean of X_2 + mean of X_3 + ...) ($N_1 + N_2 + N_3$+....)

c) Take the difference between the means of the various samples and the grand average;

d) Square these deviations and obtain the total which will give sum of squares between the samples; and

e) Divide the total obtained in step (d) by the degrees of freedom. The degrees of freedom will be one less than the number of samples, i.e. if there are 4 samples than the degrees of freedom will be 4-1=3 or v=k-1,where k=number of samples.

2. Calculate variance within the samples: The variance (or sum of square) within-*x* samples measures those inter sample differences due to chance only. It is denoted by SSE. The variance within samples (groups) measures variability around the mean of each group. Since the variability is not affected by group differences it can be considered a measure of the random variation of values with in a group. For calculating the variance within the samples we take the total of the sum of squares of the deviation of various items from the mean values of the respective samples and divide this total by the degrees of freedom. Thus, the steps in calculating variance within the samples will be:

a) Calculate the mean value of each sample, i.e., $x_1, x_2, \ldots$ etc.;

b) Take the deviations of the various items in a sample from the mean values of the respective samples;

c) Square these deviations and obtain the total which gives the sum of square within the samples and

d) Divide the total obtained in step (c) by the degrees of freedom. The degree of freedom is obtained by deduction from the total number of items the number of samples, i.e., v=N-K, where K refers to the number of samples and N refers to the total number of all observations.

3. Calculate the ratio of F as follows:

$$F=\frac{Between-column\ variance}{Within-column\ variance}$$

Symbolically,

$$F=\frac{S_1^2}{S_2^2}$$

The F-distribution (named after the famous statistician R.A Fisher) measures the ratio of the variance between the groups to the variance with in the groups. The variance between the samples means is the numerator and the variance within the sample means is the denominator. If these is no real difference from group to group, any sample difference will be explainable by random variation and the variance between groups should be close to the variance within groups. However, if there is a real difference between the groups, the variance between groups will be significantly large than the variance within groups.

Compare the calculated value of F with the table value of F for degree of freedom at a certain critical level (generally we take 5% level of significance). If the calculated value of F is greater than the table value, it is concluded that the difference in sample means is significant, i.e, it could not have arisen due to fluctuations of simple sampling or, in other words, the samples do not come from the sample population. On the other hand, if the calculated value of F is less than the table value, the difference is not significant and has arisen due to fluctuations of simple sampling

It is customary to summarise calculations for sums of squares, together with the r numbers of degree of freedom and mean squares in a table called the analysis of variance table, generally abbreviated ANOVA. The specimen of ANOVA table is given below:

Table 1

Analysis of Variance (ANOVA) Table: One-way classification

Source of Variation	**SS (Sum of squares)**	**V Degrees of freedom**	**MS (Mean Square)**	**Variance Ratio of F**
Between the samples	SSC	$v_1 = c - 1$	*SSC* /c-1	MSC/MSE
Within the samples	SSE	$v_2 = n - c$	SSE/n-c	
Total	SST	$n - 1$		

SST = Total sum of squares of variations.

SSC = Sum of squares between samples (columns)

SSE = Sum of squares within samples (rows)

MSC = Mean sum of squares between samples

MSE = Mean sum of squares within samples.

Example 1: To assess the significance of possible variation in performance in a certain test between the grammar school 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.

Table 2

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:

Table - 3

Sample 1		**Sample 2**		**Sample 3**		**Sample 4**	
X_1	X_1^2	X_2^2	X^2_3	X_3	X_3^2	X_4	X_4^2
8	64	12	144	18	324	13	169
10	100	11	121	12	144	9	81
12	144	9	81	16	256	12	144
8	64	14	196	6	36	16	256
7	49	4	16	8	64	15	225
$\Sigma X_1=45$	$\Sigma X_1^2=421$	$\Sigma X_2=50$	$\Sigma X_2^2=558$	$\Sigma X_3=60$	$\Sigma X_3^2=824$	$\Sigma X_4=65$	$\Sigma X_4^2=875$

The sum of all items of various samples

$$= \sum X_1 + \sum X_2 + \sum X_2 + \sum X_2$$

$$= 45 + 50 + 60 + 65 = 220$$

$$\text{Correction factor} = \frac{T^2}{N} = \frac{(220)^2}{20} = \frac{18,400}{20} = 2,420$$

$$\text{Total sum of squares} = \sum X_1^2 + \sum X_2^2 + \sum X_3^2 + \sum X_4^2 - \frac{T^2}{N}$$

Sum of square within sample = Total sum of square -Sum of square between samples =258- 50 = 208

ANOVA TABLE

Source of Variation	SS (Sum of squares)	V Degrees of of freedom	MS (Mean Square)	Variance Ratio of F
Between the samples	50	3	16.7	1.28
Within the samples	208	16	13	
Total	**258**	**19**		

The table value for V_1=3 and V_2=16 at 5% level of significance = 3.24. The calculated value of F is less then table value of hence, the mean difference in samples is signifi- cant. The sample could have come from same universe.

10.5 TWO WAY CLASSIFICATION

When it is believed that two independent factors might have an effect on the response variable of interest, it is possible to design the test so that an analysis of variance can be used to test for the effect of the two factors simultaneously such a test is called a two-factor analysis of variance. With the two factor analysis of variance, we can test sets of hypothesis with the same data at the same time.

In two way classification the data are classified according to two different criteria or factors. The procedure for analysis of variance is somewhat different than the one followed while dealing with problems of one-way classification.

10.5.1 Computations

In a two-way classification the analysis of variance table takes the following form:

Table 4

Sources of Variation	Sum of squares	Degrees of Freedom	Mean sum of squares	Ratio of F
Between Samples	SSC	(c-1)	MSC=SSC/ (c-1)	MSC/MSE
Between Rows	SSR	(r-1)	MSR=SSR/(r-1)	MSR/MSE
Residual or Error	SSE	(c-1)(r-1)	MSE=SSE/(c-1)(r-1)	
Total	SST	$n-1$		

SSC = Sum of squares between columns
SSR = Sum of squares between rows
SSE = Sum of squares due to error
SST = Total sum of squares

The sum of square for the source 'Residual' is obtained by subtracting from the total sum of squares the sum of squares between columns and rows, i.e., SSE = SST- [SSC + SSR]

The total number of degree of freedom = n-1 or cr-1

Where c refers to number of columns, and r refers to number of rows.

Number of degrees of freedom between columns = (c-1)

Number of degrees of freedom between rows = (r-1)

Number of degrees of freedom for residual = (c-1) (r-1)

The total number of squares, sum of square for 'between columns' and sum of squares for 'between rows' are obtained in the same way.

Residual or error sum of square = total sum of square - sum of square between column-sum of squares between rows.

The F values are calculated as follows: $F(v1, v3) = \frac{MSC}{MSE}$

Where v1=(c-1) and v2=(r-1) (c-1)

$F(v2, v3) = \frac{MSR}{MSE}$

Where v_1=(r-1) and v_2=(c-1) (r-1)

The calculated values of F are compared with the table values. If calculated value of F is greater than the table value at pre-assigned level of significance, the null hypothesis is rejected, otherwise accepted.

Example 2: The following data represent the number of units of production per day tuned out by five different workers using 4 different types of machines.

Table 5

		→ **Machines**			
		A	B	C	D
	1	44	38	47	36
	2	46	40	52	43
Workers	3	34	36	44	32
↓	4	43	38	46	33
	5	38	42	49	39

(a) Test whether the mean productivity is the same for the different machine types.

(b) Test whether the five men differ with respect to mean productivity.

Solution:

Let us take the hypothesis that (a) the mean productivity is the same for four different machines, and (b) the 5 men do not differ with respect to mean productivity.

To simplify calculations let us divide each value by 40. The coded data is given below:

Table 6

Workers	Machine type				Total
	A	B	C	D	
1	+4	-2	+7	-4	+5

2	+6	0	+12	+3	+21
3	-6	-4	+4	-8	-14
4	+3	-2	+6	- 7	0
5	-2	+2	+9	-1	+8
TOTAL	+5	-6	+38	-17	T=20

Correction factor = T^2/N = = 20

Sum of squares between machines

= $(5)^2/5 + (-6)^2/5 + (38)^2/5 + (-17)^2/5$ – correction factor

= (5+7.2+288.8+57.8)-20

=358.8-20 =338.8

V=(c-1) = (4-1) =3

Sum of squares between the workers:

$$= \frac{(5)^2}{5} + \frac{(-6)^2}{5} + \frac{(38)^2}{5} + \frac{(-17)^2}{5} \text{ - Correction factor}$$

= (5+7.2+288.8+57.8)-20

=358.8-20 =338.8

V=(c-1) = (4-1) =3

Sum of squares between workers

$$= \frac{(5)^2}{4} + \frac{(21)^2}{4} + \frac{(-14)^2}{4} + \frac{(0)^2}{4} + \frac{(8)^2}{4} - \frac{T^2}{N}$$

$$= \frac{25}{4} + \frac{441}{4} + \frac{196}{4} + \frac{0}{4} + \frac{64}{4} \text{ - 20}$$

= (6.25+110.25+49+0+16)-20

=181.5-20 =161.5

V=(r-1) = (5-1) =4

Total Sum of squares = $A^2 + B^2 + C^2 + D^2$ = 574

Residual or Remainder=Total sum of squares - (sum of squares between machines - sum of squares between workers)

=574-338.8-161.5 =73.7

Degree of freedom for remainder =19-3-4=12

(c-1)(r-1)=(3×4)=12

Table 7

Analysis of Variance Table

Sources of variation	S.S	d.f	M.S	Variance ratio or F
Between machine types	338.8	3	112.933	112.933 / 6.142=18.387
Between workers	161.5	4	40.375	40.375 / 6.142 = 6.574
Remainder or Residual	73.7	12	6.142	
	574	19		

(a) For v =12, $F_{(0.05)} = 3.49$

Since the calculated value (18.4) is greater than the table value, we conclude that the mean productivity is not same for the four different types of machines.

(b) For v = 12, $F_{(0.05)} = 3.26$

The calculated value (6.58) is greater than the table value. Hence the workers differ with respect to mean productivity.

10.6 SUMMARY

The student's t-test is used for testing the hypothesis of equality of two normal population means when sample size is small. However, in testing of equality of more than two means, t-test cannot be used. In such situations we use analysis of variance. The analysis of Variance developed by R.A. Fisher, is one of the most powerful tools of statistical analysis. The analysis of Variance is a method of splitting the total variation into different components that measure different sources of variation. The only difference between one-way and two-way ANOVA is the number of independent variables. A one-way ANOVA has one independent variable, while a two-way ANOVA has two.

- **One-way ANOVA**: Testing the relationship between shoe brand (Nike, Adidas, Saucony, Hoka) and race finish times in a marathon.
- **Two-way ANOVA**: Testing the relationship between shoe brand (Nike, Adidas, Saucony, Hoka), runner age group (junior, senior, master's), and race finishing times in a marathon.

10.7 GLOSSARY

- **ANOVA**: ANOVA is a collection of Statistical model and their associated estimation procedures used to analyse the differences among group means in a sample.
- **ONE WAY ANOVA**: Comparison of means of three or more within-subject variables.
- **ANCOVA:** Any ANOVA Model with a Covariate.
- **MANOVA:** Any ANOVA model with Multiple DVs.

10.8 SELF ASSESSMENT QUESTIONS

A. Multiple Choice Questions:

1. The technique of analysis of variance was developed by _____.
2. The sum of squares between samples is denoted by _____.
3. _____stands for mean square between samples.
4. ANOVA table stands for _____.
5. The analysis of variance procedure is appropriate for testing equivalence of a set of two or more populations

10.9 LESSON END EXERCISE

1. The following data shows the lives in hours of four batches of electric tubes.

Batches	Lives in hours
A	1700 1710 1750 1780 1800 1820 1900
B	1680 1740 1740 1800 1850
C	1560 1650 1700 1720 1740 1760 1840 1920
D	1610 1620 1630 1670 1700 1780

Perform the analysis of variance and show that four batches are homogeneous.

2. The following figure related to the production of three varieties of wheat used in 15 plots:

Variety of wheat	Yield (in kg)
A	19 22 21 21 3
B	20 16 18 20 18 19
C	23 21 23 24 20

Test whether there is any significance difference in the production of three varieties of wheat.

3. The following data represent the number of units of production per day produced by 5 different workers using four different machines.

	Machines			
Workers	M_1	M_2	M_3	M_4
W_1	54	48	57	46
W_2	56	50	62	53
W_3	44	46	54	42
W_4	53	48	56	43
W_5	48	52	59	49

Test whether the :

a) Mean productivity is the same for different machines

b) The workers differ with respect to mean productivity.

4. Using the data of (2) and (3), carry out the analysis of variance after shifting the origin and state your conclusions.

5. Distinguished between one way and two way analysis of variance.

6. How is ANOVA technique helpful in solving business problems ? Illustrate your answer with suitable examples.

7. What are the basic and common assumptions made for analysis of variance ?

_____

_____

_____

10.10 SUGGESTED READING

- Levin, R.I. Robin, D.S. *Statistics for Management*, Prentice-Hall of India. New Delhi.
- Aczel, A. D. Sounderpandian, J. *Complete Business Statistics*, Mc Graw Hill Publishing. New Delhi. downloaded
- Anderson, S., W. *Statistics for Business and Economics*, Cengage Learning. New Delhi.
- Kazmeir L. J. *Business Statistics*, Tata Mc Graw Hill. New Delhi.
- Vohra, N. D. *Business Statistics*. Tata Mc Graw Hill. New Delhi.

CONCEPT AND TERMINOLOGY OF ASSOCIATION OF ATTRIBUTES AND CONSISTENCY OF DATA

STRUCTURE

11.1 INTRODUCTION

Generally statistics deal with quantitative data only. But in behavioural sciences, one often deals with variable which is not quantitatively measurable. Literally an attribute means a quality or characteristic which is not related to quantitative measurements. Examples of attributes are health, honesty, blindness etc. They cannot be measured directly. The observer may find the presence or absence of these attributes. Statistics of attributes based on descriptive character. An attribute refers to the quality of a characteristic. The theory of attributes deals with qualitative types of characteristics that are calculated by using quantitative measurements. Therefore, the attribute needs slightly different kinds of statistical treatments, which the variables do not get. Attributes refer to the characteristics of the item under study, like the habit of smoking, or drinking. So 'smoking' and 'drinking' both refer to the example of an attribute.

In the theory of attributes, the researcher put more emphasis on quality (rather than on quantity). Since the statistical techniques deal with quantitative measurements, qualitative data is converted into quantitative data in the theory of attributes.

There are certain representations that are made in the theory of attributes. The population in the theory of attributes is divided into two classes, namely the negative class and the positive class. The positive class signifies that the attribute is present in that particular item under study, and this class in the theory of attributes is represented as A, B, C, etc. The negative class signifies that the attribute is not present in that particular item under study, and this class in the theory of attributes is represented as α, β, etc.

The assembling of the two attributes, i.e. by combining the letters under consideration (such as AB), denotes the assembling of the two attributes. This assembling of the two attributes is termed as dichotomous classification. The number of observations that have been allocated in the attributes is known as the class frequencies. These class frequencies are symbolically denoted by bracketing the attribute terminologies i.e. (B), which stands for class frequency of the attribute B. The frequencies of the class also have some levels in the attribute. For example, the class that is represented by the 'n' attribute refers to the

class that has the nth order. For example, (B) refers to the class of 2nd order in the theory of attributes.

There is also independence nature in the theory of attributes. The two attributes are said to be independent only if the two attributes are absolutely uncorrelated to each other.

In the theory of attributes, A and B are said to be associated with each other only if the two attributes are not independent, but are related to each other in some way or another.

The positive association in the two attributes exists under the following condition:

$(AB) > (A)(B)/N$.

The negative association in the two attributes exists under the following condition:

$(AB) < (A)(B)/N$.

The situation of complete association in the two attributes arises when the occurrence of attribute A is completely dependent upon the occurrence of attribute B. However, attribute B may occur without attribute A, and the same thing holds true if attribute A is the independent one.

Ordinarily, the two attributes are said to be associated if the two occur together in a number of cases.

Therefore, qualitative characteristics such as deafness, blindness, employment, beauty, hair colour, sex etc. of an individual of a universe or population are termed as attributes. The attributes are not orderable into series from least to most or vice versa.

The classification which divides a group into two classes according to one attribute called classification by Dichotomy or simple classification. The classification which divides the group into more than two classes according to one attributes is called manifold classification. For example, according to the attribute "Hair Colour" the population of a city may be divided into different classes:

i. Fair-haired people

ii. Red-haired people

iii. Brown haired people

iv. Black haired people

If several (more than two) attributes are noted, the process of classification may however, be continued indefinitely. Such type of classification may be called classification as a series of dichotomies. For example, consider two attributes, namely "Blindness and Deafness". The people of a city may be first divided into two classes according to the attribute 'Blindness' and then each of these two classes may further be classified according to the attribute 'Deafness'. And therefore, ultimately we have four classes:

i. The class of blind and deaf people.

ii. The class of blind and non-deaf people.

iii. The class of non-blind and non-deaf people.

iv. The class of non-blind and deaf people.

11.2 OBJECTIVES

After studying this lesson, you will be able to:

- Know the concept of association of attributes.
- How correlation is different from association.
- Learn about different terminology used in the study of association.
- Solve exercises based on these concepts.
- Provide the concept of consistency of data.
- Understand the concept of association and disassociation

11.3 CONCEPT OF ASSOCIATION

A pointed out earlier, statistics deals with quantitative phenomenon only. However, the quantitative character may arise in any of the following two ways:

1. In the first place, we may measure the actual magnitude or size of some phenomenon.

For example, we may measure the height of students of a class, their weight, etc. Similarly, we may study the wage structure of the workers of a particular factory, the amount of rainfall in a year. The characteristics of this type of phenomenon are known as statistics of variables. The various statistical techniques like measures of central tendency, dispersion, correlation etc. deals with such variables.

2. In the second place, there are certain phenomenons like blindness, deafness, etc. which are not capable of direct quantitative measurement. In such cases the quantitative character arises only indirectly in the process of counting. For example, we can determine out of 1000 persons, how many are blind and how many are not blind but we cannot precisely measure blindness. Such phenomenon, where direct quantitative measurement is not possible, i.e., where we can study only the presence or absence of particular characteristics, is called as statistics of attributes.

11.3.1 Difference between Correlation and Association

The tool of correlation is used to measure the degree of relationship between two such phenomena as are capable of direct quantitative measurement. On the other hand, the method of association of attributes is employed to measure the degree of relationship between two phenomena whose size we cannot measure and where we cannot only determine the presence or absence of a particular attribute.

While dealing with statistics we have to classify the data. The classification is done on the basis of presence or absence of particular attribute or characteristic. When we are studying only one attribute, two classes are formed-one possessing that attribute and another not possessing it. For example, when we are studying the attribute employment, two classes shall be formed, i.e., those who are employed and those who are not employed. When two attributes are studied, four classes shall be formed. If, besides employment, we study the gender-wise distribution, four classes shall be formed:

number of males employed, number of females employed, number of males unemployed and number of females unemployed.

It should be noted that in some cases while classifying the attributes no clear-cut definition of an attribute and line of demarcation between classes can be drawn.

For example, when the attribute 'employment' is being studied the data are classified into 'Employed' and 'Unemployed'. But there can be further category of those people who are partially employed (i.e. part-time). Also there may be some persons who are employed before the survey but on the date of survey they are unemployed. So we cannot treat them as employed and also as unemployed because there is some difference between those persons who have not got any job, and those who have got some job but were retrenched after some time. Hence, it is absolutely essential to lay down clear-cut definition of the various attributes under Study. This is often a difficult task. Hence, this limitation must be kept in mind while studying association between attributes.

11.4 NOTATION AND TERMINOLOGY

To facilitate the theory of attributes it is necessary to have some standard notations for the attributes, the classes formed and for the observations assigned to each of them. Therefore, for the sake of simplicity and convenience it is imperative to use certain symbols to represent different classes and their frequencies. It is customary to use capital letters A and B to represent the presence of attributes, may be called as positive attributes. Whereas, the Greek letters and α and β are generally used to represent absence of attributes A and B also called as negative attributes. Thus 'α' = not A and 'β' = not B. Therefore, the classes A and α, B and β are complementary to each other. Combination of the attributes will be represented by just a position of letters. For example, if A denotes blindness and B denotes deafness, then AB denotes blindness and deafness. Similarly, of A denotes 'married', B denotes 'man' and c denotes 'left handed', then:

$A\beta$ denotes married women

AB denotes married men

αB denotes unmarried men

αC denotes unmarried left handed

ABC denotes married left handed men

$A\beta\gamma$ denotes married right handed women

Any combination of letters say, A, AB, $\alpha\beta$, αB etc. by means of which we specify the

characteristics of the member of a class, may be called as class symbols. A collection of all individuals is denoted by N. Further, if A represents males then α would represent females. Similarly if B represents literates then β would represent illiterates.

Lets look at the combination

(AB): number of literate males

(Aβ): number of illiterate males

(αB): number of literate females

(α β): number of illiterate females

11.4.1 Classes and Class frequencies

Different attributes in themselves, their sub-groups and combinations are called different classes and the numbers of observations assigned to them are called their class frequencies. If two attributes are studied the number of classes will be 9, i.e., (A), (α), (B), (β), (Aβ), (α β), (αB) and N.

The number of observations or units belonging to each class is known as their frequencies are denoted within bracket. Thus (A) stands for the frequency of A and (AB) stands for the number objects possessing the attribute both A and B. Further, the class frequencies of the type (A), (B), (AC), (BC), (ABC), etc. which involve only positive attributes are called as positive frequencies. The class frequencies of the type (α), (β), (αβ), (βγ), (αβγ) etc. which involve only negative attributes are called negative frequencies. The class frequencies of the type (Aβ), (αB), (ABγ), (αBC) etc. which involve the mixture of positive and negative attributes are called as contrary frequencies.

11.4.2 Order of Classes and Class Frequencies

A class represented by n attributes is called a class of nth order and the corresponding frequency as the frequency of the nth order. Thus (A), (B), (γ)...etc. are class frequencies of order 1; (AB), (αB), (αγ), (βγ), etc. are class frequencies of second order; (ABC), (ABγ), (αβγ) etc., are frequencies of third order and so on. N, the total number of members of the population, without any specification of attributes, is known as a frequency of zero-

order. Thus, the order of classes and class frequencies depend upon the number of attributes assign to a particular class.

In general, the following rules are used to determine the class frequencies:

Rule 1: With n attributes there are in all 2^n positive classes.

Rule 2: With n attributes the number of classes is 3^n.

That is,

i. For one attribute, there are three ($3^1 = 3$) frequencies.

ii. For two attributes, there are 9 ($3^2 = 9$) frequencies.

iii. For three attributes, there are 27 ($3^3 = 27$) frequencies.

11.4.3 Ultimate Class Frequencies

The classes specified by n attributes i.e. those of highest order, are called as ultimate class frequencies. Every class frequency can be expressed as the sum of the ultimate class frequencies. If these are given, the data can be completely determined. If there are n attributes, then there will be 2n ultimate class frequencies. If two attributes are studied then the number of classes of ultimate class order shall be $2^2 = 4$. In case of three attributes, there would be $2^3 = 8$ classes of the ultimate order.

Therefore, for two attributes A and B, there are 22 = 4 ultimate class frequencies denoted with; (AB), (Aβ), (αβ), (αB). And in case of three attributes say A, B and C, the number of ultimate class frequencies would be $2^3 = 8$, i.e., (ABC), (ABγ), (αβγ), (αBC), (αβC), (αβC), (αBγ), (Aβγ).

Also, total number of ultimate class frequencies = total number of positive class frequencies.

11.4.4 Contingency Table

A table which represents the classification according to the distinct classes of two characteristics A and B is called a two way contingency table. Also, it is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables. This table provide a basic picture of the interrelation between two variables and can help find

interactions between them. Suppose the attribute A has m distinct classes denoted by A_1, A_2,.....A_m and the attribute B has n distinct classes B_1, B_2,.....B_n then there are in all mn distinct classes (called cells) in the contingency table. In the contingency table, the total of various rows A1, A.....Am and total of various columns B_1, B_2,....B_n give the first order frequencies and cells have the frequencies of second order. The grand total of all frequencies gives the total number of observations, i.e. N. The contingency table can be written as:

Table 1: Contingency Table

		Attribute A				
		A_1	A_2	A_3	A_4	Total
	B_1	(A_1B_1)	(A_2B_1)	(A_3B_1)	(A_4B_1)	(B_1)
Attribute B	B_2	(A_1B_2)	(A_2B_2)	(A_3B_2)	(A_4B_2)	(B_2)
	B_3	(A_1B_3)	(A_2B_3)	(A_3B_3)	(A_4B_3)	(B_3)
	B_4	(A_1B_4)	(A_2B_4)	(A_3B_4)	(A_4B_4)	(B_4)
	Total	(A_1)	(A_2)	(A_3)	(A_4)	N

The classification by dichotomies with two attributes A and b is generally known as 2*2 (read as two by two) contingency table. The contingency table of order (2×2) for two attributes A and B can be displayed as given below:

Table 2

Attribute	**B**	**β**	**Total**
A	(AB)	(Aβ)	(A)
α	(αB)	(αβ)	(α)
Total	(B)	(β)	(N)

With the help 2*2 contingency table, we can find the ultimate class frequencies from the positive class frequencies with two attributes.

(A) + (α) = N ; (AB) + (Aβ) = (A); (AB) + (αB) = (B); (B) + (β) = N ; (αB) + (αβ) = (B); (Aβ) + (αβ) = (β)

11.4.5 Relationship between the Class Frequencies

All the class frequencies of various orders are not independent of each other and any class frequency can always be expressed in terms of class frequencies of higher order. Thus:

$N = (A) + (\alpha) = (B) + (\beta) = (C) + (\gamma)$ etc.

In case of Two attributes A and B

Also, since each of these A's and α' s can either be B's and β' s, we have

$N = (A) + (\alpha)$; $(A) = (AB) + (A\beta)$; $(B) = (AB) + (\alpha B)$; $N = (B) + (\beta)$; $(\alpha B) + (\alpha\beta) = (B)$; $(A\beta) + (\alpha\beta) = (\beta)$

Also, $N = (A) + (\alpha) = (AB) + (A\beta) + (A\beta) + (\alpha B) + (\alpha\beta)$

Thus in case of two attributes all the class frequencies can be expressed in terms of the ultimate class frequencies.

The frequency of a lower order class can always be expressed in terms of the higher order class frequencies. If the number of attributes is n, then there will be 3^n classes and we have 2^n cell frequencies.

In case of three attributes A, B and C

$(C) = (AC) + (\alpha C) = (ABC) + (A\beta C) + (\alpha BC) + (\alpha\beta C)$

Also, $(C) = (BC) + (\beta C) = (ABC) + (\alpha BC) + (A\beta C) + (\alpha\beta C)$

$(AB) = (ABC) + (\beta C)$; $(A\beta) = (A\beta C) + (A\beta\gamma)$

$(\alpha B) = (\alpha BC) + (\alpha B\gamma)$; $(\alpha\beta) = (\alpha\beta C) + (\alpha\beta\gamma)$

From above we get:

$N = (ABC) + (A\beta\gamma) + (A\beta C) + (A\beta\gamma) + (\alpha BC) + (\alpha B\gamma) + (\alpha\beta C) + (\alpha\beta\gamma)$

$(A) = (ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma)$

$(B) = (ABC) + (AB\gamma) + (\alpha BC) + (\alpha B\gamma)$

Similarly, we can express all other class frequencies in terms of the class frequencies of the third order. The above discussion leads to the following result.

"Any class frequency can be expressed as the sum of the 2^n ultimate class frequencies."

Therefore, in dichotomous classification of attributes, the data can be specified completely by:

i. The set of all the ultimate class frequencies.

ii. The set of all the positive frequencies.

Buy this we mean that if we know all the ultimate class frequencies or all the positive class frequencies, then we can obtain the frequencies of all other classes of different orders.

11.5 CONSISTENCY OF DATA

In order to find out whether the given data are consistent or not we have to apply a very simple test. The test is to find out whether any one or more of the ultimate class-frequencies is negative or not. If none of the class frequencies is negative we can safely calculate that the given data are consistent (i.e. the frequencies do not conflict in any way each other). On the other hand, if any of the ultimate class frequencies comes out to be negative the given data are inconsistent.

Example 1: Given N = 2500, (A) = 420, (AB) = 85 and (B) = 670. Find the missing values.

Solution:

We know N = (A) + (α) = (B) + (β)

(A)= (AB) + (Aβ)

(α) = (αB) + (αβ)

(B)= (AB) + (αB)

(β) = (Aβ) + (αβ)

(β) = (Aβ) + (αβ)

From (2) 420 = 85 + (Aβ)

∴ (Aβ) = 420 –85

(Aβ) = 335

From (4) 670 = 85 + (αB)

4" (αB) = 670"85

(αB) = 585

From (1) 2500 = 420 + (α)

4" (α) = 2500"420

(α) = 2080

From (1) (β) = 2500"670

(β) = 1830

From (3) = 2080 = 585 + (αβ)

4" (αβ) = 1495

Example 2: Test the consistency of the following data with the symbols having their usual meaning. N = 1000 (A) = 600 (B) = 500 (AB) = 50

Solution: We are given with:

N = 1000; (A) = 600; (B) = 500, (AB) = 50

We have to find the missing class frequencies with the help contingency table as under:

Attributes	**B**	**β**	**Total**
A	(AB) = 50	(Aβ)= 600-500= 100	(A) = 600
α	(αB) = 500-50 = 450	(αβ) = 400-450 = -50	(α) = 1000-600 = 400
Total	(B) = 500	(β) = 1000-500 = 500	N = 1000

Since (αβ) = " 50, the given data is inconsistent

Example 3: Examine the consistency of the given data. N = 60 ; (A) = 51; (B) = 32; (AB) = 25.

Solution: We are given with:

N = 60; (A) = 51; (B) = 52, (AB) = 50

We have to find the missing class frequencies with the help contingency table as under:

Attributes	**B**	**β**	**Total**
A	(AB) = 50	(Aβ)= 51-50= 1	(A) = 51
α	(αB) = 52-50 = 2	(αβ) = 9- 2 = 7	(α) = 60-51 = 9
Total	(B) = 52	(β) = 60-32 = 28	N = 60

Since all the frequencies are positive, it can be concluded that the given data are consistent.

11.6 ASSOCIATION AND DISASSOCIATION

Two attributes A and B are said to be associated if they are not independent but are related in some way or the other. There are three kinds of associations, which possibly occur between attributes.

a. Positive association

b. Negative association or disassociation

c. No association or independence.

In positive association, the presence of one attribute is accompanied by the presence of other attribute. For example, health and hygiene are positively associated.

In negative association, the presence of one attribute say A ensures the absence of another attribute say B or vice versa. For example, vaccination and occurrence of disease for which vaccine is meant are negatively associated. Therefore, disassociation doesn't indicate that there is no association between two or more attributes but rather it indicate negative association.

If two attributes are such that presence or absence of one attribute has nothing to do with the absence or presence of another, they are said to independent or not associated. For example, Honesty and Boldness.

In statistics two attributes A and B are associated if:

$(AB) \neq (A)(B)/N$

and Disassociated if:

$$(AB) < (A)(B)/N$$

11.7 SUMMARY

In this lesson, we have discussed:

1. Meaning of association in Statistics is different from the general meaning of association. In Statistics, attributes A and B are associated only if they appear together in greater number of cases than is to be expected if they are independent. In common language association means if A and B occur together a number of times then A and B are associated;
2. In association, we study the relationship between two attributes, which are not quantitatively measured;
3. Correlation coefficient measures the extent of relationship between two quantitative variables, whereas coefficient of association only suggests that the association is positive or negative;
4. If there exist no relationship of any kind between two attributes then they are said to be independent otherwise are said to be associated.

11.8 GLOSSARY

- **Attributes:** A qualitative characteristics are termed as attributes.
- **Positive attributes:** The presence of attribute is termed as positive attributes.
- **Class Frequency:** The number of observations assigned to any class is called as class frequency.
- **Negative Attributes:** The absence of attributes is called as negative attributes.
- **Ultimate Class Frequency:** The Class frequency specified by highest order is

called as ultimate class frequency.

- **Consistency:** The necessary and sufficient condition for the consistency of set of class frequencies is that none of the ultimate class frequency should be negative.
- Disassociation: In statistics, disassociation means negative relationship between two attributes.
- Association: In Statistics, association tells us whether two or more attributes are related.

11.9 SELF ASSESSMENT QUESTIONS

1. Fill in the blanks:

a) Association of attributes helps us to study the relationship between phenomena which are of .....nature.

b) (AB) denotes the number of individuals possessing attributes.....

c) When we study two attributes, the total frequencies are.....

d) The order of classes depends upon the number of .....under study.

2. Tick (✓) the correct option:-

1. Which of the following describe the middle part of a group of numbers?

   a) Measure of Variability b) Measure of Central Tendency

   c) Measure of Association d) Measure of Shape

2. If an attribute has two classes, it is called:

   a) Trichotomy b) Simple classification

   c) Dichotomy d) Manifold classification.

3. If an attribute has more than two classes, it is said to be:

a) Mainfold classification b) Trichotomy

c) Dichotomous d) All of the above

5. The total of all frequencies n is of order:

a) Zero b) One

c) Two d) Three

6. In case of consistent data, no class frequency can be:

a) Positive b) Negative

c) Both (a) and (b) d) Neither (a) and (b)

7. With two attributes A and B, the total number of ultimate frequencies is:

a) Two b) Four

c) Six d) Nine

11.10 LESSON END EXERCISE

1. What is the difference between variables and attributes?

_____

_____

_____

2. Differentiate between correlation and association.

_____

_____

_____

3. Explain consistency of data.

_____

_____

_____

4. For three attributes A, B and C. Write down all the class frequencies of order zero,

one, two and three.

5. Is there any inconsistency in the data given below:

   i. N= 1000, (A)=150, (B)=300, (AB)=200

   ii. N=1000, (A)=50, (B)=60, (AB)=20

6. Out of total population of 1000 the number of vaccinated persons was 600. In all 200 had an attack of smallpox and out of these 30 were those who were vaccinated. Do you find any association between vaccination and freedom from attack?

7. From the following given frequencies. Find the other missing frequencies:

   $(AB) = 250$, $(A\beta) = 120$, $(\alpha B) = 200$, $(\alpha\beta) = 70$

11.11 SUGGESTED READING

- Gupta, S.P.: *Statistical Methods*, Sultan Chand & Sons, New Delhi.
- Gupta, S.C. and V.K. Kapoor : *Fundamentals of Applied Statistics.*
- Levin, Richard and David S Rubin: *Statistics for Management*, Prentice Hall, Delhi.
- Levin and Brevson: *Business Statistics,* Pearson Education, New Delhi.
- Hooda, R.P.: *Statistics for Business and Economics,* Macmillan, New Delhi.

METHODS OF ATTRIBUTES COMPARISON, PROPORTION, YULE'S COEFFICIENT OF ASSOCIATION, COLLIGATION AND CONSISTENCY METHOD

STRUCTURE

12.1 INTRODUCTION

Association of Attributes is when data is collected on the basis of some attribute or qualitative data we have statistics commonly termed as statistics of attributes. It is not necessary that the objects may process only one attribute; rather it would be found that the objects possess more than one attribute. In such a situation our interest may remain in knowing whether the attributes are associated with each other or not. Technically, we say that the two attributes are associated if they appear together in a greater number of cases than is to be expected if they are independent and not simply on the basis that they are appearing together in a number of cases as is done in ordinary life. The association may be positive or negative (negative association is also known as disassociation). Independent, positive association, negative association etc. are known as nature of association.

In this lesson we have discuss about the nature of association, i.e., whether two or more attributes are independent, positively associated and negatively associated with the help of comparison and proportion method. But the statistician is more interested in the degree of association along with its nature. Therefore, further this lesson would explain the degree of association between two attributes by using an important method, i.e., Yule's coefficient of association along with coefficient of colligation and coefficient of contingency. Therefore, in this lesson our purpose is to explore the nature as well as the degree of association between two attributes by using the methods mentioned here above.

12.2 OBJECTIVES

After studying this lesson, you will be able to:

- Assess what kind of associations among attributes is likely to occur.
- Find the nature of association between two attributes by using comparison and proportion method.
- Explain the degree of association between two attributes by using Yule's coefficient of association
- Describe the concept of contingency table for manifold classification;

- Compute the expected frequencies for different cells, which are necessary for the computation of chi-square;
- Calculate coefficient of contingency and interpret the level of association with the help of it.
- Clarify the difference between coefficient of colligation and Yule's coefficient of colligation.

12.3 METHODS OF ATTRIBUTES : COMPARISON METHOD

This method is also known as comparison of observed and expected frequency method. This is so because when comparison method is applied, the actual observations are compared with expected observations. If actual observation is equal to the expectation, the attributes are said to be independent; if actual observation is more than the expectation, the attributes are said to be positively associated and if the actual observation is less than the expectation, the attributes are said to be negatively associated.

Symbolically, attributes A and B are:

(i) Independent if $(AB) = \frac{(A)\times(B)}{N}$

(ii) Positively associated if $(AB) > \frac{(A)\times(B)}{N}$

(iii)Negatively associated if $(AB) < \frac{(A)\times(B)}{N}$

The same is true for attributes α and B; α and β; A and β. Thus, attributes α and β are called:

(i) Independent if $(\alpha\ \beta) = \frac{(\alpha)\times(\beta)}{N}$

(ii) Positively associated if $(\alpha\ \beta) > \frac{(\alpha)\times(\beta)}{N}$

(iii)Negatively associated if $(\alpha\ \beta) < \frac{(\alpha)\times(\beta)}{N}$

Example 1: From the following data, find out whether: A and B are independent, associated and disassociated, if N = 100, (A) = 40, (B) = 80, (AB) = 30.

Solution: With comparison method, attributes shall be:

(i) Independent if (AB) = $\frac{(A)\times(B)}{N}$

(ii) Positively associated if (AB) > $\frac{(A)\times(B)}{N}$

(iii)Negatively associated or Disassociated if (AB) < $\frac{(A)\times(B)}{N}$

Therefore, here in our case, (AB) = 30, (A) = 40, (B) = 80 and N = 100

$$\frac{(A) \times (B)}{N} = \frac{40 \times 80}{100} = 32$$

Thus, (AB) = 30 is less than $\frac{(A)\times(B)}{N} = 32$

Hence, attributes A and B are disassociated.

Example 2: From the following ultimate class frequencies, find the frequencies of the positive and negative classes and the total number of observations.

(AB) = 100, (αB) = 80, (Aβ) = 50, (αβ) = 40.

Solution: By putting these values in the nine square table, we can find the desired information.

	A	**α**	**Total**
β	100	80	180
β	50	40	90
Total	150	120	270

N = (AB) + (Aβ) + (αB) + (αβ)
= 100 + 50 + 80 + 40 = 270

(A) = (AB) + (Aβ) = 100 + 50 = 150

(B) = (AB) + (αB) = 100 + 80 = 180

(α) = (αB) + (αβ) = 80 + 40 = 120

(β) = (Aβ) + (αβ) = 50 + 40 = 90

12.3.1 Limitations

With the help of this method we can only determine the nature of association (i.e., whether there is positive or negative association or no association) and not the degree of association (i.e., where association is high or low). Yules's coefficient is superior because it provides information not only on the nature but also on the degree of association.

12.4 PROPORTION METHOD

If there is no relationship of any kind between two attributes A and B, we expect to find the same proportion of A's amongst the B's as amongst the β' s. Thus, if a coin is tossed we expect the same proportion of heads irrespective of whether the coin is tossed by the right hand or the left hand.

Symbolically, two attributes may be termed as:

(i) Independent if $= \frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)}$

(ii) Positively associated if $>$ if $\frac{(AB)}{(B)} > \frac{(A\beta)}{(\beta)}$

Negatively associated or disassociated if $< \frac{(AB)}{(B)} < \frac{(A\beta)}{(\beta)}$

This case is application when B is taken as proportion.

Similarly, when A is taken as proportion, the attributes are:

(i) Independent if $= \frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)}$

(ii) Positively associated if $> \frac{(AB)}{(A)} > \frac{(\alpha B)}{(\alpha)}$

(iii) Negatively associated or disassociated if $< \frac{(AB)}{(A)} < \frac{(\alpha B)}{(\alpha)}$

Example 3: In a population of 500 students the numbers of married are 200. Out of 150 students who failed 60 belonged to the married group. It is required to find out whether the attributes marriage and failure are independent, positively associated or negatively associated.

Solution: Let A denote married students.

$\therefore \boldsymbol{\alpha}$ represents unmarried students.

Let B denote failure

$\therefore \beta$ represent success

And N represents total number of students.

On the basis of the information given in question we have:

N = 500, (A) = 200, (β) = 150, (Aβ) = 60.

Applying the proportion method:

Attributes A and b shall be independent if $\frac{(\mathbf{AB})}{(\mathbf{A})} < \frac{(\boldsymbol{\alpha}\mathbf{B})}{(\boldsymbol{\alpha})}$

In other words, if the proportion of married students who failed is the same as the proportion of unmarried students who failed, we can say that the attributes, marriage and failure are independent.

Proportion of married students who failed: i.e., = $\frac{(AB)}{(A)}$ 60/ 200 = 0.3 or 30%

Proportion of unmarried students who failed: i.e. $\frac{(\boldsymbol{\alpha}\mathbf{B})}{(\boldsymbol{\alpha})} = \frac{\mathbf{90}}{\mathbf{300}} = \mathbf{0.3\ or\ 30\%}$

Since the two proportions are the same we conclude that the attributes, marriage and failure are independent.

Example 4: Find whether A and B are independent in the following case by proportion method: (AB) = 256, (αB) = 768, (Aβ) = 48, (αβ) = 144.

Solution: Attributes A and A shall be independent if :

$$\frac{(\mathbf{AB})}{(\mathbf{A})} = \frac{(\boldsymbol{\alpha}\mathbf{B})}{(\boldsymbol{\alpha})}$$

For finding (A) and (α) let us prepare a nine square table:

	A	α	Total
B	256	768	1024
β	48	144	192
Total	304	912	1216

Therefore, by applying the above equation, $= \mathbf{256/304} = \frac{\mathbf{768}}{\mathbf{912}}$

Since left hand side and right hand sides are equal i.e., $= \frac{\mathbf{(AB)}}{\mathbf{(A)}} = \frac{\mathbf{(\alpha B)}}{\mathbf{(\alpha)}}$

Hence attributes A and B are independent.

12.4.1 Limitations

Just like the previous method, in this method also we can only determine the nature of association and not the degree of association.

Example 5: In a certain class it was found that 70% of teh students passed in half yearly examination, 30% students passed half yearly and annual examination, while 28% were such who passed in annual but failed in half yearly examination. Find the percentage of students who:

(a) Passed in annual examination.

(b) Passed in half yearly but failed in annual examination, and

(c) Failed in both the examination.

Solution: Let A denote those passing annual examination

Let B denote those passing the half yearly examination. α and β will represent respectively those who fail in the annual examination and those failing in the half yearly examination.

(i) (A)
(ii) (αB)
(iii) (αβ)
(i) (A) = (AB) + (Aβ) = 30 + 28 = 58%

Hence the percentage of students who passed in the annual examination is 58.

(ii) $(\alpha B) = (B) - (AB) = 70 - 30 = 40\%$

Hence the percentage of students who failed in annual examination nut passed in the half yearly is 40.

(iii) $(\alpha\beta) = (\beta) - (A\beta) = N - (B) - (A\beta) = 100 - 70 - 28 = 2\%$

Hence 2% students failed in both the examinations.

Example 6: A survey was conducted in respect of marital status and success in examinations. Out of 2000 persons who appeared for an examination, 80% of them were boys, and the rest were girls. Among 300 married boys, 140 were successful, 1100 boys were successful among unmarried boys. In respect of 10 married girls 40 were successful, 200 unmarried girls were successful. Construct two separate nine- square tables and determine the association between marital status and passing of examination.

Solution: Let A denote married boys

A would denote unmarried boys

Let B denote those who were successful

B denote those who were unsuccessful

In respect of boys, we are given the following information:

N =1600, (A) = 300, (AB) = 140, (αB) = 1100

We can find out the missing values from the nine square table.

	A	**α**	**Total**
B	140	1100	1240
B	160	200	360

Expectation of (AB) = $\frac{A \times B}{N} = \frac{300 \times 1240}{1600} = 232.5$

Since (AB) the actual observation (140) is less than the expected observation (232.5), the attributes marriage and success in the examination are negatively associated.

We can construct another table in respect of girls. Taking A as married girls, the given information is:

N = 400, (AB) = 40, (A) = 100, (αβ) = 200

We can find out the missing frequencies from the nine square table as below,

	A	**α**	**Total**
B	40	200	240
B	60	100	160
Total	100	300	400

Expectation of (AB) = $\frac{\boldsymbol{A \times B}}{\boldsymbol{N}} = \frac{\mathbf{100 \times 240}}{\mathbf{400}} = 60$

Since (AB) actual observation (40) is less than the expectation (60), the attributes marriage and success in the examination are negatively associated.

Example 7: Out of 3000 unskilled workers of a factory, 2000 come from rural areas and out of 1200 skilled workers, 300 come from rural areas. Determine the association between skill and residence by the method of proportion.

Solution: Let A denote skilled workers

α will denote unskilled workers

Let B denote workers from rural areas

β will denote workers from urban areas

We are given: (A) = 1200, (α) = 3000, (αB) = 2000, (AB) = 300

According to the method of proportions, two attributes A and B are said to be independent if : $= \frac{\mathbf{(AB)}}{\mathbf{(A)}} = \frac{\mathbf{(\alpha B)}}{\mathbf{(\alpha)}}$

In the given case: $\frac{\mathbf{(AB)}}{\mathbf{(A)}} = \frac{\mathbf{300}}{\mathbf{1200}} = 0.25$ and $\frac{\mathbf{(\alpha B)}}{\mathbf{(\alpha)}} = \frac{\mathbf{2000}}{\mathbf{3000}} = \mathbf{0.67}$

Since $\frac{\mathbf{(AB)}}{\mathbf{(A)}}$ is less than $\frac{\mathbf{(\alpha B)}}{\mathbf{(\alpha)}}$ there is negatibe association between skill and residence.

Example 8: Use proportion method to determine the nature of association between A and B:

	B	**β**	**Total**
A	30	50	80
α	20	100	120
Total	50	150	200

Solution: According to the proportion method if :

$\frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)}$ when A is taken as proportion, the attributes A and B are independent.

We have (AB) = 30, (A) = 80, (αB) = 20, (α) = 120

$\frac{(AB)}{(A)} = \frac{30}{80} = 0.375$ and $\frac{(\alpha B)}{(\alpha)} = \frac{20}{120} = 0.167$

Since $\frac{(AB)}{(A)} > \frac{(\alpha B)}{(\alpha)}$ the attributes A and B are positively associated.

12.5 YULE'S COEFFICIENT OF ASSOCIATION

The most popular method of studying association is Yule's coefficient because here we can not only determine the nature of association, i.e., whether the attributes are positively association, negatively associated or independent, but also the degree or extent to which the two attributes are associated. The Yule's coefficient is denoted by the symbol Q and is obtained by applying the formula as given below:

$$\mathbf{Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}}$$

The value Yule's lies between 1. When the value of:

Q = + 1, there is a perfect positive association between the attributes.

Q = - 1, there is a perfect negative association between the attributes or perfect disassociation.

Q = 0, attributes are independent.

The coefficient of association can be used to compare the intensity of association between two attributes with the intensity of association between two other attributes.

Example 9: Find the Yule's coefficient of association for the following data:

N = 1500, (α) = 1117, (B) = 360, (AB) = 35

Solution: By putting the known values of frequencies into 2*2 contingency table and finding the remaining unknown frequencies:

	A	α	Total
B	35	325	360
β	348	792	1140
Total	383	1117	1500

Yule's Coefficient of Assocaition is :

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$= \frac{(35)(792) - (325)(348)}{(35)(792) + (325)(348)}$$

$$= \frac{27720 - 113100}{27720 + 113100}$$

$$= \frac{-85380}{140820} = -0.606$$

Example 10: In a group of 800 students, the number of married is 320. Out of 240 students who failed, 96 belonged to the married group. Find out whether the attributes marriage and failure are independent.

Solution: Let A stand for married students and B for those who failed. We are given N = 800, (A) = 320, (B) = 240, (AB) = 96. By putting the information in nine-square table, we have

	A	α	Total
B	96	144	240
β	224	336	560
Total	320	480	800

Calculating Yule's Coefficient of association:

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} = \frac{(96)(336) - (224)(144)}{(96)(336) + (224)(144)} = \frac{32256 - 32256}{32256 + 32256} = 0$$

Example 11: Investigate the association between eye colour of husbands and eye colour of wives from the data given below:

Husbands with light eyes and wives with light eyes = 309
Husbands with light eyes and wives with non-light eyes = 214
Husbands with non- light eyes and wives with light eyes = 132
Husbands with non-light eyes and wives with non-light eyes= 119

Solution :

Applying Yule's method: $Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} = \frac{(309)(119) - (214)(132)}{(309)(119) + (214)(132)} = \frac{8523}{65019} = 0.131$

Thus, there is a very little association between the eye colour of husband and wife.

12.6 COEFFICIENT OF COLLIGATION

Yule has given another method of studying association between two attributes. This method is called as coefficient of colligation. It is denoted by γ (gamma) and is obtained by applying the following formula:

$$\gamma = \frac{1 - \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}{1 + \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}$$

12.6.1 Yule's Coefficient of Colligation

From Coefficient of colligation, i.e, γ, we also obtain Q as follows:

$Q = \frac{2\gamma}{1+\gamma 2}$ (here, γ2 is gamma square)

It should be noted that though γ and Q serve the same purpose, these coefficients are not directly comparable with each other. Further, in practice Q is more popularly used than γ as a measure of association.

Example 12: A teacher examined 280 students in economics and auditing and found that 160 failed in economics, 140 failed in Auditing and 80 failed in both the subjects. Is there any association between failure in economics and auditing? Use coefficient of colligation.

Solution: Let A denote students who failed in economics and B denote students who failed in auditing.

Putting the given information in a nine-square table, we have

	A	**α**	**Total**
B	80	60	140
β	80	60	140
Total	160	120	280

$$\gamma = \frac{1-\sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}{1+\sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}$$

$$\gamma = \frac{1-\sqrt{\frac{(80)(60)}{(80)(60)}}}{1+\sqrt{\frac{(80)(60)}{(80)(60)}}} = \frac{1-\sqrt{1}}{1+\sqrt{1}} = \frac{1-1}{1+1} = \frac{0}{2} = 0$$

Since the value of gamma is 0, therefore there is no association between failure in economics and auditing.

Example 13: In class-test in which 135 candidates were examined for proficiency in economics and English, it was discovered that 75 students failed in English, 90 failed in economics and 50 failed in both. Find if there is any association between failing in English and economics and also state the magnitude of association by using coefficient of colligation.

Solution: Let A denote students who failed in English and α will denote those who passed in english and B denote students who failed in economics and β will denote students who passed in economics.

Hence the given information can be put in nine-square table and their values determined.

	A	**α**	**Total**
B	50	40	90
β	25	20	45
Total	75	60	135

$$\gamma = \frac{1-\sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}{1+\sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}$$

$$\gamma = \frac{1-\sqrt{\frac{(25)(40)}{(50)(20)}}}{1+\sqrt{\frac{(25)(40)}{(50)(20)}}} = \frac{1-\sqrt{1}}{1+\sqrt{1}} = \frac{1-1}{1+1} = \frac{0}{2} = 0$$

Since the value of gamma is 0, therefore there is no association between failure in economics and English.

12.7 COEFFICIENT OF CONTINGENCY

As we know that the classification of the data can be dichotomous or manifold. If an attribute has only two classes it is said to be dichotomous and if it has many classes, it is called manifold classification. For example the criterion 'location' can be divided into big city and small town. The other characteristic 'nature of occupancy' can be divided into 'owner occupied', 'rented to private parties'. This is dichotomous classification. Now suppose we have N observations classified according to both criteria.

Table 1

Nature of Occupancy	Location		Total
	Big Town	Small Town	
Owner occupied	54	67	121
Rented to parties	107	22	129
Total	161	89	250

Here we have classification by two criteria - one location (two categories) and the other nature of occupancy (two categories). Such a two-way table is called contingency table. The table above is 2X 2 contingency table where both the attributes have two categories each. The table has 2 rows and 2 columns and 2 2 = 4 distinct cells. We also discussed in the previous lesson that the purpose behind the construction of such table is to study the relationship between two attributes i.e. the two attributes or characteristics appear to occur independently of each other or whether there is some association between the two. In the above case our interest lies in ascertaining whether both the attributes i.e. location and nature of occupancy are independent. In practical situations, instead of two classes, an attribute can be classified into number of classes. Such type of classification is called manifold classification.

For example stature can be classified as very tall, tall, medium, short and very short. In the present lesson, we shall discuss manifold classification; related contingency table and methodology to test the intensity of association between two attributes, which are classified into number of classes. The main focus of this lesson would be the computation of Yule's' s coefficient of association, coefficient of colligation, chi-square and the coefficient of contingency, which would be used to measure the degree of association between two attributes.

12.7.1 Contingency Table: Manifold Classification

We have already learnt that if an attribute is divided into more than two parts or groups, we have manifold classification. For example, instead of dividing the universe into two parts-heavy and not heavy, we may sub-divide it in a large number of parts very heavy, heavy, normal, light and very light. This type of sub division can be done for both the attributes of the universe. Thus, attribute A can be divided into a number of groups A_1,

A_2,…, A_r. Similarly, the attribute B can be sub-divided into B_1, B_2,…, B_r. When the observations are classified according to two attributes and arranged in a table, the display is called contingency table. This table can be 3 3, 4 4, etc. In 33 table both of the attributes A and B have three sub-divisions. Similarly, in 4 4 table, each of the attributes A and B is divided into four parts, viz. A_1, A_2, A_3, A_4 and B_1, B_2, B_3, B_4. The number of classes for both the attributes may be different also. If attribute A is divided into 3 parts and B into 4 parts, then we will have 34 contingency table. In the same way, we can have 3 5, 4 3, etc. contingency tables. It should be noted that if one of the attributes has two classes and another has more than two classes, even then the classification is manifold. Thus, we can have 2 3, 2 4, etc. contingency tables. We shall confine our attention to two attributes A and B, where A is sub-divided into r classes, A_1, A_2,…, A_r and B is subdivided into s classes B_1, B_2, …, B_s. Following is the layout of r s contingency table:

Table 2: rxs Contingency Table

	Attribute A					
		A_1	A_2	A_3	A_4	Total
	B_1	(A_1B_1)	(A_2B_1)	(A_3B_1)	(A_4B_1)	(B_1)
Attribute B	B_2	(A_1B_2)	(A_2B_2)	(A_3B_2)	(A_4B_2)	(B_2)
	B_3	(A_1B_3)	(A_2B_3)	(A_3B_3)	(A_4B_3)	(B_3)
	B_4	(A_1B_4)	(A_2B_4)	(A_3B_4)	(A_4B_4)	(B_4)
	Total	(A_1)	(A_2)	(A_3)	(A_4)	N

In the above table sum of columns A_1, A_2, etc. and the sum of rows B_1, B_2, etc. would be first order frequencies and the frequencies of various cells would be second order frequencies. The total of either A_1, A_2, etc. or B_1, B_2, etc. would give grand total N.

In the table:

$(A_1) = (A_1B_1) + (A_1B_2) + \ldots + (A_1B_s)$,

$(A_2) = (A_2B_1) + (A_2B_2) + \ldots + (A_2B_s)$,

etc. Similarly,

$(B_1) = (A_1B_1) + (A_2B_1) + \ldots + (A_rB_1)$,

$(B_2) = (A_1B_2) + (A_2B_2) + \ldots + (ArB_2)$,

etc. And

$N = (A_1) + (A_2) + \ldots + (A_r)$ or

$N = (B_1) + (B_2) + \ldots + (B_s)$

In the following section you will learn how to find degree of association between attributes in r s contingency table.

12.7.2 Chi - Square and Coefficient of Contingency

Association between two attributes in case of manifold classification and the resulting contingency table can be studied as explained below:

We can have manifold classification of the two attributes in which case each of the two attributes are first observed and then each one is classified into two or more subclasses, resulting into what is called as contingency table. The following is an example of 4 × 4 contingency table with two attributes A and B, each one of which has been further classified into four sub-categories.

		Attribute A				
		A_1	A_2	A_3	A_4	Total
	B_1	(A_1B_1)	(A_2B_1)	(A_3B_1)	(A_4B_1)	(B_1)
Attribute B	B_2	(A_1B_2)	(A_2B_2)	(A_3B_2)	(A_4B_2)	(B_2)
	B_3	(A_1B_3)	(A_2B_3)	(A_3B_3)	(A_4B_3)	(B_3)
	B_4	(A_1B_4)	(A_2B_4)	(A_3B_4)	(A_4B_4)	(B_4)
	Total	(A_1)	(A_2)	(A_3)	(A_4)	N

Association can be studied in a contingency table through Yule's coefficient of association as stated above, but for this purpose we have to reduce the contingency table into 2 × 2 table by combining some classes. For instance, if we combine $(A_1) + (A_2)$ to form (A) and $(A_3) + (A_4)$ to form (α) and similarly if we combine $(B_1) + (B_2)$ to form (B) and $(B_3) + (B_4)$ to form (β) in the above contingency table, then we can write the table in the form of a 2 × 2 table as shown in Table Below:

		Attribute		
		A	**α**	**Total**
Attribute	*B* *β*	*(AB)* *(Aβ)*	*(αβ)* *(αβ)*	*B* *(β)*
	Total	**(A)**	**(a)**	**N**

After reducing a contingency table in a two-by-two table through the process of combining some classes, we can work out the association as explained above. But the practice of combining classes is not considered very correct and at times it is inconvenient also. Therefore, Karl Pearson has suggested a measure known as Coefficient of mean square contingency or simply coefficient of contingency for studying association in contingency tables. This can be obtained as under:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

where

C = Coefficient of contingency

χ^2 = Chi-square value which is $= \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

N = number of items.

While finding out the value of c we proceed on the assumption of null hypothesis, i.e., the two attributes are independent and exhibit no association.

For calculation of c we have to determine the value of x^2 (pronounced as chi-square). The steps in calculating the value of x^2 are:

(i) Find the expected or independent frequency for each cell. Thus, for cell (A_1B_1), the expectation if $\frac{A1 \times B1}{N}$.

(ii) Obtain the difference between observed and expected frequencies in each cell, i.e., find (O-E).

(iii) Square (O – E) and divide the figure by E. The expected frequency for each cell.

(iv) Add up the figures obtained in step (iii). This would give the value of x^2.

Thus $x^2 = \Sigma \frac{(O-E)2}{E}$

Once the value of x^2 is obtained it is easy to determine the value of C.

Example 14: The following table gives the association among 1000 advocates between their weights and mental level. Determine the coefficient of contingency between the two.

Table 2

Mental Level	**Weight In Ponds**					
	100-120 B_1	**120-130 B_2**	**130-140 B_3**	**140-150 B_4**	**Above B_4 B_5**	**Total**
Normal A_1	50	102	198	210	240	800
Weak A_2	30	38	72	30	30	200
Total	80	140	270	240	270	1000

Solution: The expected frequency corresponding to Cell (A_1B_1) is

$$E_{11} = \frac{(A1)(B1)}{N} = \frac{(80)(800)}{1000} = 64$$

The expected frequency corresponding to Cell (A_1B_2) is

$$E_{12} = \frac{(A1)(B2)}{N} = \frac{(140)(800)}{1000} = 112$$

The expected frequency corresponding to Cell (A_1B_3) is

$$E_{13} = \frac{(A1)(B3)}{N} = \frac{(800)(270)}{1000} = 216$$

The expected frequency corresponding to Cell (A_1B_4) is

$$E_{14} = \frac{(A1)(B4)}{N} = \frac{(800)(240)}{1000} = 192$$

The expected frequency corresponding to Cell (A_1B_5) is

$E_{15} = \frac{(A1)(B5)}{N} = = 216$

Similarly, the expected frequencies for the cells (A_2B_1), (A_2B_2), (A_2B_3), (A_2B_4), (A_2B_5) are:

$E_{21} = \frac{(A2)(B4)}{N} = \frac{(200)(80)}{1000} = 16$

$E_{22} = \frac{(A2)(B2)}{N} = \frac{(200)(140)}{1000} = 28$

$E_{23} = \frac{(A2)(B3)}{N} = \frac{(200)(270)}{1000} = 54$

$E_{24} = \frac{(A2)(B4)}{N} = \frac{(200)(240)}{1000} = 48$

$E_{25} = \frac{(A2)(B5)}{N} = \frac{(200)(270)}{1000} = 54$

Now we compute the value of $x^2 = \Sigma (O-E)^2 / E$

$= (54-64)^2 / 64 + (102-112)^2 / 112 + (198-216)^2 / 216 + (210-192)^2 / 192 + (240-216)^2 / 216 + (30-16)^2 / 16 + (38-28)^2 / 28 + (72-54)^2 / 54 + (30-48)^2 / 48 + (30-54)^2 / 54$

$= 3.0625 + 0.8929 + 1.500 + 1.6975 + 2.6667 + 12.25 + 3.5714 + 6.00 + 6.75 + 10.6667$

$= 49.0477$

Therefore, coefficient of contingency C is given by:

$C = \sqrt{\frac{x2}{N + x2}}$

$= \sqrt{\frac{49.0477}{1000+49.0477}}$

$= \sqrt{0.04676}$

$= 0.2162$

12.8 SUMMARY

In this lesson, we have discussed the nature as well as the degree of association between two attributes with the help of various methods. In statistics, attributes A and B are associated only if they appear together in greater number of cases than is to be expected if they are independent. In common language association means if A and B occur together a number of times. If there exist no relationship of any kind between two attributes then they are said to be independent otherwise are said to be associated. Attributes A and B are said to be Positively associated if A)(B)/N < (AB) Negatively associated if (A)(B)/ N > (AB) Independent if (A)(B)/N = (AB).

Sometimes only the knowledge of the association (whether positive or negative) or independence between attributes is not sufficient. We are interested in finding the extent or degree of association between attributes, so that we can take decision more precisely and easily. In this regard, we have discussed Yule's coefficient of association, coefficient of colligation and coefficient of contingency in this unit. The value of Yule's coefficient of association lies between –1 to +1. If Q = +1, A and B are perfectly associated. In between –1 to +1, are lying different degrees of association; another important coefficient of association is coefficient of colligation. We have also discussed: Contingency table is a table of joint frequencies of occurrence of two variables classified into categories, X^2 is used for finding association and relationship between attributes, The calculation of X^2 is based on observed frequencies and theoretically determined (expected) frequencies, We have seen that if the observed frequency of each cell is equal to the expected frequency of the respective cell for whole contingency table, then the attributes A and B are completely independent and if they are not same for some of the cells then it means there exists some association between the attributes; The degree or the extent of association between attributes in r s contingency table could be found by computing coefficient of mean square contingency

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

where

C = Coefficient of contingency

χ^2 = Chi-square value which is $= \Sigma \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

N = number of items.

The value of C lies between 0 and 1 but it never attains the value unity. A value near to 1 shows great degree of association between two attributes and a value near 0 shows no association.

12.9 GLOSSARY

- **Association**: In the theory of attributes, the attributes A and B are said to be associated with each other only if the two attributes are not independent, but are related to each other in some way or another.
- **Correlation** - A common statistical analysis, usually abbreviated as r, that measures the degree of relationship between pairs of interval variables in a sample. The range of correlation is from -1.00 to zero to +1.00. Also, a non-cause and effect relationship between two variables.
- **Statistical Significance** - The probability that the difference between the outcomes of the control and experimental group are great enough that it is unlikely due solely to chance. The probability that the null hypothesis can be rejected at a predetermined significance level (0.05 or 0.01).
- **Theory** - A general explanation about a specific behaviour or set of events that is based on known principles and serves to organise related events in a meaningful way. A theory is not as specific as a hypothesis.
- **Comparison method**- By comparison method, the nature of association between two attributes is calculated by comparing observed and expected frequencies.

- **Yule's Method**: Yule's Y, also known as the coefficient of colligation, is a measure of association between two binary variables. The measure was developed by George Yule in 1912, and should not be confused with Yule's coefficient for measuring Skewness based on quartiles.
- **Coefficient of contingency**: The contingency coefficient is a coefficient of association that tells whether two variables or data sets are independent or dependent of each other. It is also known as Pearson's Coefficient (not to be confused with Pearson's Coefficient of Skewness)
- **Contingency Table**: A table showing the distribution of one variable in rows and another in columns, used to study the correlation between the two variables.

12.10 SELF ASSESSMENT QUESTIONS

A. Tick √ the correct answer:-

1. The value of chi-square statistic is always:

   (a) Negative (b) Zero (c) Non-negative (d) One

2. The shape of the chi-square distribution depends upon:

   (a) Parameters (b) Degree of freedom (c) Number of cells (d) Standard deviation

3. For a 3 x 3 contingency table, the numbers of cells in the table are:

   (a) 3 (b) 6 (c) 9 (d) 4

4. If all frequencies of classes are same, the value of Chi-square is:

   (a) Zero (b) One (c) Infinite (d) All of the above

5. The eyes colour of 100 women is:

   (a) Variable (b) Constant (c) Attribute (d) Discrete

6. If two attributes A and B have perfect positive association the value of coefficient of association is equal to:

(a) +1 (b) -1 (c) 0 (d) (r-1)(c-1)

B. True/False:-

1. Yule's coefficient of association is used to study only the nature of association between two attributes. T/F

2. In comparison method, we compare observed and expected frequencies. T/F

3. (AB) indicate actual or observed frequency. T/F

4. Association is the relationship between two or more variables. T/F

12.11 LESSON END EXERCISE

1. With three attributes A, B and C write down:

   i. Number of positive class frequencies.
   ii. Number of ultimate class frequencies.
   iii. Number of all the class frequencies.
   iv. All the class frequencies in symbols.

2. Find the missing frequencies from the following frequencies:

   N= 1000, (A)= 877, (B)= 1086

3. Show whether A and B are independent, positively associated or negatively associated in each of the following cases: Use comparison method.

   (i) N = 1000; (A) = 450; (B) = 600; (AB) = 340
   (ii) (aB) = 383 (á)= 585; (A) = 480; (AB) = 290

4. Find whether A and B are independent, positively associated or negatively associated by using proportion method.

   N = 1000; (A) = 500; (B) = 400; (AB) = 200

5. From the following data prepare 22 table and using comparison method, discuss whether there is any association between literacy and unemployment.

   Illiterate Unemployed 250 persons
   Literate Employed 25 persons
   Illiterate Employed 180 persons
   Total number of persons 500 persons

6. Can vaccination be regarded as a preventive measure for small pox from the data given below:

   'Of 1482 persons in a locality exposed to small pox 368 in all were attacked'
   'Of 1482 persons, 343 had been vaccinated and of these only 35 were attacked'

   Find the Yule's Coefficient of association from the following data: N= 800, (A)= 470, (β) =450 and (AB) = 230

7. When are two attributes said to be independent? From the following data check whether attributes A and B independent or not by using coefficient of colligation method: N = 100, (A) = 50, (B) = 70 and (AB) = 30

8. The male population of certain state is 250 lakhs. The number of literate males is 26 lakhs and the total number of male criminals is 32 thousand. The number of literate male criminal is 3000. Do you find any association between literacy and criminality?

9. Out of total population of 1000 the number of vaccinated persons was 600. In all 200 had an attack of smallpox and out of these 30 were those who were vaccinated. Do you find any association between vaccination and freedom from attack? Use coefficient of colligation.

10. In an area with a total population of 7000 adults, 3400 are males and out of a total 600 graduates, 70 are females. Out of 120 graduate employees, 20 are females. (i) Is there any association between sex and education? (ii) Is there any association between appointment and sex? Use Coefficient of contingency.

11. Find if A and B are independent, positively associated or negatively associated in each of the following cases:

(i) N = 100; (A) = 47; (B) = 62 and (AB) = 32
(ii) (A) = 495; (AB) = 292; (a) = 572 and (ab)= 380
(iii) (AB) = 2560; (αB) = 7680; (Aβ)= 480 (aβ)= 1440

12. Eighty-eight residents of an Indian city, who were interviewed during a sample survey, are classified below according to their smoking and tea drinking habits. Calculate Yule's coefficient of association and comment on its value.

	Smokers	Non-Smokers
Tea Drinkers	40	33
Non-tea Drinkers	3	12

13. Find the association between Literacy and Unemployment from the following figures:

Total Adults: 10,000
Literates: 1,290
Unemployed: 1,390
Literate Unemployed: 820

14. Among the adult population of a certain town 50% are male; 60% are wages earners and 50% are 45 years of age or above. 10% of male are not wage earners and 40% of the male are under 45 years of age. Can we infer anything about the percentage of the population of 45 or over are wage earners.

12.12 SUGGESTED READINGS

- Gupta, S.P.: *Statistical Methods*, Sultan Chand & Sons, New Delhi.
- Gupta, S.C. and V.K. Kapoor : *Fundamentals of Applied Statistics.*

- Anderson, Sweeney and Williams: *Statistics for Business and Economics*-Thompson, New Delhi.
- Hooda, R.P.: *Statistics for Business and Economics,* Macmillan, New Delhi.
- Lawrence B. Morse: *Statistics for Business and Economics,* Harper Collins.

PARTIAL CORRELATION

STRUCTURE

13.1 INTRODUCTION

The correlation and regression coefficients measure the degree and nature of the effect of one variable on another variable. While it is useful to know how one phenomenon is influenced by another? It is also important to know how one phenomenon is affected by several other variables. Its nature, relationship tends to be complex rather than simple.

Also, while learning about correlation, we understood that it indicates relationship between two variables. Indeed, there are correlation coefficients that involve more than two variables. It sounds unusual and you would wonder how to do it? Under what circumstance it can be done? Let me give you two examples. The first is about the correlation between cholesterol level and bank balance for adults. Let us say that we find a positive correlation between these two factors. That is, as the bank balance increases, cholesterol level also increases. But this is not a correct relationship as Cholesterol level can also increase as age increases. Also as age increases, the bank balance may also increase because a person can save from his salary over the years. Thus there is age factor which influences both cholesterol level and bank balance. Suppose we want to know only the correlation between cholesterol and bank balance without the age influence, we could take persons from the same age group and thus control age, but if this is not possible we can statistically control the age factor and thus remove its influence on both cholesterol and bank balance. This if done is called partial correlation.

Further, in multivariate correlation models one is naturally very interested in relationships among the variables. One set of measures useful to this end consists of the coefficients of multiple correlation and the coefficients of partial correlation. All partial correlation coefficients measure the correlation between two variables. Also, in industry and business today, large amounts of data are continuously being generated. This may be data pertaining to a company's annual production, annual sales, turnover profits or some other variable of direct interest to management. The accumulated data may be used to gain information about the system (as for instance what happens to the output of the plant when temperature is reduced to half) or to visually depict the past pattern of our interest in regression is primarily for the first purpose, mainly to extract the main features of the relationships hidden in or implied by the mass of data. There are three types of correlations, viz., simple, partial and multiple. Simple correlation is used when there are only two variables and we may be interested to find out the degree of relationship between these two variables. Simple correlation ignores the effect of all other variables even though these variables might be quite closely related to the independent variable. When there are more than two variables, we may apply either partial correlation or multiple correlations. In this lesson, we will be learning about partial correlation.

13.2 OBJECTIVES

After studying this lesson, you will be able to:

- Describe the concept of correlation, simple and partial correlation.
- Define partial correlation coefficient.
- Derive the partial correlation coefficient formula.

13.3 PARTIAL CORRELATION

It is often important to measure the correlation between a dependent variable and one particular independent variable when all other variables involved are kept constant, i.e., when the effects of all other variables are removed referred to as 'others things being equal'. This can be obtained simply by calculating coefficient of partial correlation.

Therefore, partial correlation analysis measures the strength of the relationship between Y known as dependent variable and one independent variable in such a way that variations in the other independent variables are taken into account.

13.3.1 Meaning of Partial Correlation

Partial correlation coefficient provides a measure of the relationship between the dependent variable and other variables, with the effect of the most of variables eliminated.

We denote the partial correlation coefficient by $r_{12.3}$ between X_1 and X_2, keeping X_3 constant. We find that:

$$r_{12.3} = \frac{r12 - r13\, r23}{\sqrt{(1 - r13^2)} \times \sqrt{(1 - r23^2)}}$$

Similarly,

$$r_{13.2} = \frac{r13 - r12\, r32}{\sqrt{(1 - r12^2)} \times \sqrt{(1 - r23^2)}}$$

Where, $r_{13.2}$ is the coefficient of partial correlation between X_1 and X_3, keeping X_2 constant.

$$r_{23.1} = \frac{r23 - r21\, r31}{\sqrt{(1-r21^2)} \times \sqrt{(1-r31^2)}}$$

Where, $r_{23.12}$ is the coefficient of partial correlation between X_2 and X_3, keeping X_1 constant.

Thus, for three variables, X_1, X_2, and X_3, there will are three coefficients of partial correlation each studying the relationship between two variables when the third is held constant.

The partial correlation coefficient thus helps us to answer questions such as: Is the correlation between, say, X_1 and X_2, merely due to the fact that both are affected by X_3 or is there a no co-variation between X_1 and X_2 over and above the association due to the common influence of X_3? Therefore, in determining the partial correlation coefficient between X_1 and X_2, we attempt to remove the influence of X_3 from each of the two variables so as to ascertain whether net relationship exists between the "Unexplained" residuals that remain.

Also, it should be noted that the value of a partial correlation coefficient is always interpreted via the corresponding coefficient, partial determination, i.e., by squaring the partial correlation coefficient. Thus, if X_1, X_2, and X3 represent sales, advertisement expenditure and price respectively, we get $r^2_{12.3} = 0.912$. This means that more than 91% of the variation in sales which is not associated with price is associated with advertisement expenditure.

13.3.2 Uses of Partial Correlation

Partial correlation analysis is the measurement of relationship between two factors, with the effects of two or more other factors eliminated. If the assumptions of the method are true for a series of data, the power of partial analysis is great. The problem of holding certain variables constant, while the relationship between the others is measured, often presents difficult itself in statistical analysis. Partial correlation is especially useful in the analysis of interrelated series. It is particularly pertinent to uncontrolled experiments of various kinds, in which such interrelationships usually exist. Most economic data fall in this category.

Partial correlation is of great value when used in conjunction with gross and multiple correlation in the analysis of factors affecting variations phenomenon.

Partial analysis, like all correlation, has the advantage that the relationships are expressed

concisely in a few well defined coefficients. Also it is adaptable to small amounts of data and the reliability of the results can be rather easily

tested.

13.3.3 Limitations

1. The usefulness of the partial correlation coefficient is somewhat limited by the following assumption:

   i. The zero order correlation must have linear regression.

   ii. The effects of the independent variables must be additively and not jointly related.

   iii. Because the reliability of partial coefficient decreases as its order increases, the number of observations in gross correlations should be fairly large. Very often the students carry the analysis beyond the limits of the data. Thus, weakness to some extent can be guarded against by the test of reliability.

2. When the above mentioned assumptions have been satisfied, partial correlation possesses the disadvantage of laborious calculations and difficult interpretation even for statisticians.

The interpretation of the partial and multiple correlation results tends to assume that the independent variables have causal effects on dependent variable.

This assumption is sometimes true, but more often untrue in varying degrees.

13.4 COMPUTATION OF PARTIAL CORRELATION

Example 1: On the basis of the following information compute:

(i) $r_{23.1}$ (ii) $r_{13.3}$ (iii) $r_{12.3}$ when $r_{12} = 0.70$, $r_{13} = 0.61$, $r_{23} = 0.40$

Solution: $r_{23.1} = \frac{r23 - r21\, r31}{\sqrt{(1 - r21^2)} \times \sqrt{(1 - r31^2)}}$

Substituting the given values:

$$r_{23.1} = \frac{0.4 - 0.7 \times 0.61}{\sqrt{(1 - 0.7^2)} \times \sqrt{(1 - 0.61^2)}}$$

$$= \frac{0.4 - 0.427}{\sqrt{(0.51)} \times \sqrt{(1 - 0.3721)}} = \frac{-0.027}{0.714 \times 0.792} = \mathbf{-0.048}$$

$$r_{13.2} = \frac{r13 - r12\, r32}{\sqrt{(1 - r12^2)} \times \sqrt{(1 - r23^2)}}$$

Substituting the given values:

$$r_{13.2} = \frac{0.61 - 0.7 \times 0.4}{\sqrt{(1 - 0.7^2)} \times \sqrt{(1 - 0.4^2)}}$$

$$= \frac{0.61 - 0.28}{\sqrt{(1 - 0.49)} \times \sqrt{(1 - 0.16)}} = \frac{0.33}{\sqrt{0.51} \times \sqrt{0.84}} = 0.504$$

$$r_{12.3} = \frac{r12 - r13\, r23}{\sqrt{(1 - r13^2)} \times \sqrt{(1 - r23^2)}}$$

Substituting the given values:

$$r_{12.3} = \frac{0.70 - (0.61 \times 0.4)}{\sqrt{(1 - 0.61^2)} \times \sqrt{(1 - 0.4^2)}}$$

$$= \frac{0.7 - 0.244}{\sqrt{(1 - 0.3721)} \times \sqrt{(1 - 0.16)}} = 0.629$$

Example 2: On the basis of observations made on 39 cotton plants, the total correlation of yield of cotton (X_1), number of balls, i.e., seed vessels (X_2) and height (X_3) are found to be:

$r_{12} = 0.80$, $r_{13} = 0.65$, $r_{23} = 0.70$

Comment on the partial correlation between yield of cotton and the number of balls, eliminating the effect of height.

Solution: We have to find the partial correlation between yield of cotton and the number

of bolts, eliminating the effect of height, i.e, in terms of symbols, we have to calculate $r_{12.3}$.

$$r_{12.3} = \frac{r12 - r13\, r23}{\sqrt{(1 - r13^2)} \times \sqrt{(1 - r23^2)}}$$

Substituting the given values:

$$r_{12.3} = \frac{0.80 - (0.65 \times 0.7)}{\sqrt{(1 - 0.65^2)} \times \sqrt{(1 - 0.7^2)}}$$

$$= \frac{0.8 - 0.455}{\sqrt{(1 - 0.4225)} \times \sqrt{(1 - 0.49)}} = 0.635$$

Example 3: If $r_{12} = 0.86$, $r_{13} = 0.65$, and $r_{23} = 0.72$, find the partial correlation coefficient by keeping third variable constant.

Solution: Here in this question we have to find the value of $r_{12.3}$.

$$r_{12.3} = \frac{r12 - r13\, r23}{\sqrt{(1 - r13^2)} \times \sqrt{(1 - r23^2)}}$$

Substituting the given values:

$$r_{12.3} = \frac{0.86 - (0.65 \times 0.72)}{\sqrt{(1 - 0.65^2)} \times \sqrt{(1 - 0.72^2)}}$$

$$= \frac{0.8 - 0.468}{\sqrt{(.5775)} \times \sqrt{(.4816)}} = 0.743$$

13.4.1 Order of Partial Correlation Coefficient

The order of partial correlation coefficient depends upon the number of variables are help to be constant. If no variable is held to be constant, it is known as zero order correlation or simple correlation coefficient. If one variable is held to be constant it is the case of first order partial correlation and in case two variables are kept as constant, it is known as second order partial correlation coefficient, and so on. Thus, partial coefficients such as $r_{12.3,}$ $r_{13.2}$ are often referred to as first order coefficients, since one variable has been held constant. $r_{12,}$ r_{13} etc. are the examples of simple correlation coefficient.

13.5 SUMMARY

In this lesson, we have discussed about the partial correlation analysis. Partial correlation analysis is aimed at finding correlation between two variables after removing the effects of other variables. The central concept in partial correlation analysis is the partial correlation coefficient $r_{xy.z}$ between variables x and y, adjusted for a third variable z . In probability theory and statistics, partial correlation measures the degree of association between two random variables, with the effect of a set of controlling random variables removed. Like the correlation coefficient, the partial correlation coefficient takes on a value in the range from –1 to 1

13.6 GLOSSARY

- **Correlation:** A common statistical analysis, usually abbreviated as r, that measures the degree of relationship between pairs of interval variables in a sample. The range of correlation is from -1.00 to zero to +1.00. Also, a non-cause and effect relationship between two variables.
- **Partial Correlation:** Partial correlation measures the degree of relationship between two random variables, with the effect of a set of other independent variables is held to be constant.
- **Positive Correlation:** A positive correlation is a relationship between two variables that move in tandem-that is, in the same direction. A positive correlation exists when one variable decreases as the other variable decreases, or one variable increases while the other increases.
- **Negative Correlation:** Negative correlation is a relationship between two variables in which one variable increases as the other decreases, and vice versa.
- **First Order partial correlation:** When in partial correlation only one variable is held to be constant, it is known as first order partial correlation. For e.g., $r_{12.3}$.
- **Second Order partial correlation:** When in partial correlation two variables are held to be constant, it is known as second order partial correlation. For e.g., $r_{12.34}$.

13.7 SELF ASSESSMENT QUESTIONS

A. Multiple Choice Questions:

1. In partial correlation coefficient all other variables are held to be:

   a. Deviated b. Not Fixed
   c. Residual Mean d. Constant

2. Correlation analysis is a .....

   a. Univariate analysis b. Bivariate analysis
   c. Multivariate analysis d. Both b and c

3. When the values of two variables move in the same direction, correlation is said to be .....

   a. Linear b. Non-linear
   c. Positive d. Negative

4. When the values of two variables move in the opposite directions, correlation is said to be .....

   a. Linear b. Non-linear
   c. Positive d. Negative

5. When the amount of change in one variable leads to a constant ratio of change in the other variable, then correlation is said to be .....

   a. Linear b. Non-linear
   c. Positive d. Negative

13.8 LESSON END EXERCISE

1. Find the partial correlation coefficient between first and second variable from the following data: $r_{12} = 0.40$; $r_{23} = 0.60$; $r_{13} = 0.70$.

2. What is partial correlation? Under what circumstances it is to be preferred to the total correlation.

3. Given $r_{12} = 0.50$; $r_{23} = 0.1$; $r_{13} = 0.40$, find $r_{12.3}$ and $r_{23.1}$.

4. What is meant by multi-collinearity

5. Distinguish between simple and partial correlation coefficients.

13.9 SUGGESTED READINGS

- Gupta, S.C. and V.K. Kapoor : *Fundamentals of Applied Statistics.*
- Hooda, R.P.: *Statistics for Business and Economics,* Macmillan, New Delhi.
- Hien, L.W: *Quantitative Approach to Managerial Decisions,* Prentice Hall, New Jersy, India, Delhi.
- Lawrence B. Morse: *Statistics for Business and Economics,* Harper Collins.
- Mc Clave, Benson and Sincich: *Statistics for Business and Economics,* Eleventh Edition, Prentice Hall Publication.

MULTIPLE CORRELATION

STRUCTURE

14.1 INTRODUCTION

Multiple correlation analysis is an extension of simple and partial correlation analysis, as described in lesson 13, to situations involving two or more independent variables and their degree of association with the dependent variable. As is the case for partial correlation

analysis described in lesson 13, the dependent variable is designated by Y, while the several independent variables are designated sequentially beginning with X_1. (Note: In some textbooks and computer software the dependent variable is designated by X_1, in which case the independent variables are designated sequentially beginning with X_2.)

The coefficient of multiple correlations, which is designated by the uppercase $R_{1.23}$: for the case of two independent variables, is indicative of the extent of the relationship between two independent variables taken as a group and the dependent variable. It is possible that one of the independent variables alone could have a positive relationship with the dependent variable while the other independent variable could have a negative relationship with the dependent variable. All R values are reported as absolute values, without an arithmetic sign. In this lesson, we will be discussing about concept of multiple correlation, advantages and disadvantages, computational procedure for finding the value of coefficients of multiple correlation.

14.2 OBJECTIVES

After studying this lesson, you will be able to:

- Describe and explain concept of multiple correlation.
- Derive the multiple correlation coefficient formula.
- Explain the properties of multiple correlation coefficients.
- Compute and interpret multiple correlations.
- Know the advantages and disadvantages of multiple correlation

14.3 MULTIPLE CORRELATION

If information on two variables like height and weight, income and expenditure, demand and supply, etc. are available and we want to study the linear relationship between two variables, correlation coefficient serves our purpose which provides the strength or degree of linear relationship with direction whether it is positive or negative. But in biological, physical and social sciences, often data are available on more than two variables and value of one variable seems to be influenced by two or more variables. For example, crimes in a city may be influenced by illiteracy, increased population and unemployment in the city,

etc. The production of a crop may depend upon amount of rainfall, quality of seeds, quantity of fertilizers used and method of irrigation, etc. Similarly, performance of students in university exam may depend upon his/her IQ, mother's qualification, father's qualification, parent's income, number of hours of studies, etc. Whenever we are interested in studying the joint effect of two or more variables on a single variable, multiple correlations gives the solution of our problem. In fact, multiple correlation is the study of combined influence of two or more variables on a single variable. Suppose, X_1 , X_2 and X_3 are three variables having observations on N individuals or units. Then multiple correlation coefficient of X_1 on X_2 and X_3 is the simple correlation coefficient between X_1 and the joint effect of X_2 and X_3. It can also be defined as the correlation between X_1 and its estimate based on X_2 and X_3.

14.3.1 Concept

In multiple correlations we are dealing with the situations that involve three or more variables. For example, we may consider the association between the yield of wheat per acre and both the amount of rainfall and the average daily temperature. We are trying to make estimates of the value of one of these variables based on the values of all the others variables. The variables whose value we are trying to estimate is called the dependent variable and the other variables on which our estimates are based are known as independent variables. The statistician himself chooses which variable is to be dependent and which variables are to be independent. It is merely a question of problem being studied.

The coefficient of multiple linear correlations is represented by R_1 and it is common to add subscripts designating the variables involved. Thus, $R_{1.234}$ would represent the coefficient of multiple linear correlations between X_1 on the one hand and X_2, X_3, and X_4 on the other. The subscript of the dependent variable is always left to the left of the point.

The coefficient of multiple correlations can be expressed in terms of r_{12}, r_{13}, and r_{23} as follows:

$$R_{1.23} = \frac{\sqrt{r12^2+r13^2-2r12.r13.r23}}{\sqrt{(1-r23^2)}}$$

$$R_{3.12} = \frac{\sqrt{r31^2+r32^2-2r12.r13.r23}}{\sqrt{(1-r12^2)}}$$

$$R_{2.31} = \frac{\sqrt{r23^2 + r21^2 - 2r12.r13.r23}}{\sqrt{(1 - r31^2)}}$$

It must noted that $R_{1.23}$ is the same as $R_{1.32}$.

By squaring $R_{1.23,}$ we obtain the coefficient of multiple determination.

14.3.1.1 Properties of Multiple Correlation Coefficients

The following are some of the properties of multiple correlation coefficients:

1. Multiple correlation coefficient is the degree of association between observed value of the dependent variable and its estimate obtained by multiple regression,
2. Multiple Correlation coefficient lies between 0 and 1 and is always positive in sign.
3. If multiple correlation coefficient is 1, then association is perfect and multiple regression equation may said to be perfect prediction formula,
4. If multiple correlation coefficient is 0, dependent variable is uncorrelated with other independent variables. From this, it can be concluded that multiple regression equation fails to predict the value of dependent variable when values of independent variables are known,
5. Multiple correlation coefficient is always greater or equal than any total correlation coefficient. If $R_{1.23}$ is the multiple correlation coefficient than $R_{1.23} \geq r_{12}$ or r_{13} or r_{23}, and
6. Multiple correlation coefficient is obtained by method of least squares would always be greater than the multiple correlation coefficient obtained by any other method.

14.3.2 Advantages

The multiple correlation serves the following purposes:

1. It serves as a measure of the degree of association between one variable taken as dependent variable and a group of other variables taken as the independent variables.
2. It also serves as a measure of goodness of fit of the calculated plane of regression and consequently as a measure of the general degree of accuracy of estimates made

by reference to equation for the plane of regression.

14.3.3 Disadvantages

1. Multiple correlation analysis is based on the assumption that the relationship between the variables is linear. In other words, the rate of change in one variable in terms of others is assumed to be constant for all values. In practice most relationships are not linear but follow some other pattern. This limits somewhat the use of multiple correlation analysis. The linear regression coefficients are not accurately descriptive of curvilinear data.

2. Another limitation is the assumption that effects of independent variables on the dependent variables are separate, distinct and additive. When the effects of variables are additive, a given change in one has the same effect on the dependent variable regardless of the sizes of the other two independent variables.

3. Linear multiple correlation involves a great deal of work relative to the results frequently obtained. When the results are obtained, only a few students are able to interpret them. The misuse of correlation results has probability led to more doubt on the method than is justified. However, this lack of understanding and resulting misuse are due to the complexity of the method.

14.4 Computation of Coefficients of Multiple Correlations

Example 1: The following zero-order correlation coefficients are given, $r_{12} = 0.98$, $r_{13} = 0.44$ and $r_{23} = 0.54$. Calculate multiple correlation coefficient treating first variable as dependent and second and third variables as independent.

Solution: We have to calculate the multiple correlation coefficient treating first variable as dependent and second and third variables as independent, i.e., we have to find $R_{1.23}$.

$$R_{1.23} = \frac{\sqrt{(r12^2 + r13^2 - 2r12r13r23)}}{\sqrt{(1 - r23^2)}}$$

Substituting the given values:

$$R_{1.23} = \frac{\sqrt{(.98)^2 + (.44)^2 - 2\times0.98\times0.44\times0.54)}}{\sqrt{(1 - .54^2)}}$$

$$= \frac{\sqrt{0.9604 + 0.1936 - 0.4657}}{\sqrt{0.7084}} = 0.986$$

Example 2: If $r_{12} = 0.9$, $r_{13} = 0.75$ and $r_{23} = 0.7$, Find $R_{1.23}$.

Solution: $R_{1.23} = \frac{\sqrt{(r12^2 + r13^2 - 2r12r13r23)}}{\sqrt{(1 - r23^2)}}$

Substituting the given values

$$R_{1.23} = \frac{\sqrt{(0.9)^2 + (0.75)^2 - 2\times0.9\times0.75\times0.7)}}{\sqrt{(1 - (0.7)^2}}$$

$$= \frac{\sqrt{0.81 + 0.5625 - 0.945}}{\sqrt{0.51}} = \frac{\sqrt{0.4275}}{\sqrt{0.51}} = 0.961$$

Example 3: If $r_{12} = 0.77$, $r_{13} = 0.72$ and $r_{23} = 0.52$, Find $R_{3.12}$.

Solution: $R_{3.12} = \frac{\sqrt{(r31^2 + r32^2 - 2r12r13r23)}}{\sqrt{(1 - r12^2)}}$

Substituting the given values

$$R_{3.12} = \frac{\sqrt{(0.72)^2 + (0.52)^2 - 2\times0.77\times0.72\times0.52)}}{\sqrt{(1 - (0.77)^2}}$$

$$= \frac{\sqrt{0.51 + 0.27 - 0.5766}}{\sqrt{0.407}} = \frac{\sqrt{0.2034}}{\sqrt{0.407}} = 0.\ 7068.$$

14.5 SUMMARY

In this lesson we have learned about the interesting procedures of computing the multiple correlations. Especially, when we are interested in controlling for one or more variable. The multiple correlations provide us with an opportunity to calculate correlations between a variable and a linear combination of other variable. You practice them by solving some of the example given below, and you will understand the use of them. Therefore, in this lesson we have discussed:

1. The multiple correlations, which is the study of joint effect of a group of two or more variables on a single variable which is not included in that group.
2. The estimate obtained by regression equation of that variable on other variables, 3. Limit of multiple correlation coefficient, which lies between 0 and +1.
4. The numerical problems of multiple correlation coefficients.
5. The properties of multiple correlation coefficients.

14.6 GLOSSARY

- **Multiple Correlations:** A statistical technique that predicts values of one variable on the basis of two or more other variables.
- **Coefficient:** A numerical or constant quantity placed before and multiplying the variable in an algebraic expression.
- **Dependent Variable:** The dependent variable is the variable that is being measured or tested in an experiment.

14.7 SELF ASSESSMENT QUESTIONS

Fill in the Blanks.

1. By squaring $R_{1.23}$ we obtain the coefficient of.....
2. The coefficient of multiple correlation lies between.....
3. The dependent variable is always denoted by.....
4. The closer the coefficient of multiple correlation is to 1, the.....the relationship between variables.

14.8 LESSON END EXERCISE

1. Distinguish between partial and multiple correlations by giving suitable example.

2. Write down the expressions for $r_{12.4}$ and $R_{1.23}$ in terms of r_{12}, r_{23}, and r_{31}. What are the limits between which $r_{12.3}$ and $r_{1.23}$ must lie? Also state the relation between $r_{12.3}$, $b_{1.23}$ and $b_{21.3}$.

3. Find the multiple correlation coefficient $R_{1.23}$, when, $r_{12} = 0.90$; $r_{13} = 0.75$, $r_{23} = 0.70$.

4. Explain the properties of coefficient of multiple correlation.

14.9 SUGGESTED READINGS

- Gupta, S.P.: *Statistical Methods,* Sultan Chand & Sons, New Delhi.
- Gupta, S.C. and V.K. Kapoor: *Fundamentals of Applied Statistics*.
- Anderson, Sweeney and Williams: *Statistics for Business and Economics*-Thompson, New Delhi.
- Lawrence B. Morse: *Statistics for Business and Economics,* Harper Collins.
- Mc Clave, Benson and Sincich: *Statistics for Business and Economics,* Eleventh Edition, Prentice Hall Publication.

MULTIPLE REGRESSION

STRUCTURE

15.1 INTRODUCTION

The term 'regression' was first used by a British Biometrician Sir Francis Gallon in the later part of nineteenth century, in connection with the height of parents and their off springs. He found that the offspring of tall or short parents tend to regress to the average height. In

other words though tall fathers do tend to have taIl sons, yet the average height and the average heights of short fathers is less is "stepping back". But now a days the term 'regression' stands for some sort of functional relationship between two or more related variables. After having established the fact of two variables are closely related we may be interested in estimating the value of the one variable given the value of other variable.

Linear regression is a common statistical data analysis technique. It is used to determine the extent to which there is a linear relationship between a dependent variable and one or more independent variables. There are two types of linear regression, simple linear regression and multiple linear regressions. In simple linear regression a single independent variable is used to predict the value of a dependent variable. In multiple linear regressions two or more independent variables are used to predict the value of a dependent variable. The difference between the two is the number of independent variables. In both cases there is only a single dependent variable. Simple linear regression establishes the relationship between two variables using a straight line. Linear regression or simple regression attempts to draw a line that comes closest to the data by finding the slope and intercept that define the line and minimise the regression errors. Many data relationships do not follow a straight line, so statisticians use nonlinear regression instead. The two are similar in that both track a particular response from a set of variables graphically. But nonlinear models are more complicated than linear models because the function is created through a series of assumptions that may stem from trial-and-error. In simple regression, two variables exists, one is dependent and the other one is independent. For example, if X and Y are two variables, we shall have two simple regression equations, i.e., regression equation of X on Y by taking X as dependent and Y as independent and regression equation of Y on X by taking Y as dependent variable. One thing must noted here is that, the regression lines cut each other at the point of average of X and Y, i.e., if from the point where both the regression lines cut each other a perpendicular is drawn on the X-axis, we will get the mean value of X and if from that point a horizontal line is drawn on the Y-axis, we will get the mean value of Y.

As we know the correlation coefficient and regression analysis measure the degree of dependence and nature of the effect of one variable on the other variable. If we deal the combined effect of a group of variables (at least two variables) upon a single variable

which is not included in that group, our study is that of multiple regression and multiple correlation. If however, we wish to examine the effect of one variable on another after eliminating the effect of remaining variables; our study is of partial correlation and partial regression. Therefore, in this lesson we will be discussing about multiple regression equations.

15.2 OBJECTIVES

After reading this lesson, you will be able to:

- Understand the role of regression in establishing mathematical relationship between dependent and independent variables.
- Find and interpret the least-squares multiple regression equation.
- Forecast the value of dependent variable for the given value of independent variable
- Calculate and interpret the coefficient of multiple determination (R^2).

15.3 CONCEPT OF MULTIPLE REGRESSION

It is rare that a dependent variable is explained by only one variable. In this case, an analyst uses multiple regression, which attempts to explain a dependent variable using more than one independent variable. Multiple regressions can be linear and nonlinear. Multiple regressions are based on the assumption that there is a linear relationship between both the dependent and independent variables. It also assumes no major correlation between the independent variables. Multiple regression is an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable). The variables we are using to predict the value of the dependent variable are called the independent variables (or sometimes, the predictor, explanatory or regressor variables).

For example, you could use multiple regression to understand whether exam performance can be predicted based on revision time, test anxiety, lecture attendance and gender. Alternately, you could use multiple regression to understand whether daily cigarette consumption can be predicted based on smoking duration, age when smoking, smoker type, income and gender started. Also, multiple regressions allow you to determine the

overall fit (variance explained) of the model and the relative contribution of each of the predictors to the total variance explained. For example, you might want to know how much of the variation in exam performance can be explained by revision time, test anxiety, lecture attendance and gender "as a whole", but also the "relative contribution" of each independent variable in explaining the variance

15.3.1 Objectives of Multiple Regression Analysis

The following are the three main objectives of multiple regression analysis:

1. To derive an equation which provide estimates of the dependent variable from the values of two or more independent variables.
2. To obtain a measure of the error involved in using regression equation as a basis for estimation.
3. To obtain a measure of the proportion of variance in the dependent variable accounted for or explained by the independent variables.

15.3.2 Relevance of Multiple Regression

The significance or relevance of multiple regression analysis can be reflected from the following points:

1. The first is the ability to determine the relative influence of one or more predictor variables to the criterion value. The real estate agent could find that the size of the homes and the number of bedrooms have a strong correlation to the price of a home, while the proximity to schools has no correlation at all, or even a negative correlation if it is primarily a retirement community.
2. The second advantage is the ability to identify outliers, or anomalies. For example, while reviewing the data related to management salaries, the human resources manager could find that the number of hours worked, the department size and its budget all had a strong correlation to salaries, while seniority did not. Alternatively, it could be that all of the listed predictor values were correlated to each of the salaries being examined, except for one manager who was being overpaid compared to the others.

3. Multiple regression offers a degree of flexibility, the interactions between variables can be incorporated and those can be eliminated which provide least insight into the model.

4. Multiple regression allows a statistician to explore the effect of more than one variable on the outcome he wants to study.

5. It uses data very efficiently and can make useful predictions with small data sets.

6. It can allow to construct easy equations of various parameters which can help give predictions and can accordingly be optimized.

15.3.3 Assumptions Multiple Regression Analysis

When you choose to analyse your data using multiple regression, part of the process involves checking to make sure that the data you want to analyse can actually be analysed using multiple regression. You need to do this because it is only appropriate to use multiple regression if your data assumptions that are required for multiple regression to give you a valid results. These assumptions are:

- Dependent variable should be measured on a continuous scale (i.e., it is either an interval or ratio variable). Examples of variables that meet this criterion include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg) etc.
- Two or more independent variables, which can be either continuous (i.e., an interval or ratio variable) or categorical (i.e., an ordinal or nominal variable). Examples of nominal variables include gender (e.g., 2 groups: male and female), ethnicity (e.g., 3 groups: Caucasian, African American and Hispanic), physical activity level (e.g., 4 groups: sedentary, low, moderate and high), profession (e.g., 5 groups: surgeon, doctor, nurse, dentist, therapist), and so forth.
- Independence of observations i.e., independence of residuals.
- There needs to be a linear relationship between the dependent variable and each of independent variables.
- Data needs to show homoscedasticity, which is where the variances along the line of best fit remain similar as we move along the line.

- Data must be free from multi collinearity, which occurs when you have two or more independent variables that are highly correlated with each other.
- The residuals (errors) are approximately normally distributed.

15.4 COMPUTATION OF MULTIPLE REGRESSION EQUATIONS

The multiple regression equation describes the average relationship between these variables and this relationship is used to predict or control the dependent variable.

A regression equation is an equation for estimating a dependent variable, Say X_1 from the independent variables say X_2, X_3, ...and is called a regression equation of X_1 on X_2 and X_3,....In functional notation, this is sometimes written briefly as $X_1 = F(X_2, X_3, X_4....)$ read as "X_1 is a function of X_2, X_3 and so on".

In case of three variables, the regression equation of X_1 on X_2 and X_3 has the form:

$$X_{1.23} = a_{1.23} + b_{12.3} X_2 + b_{13.2} X_3$$

$X_{1.23}$ is the computed value of dependent variable and X_2, X_3 are the independent variables.

The constant $a_{1.23}$ is the intercept made by the regression plane. It gives the value of the dependent variable when all the independent variables assume a value equal to zero. $b_{12.3}$ and $b_{13.2}$ are called partial regression coefficients or the net regression coefficients. $b_{12.3}$ measures the amount by which a unit change in X_2 is expected to affect X_1 when X_3 is held constant and $b_{13.2}$ measures the amount of change in X_1 per unit change in X_3 when X_2 is held constant.

Due to the fact the X_1 varies partially because of variation in X_2 and partially because of variation in X_3, we call $b_{12.3}$ and $b_{13.2}$ the partial regression coefficients of X_1 on X_2 keeping X_3 constant and of X_1 on X_3 keeping X_2 constant.

Case I: Multiple Regression Equations When Deviation are Taken From Actual Mean

When the deviations are already taken from their actual mean, the value of constant or intercept is zero. In this case the multiple regression equation of X_1 on X_2 and X_3 is :

$$x_{1.23} = b_{12.3} x_2 + b_{13.2} x_3$$

Where, $x_1 = (X_1 - X_1)$, $x_2 = (X_2 - X_2)$, $x_3 = (X_3 - X_3)$,

$$b_{12.3} = \frac{\sigma 1}{\sigma 2} \times \left(\frac{r12 - r13\, r23}{(1 - r23^2)} \right)$$

$$b_{13.2} = \frac{\sigma 1}{\sigma 3} \times \left(\frac{r13 - r12\, r23}{(1 - r23^2)} \right)$$

Case II: When it is asked in the question to find least square regression equation or when S_1, S_2, S_3 are given in the question along with means of the variables.

1. In this case the regression equation of X_1 on X_2 and X_3 can be expressed as:

$$(X_1 - X_1) = b_{12.3}(X_2 - X_2) + b_{13.2}(X_3 - X_3)$$

Where, $b_{12.3} = \frac{S1}{S2} \times \left(\frac{r12 - r13\, r23}{(1 - r23^2)} \right)$

$$b_{13.2} = \frac{S1}{s3} \times \left(\frac{r13 - r12\, r23}{(1 - r23^2)} \right)$$

2. In this case the regression equation of X_2 on X_1 and X_3 can be expressed as:

$$(X_2 - X_2) = b_{21.3}(X_1 - X_1) + b_{23.1}(X_3 - X_3)$$

Where, $b_{21.3} = \frac{S2}{S1} \times \left(\frac{r21 - r23\, r13}{(1 - r13^2)} \right.$

$$b_{23.1} = \frac{S2}{s3} \times \left(\frac{r23 - r21\, r31}{(1 - r13^2)} \right)$$

3. In this case the regression equation of X_3 on X_1 and X_2 can be expressed as:

$$(X_3 - X_3) = b_{31.2}(X_1 - X_1) + b_{32.1}(X_2 - X_2)$$

Where, $b_{31.2} = \frac{S3}{S1} \times \left(\frac{r13 - r32\, r21}{(1 - r21^2)} \right)$

$$b_{32.1} = \frac{S3}{s2} \times \left(\frac{r32 - r12\, r31}{(1 - r21^2)} \right)$$

The above method of obtaining regression equations is much simpler as compared to one where simultaneously several normal equations are to be solved. For calculating regression equation for three variables when the above procedure is used, we need the following: S_1, S_2, S_3, r_{12}, r_{23}, r_{13}, X_1, X_2, X_3

Example 1: Given the following, determine the regression equation of:

(1) x_1 on x_2 and x_3

(2) x_2 on x_1 and x_3

$r_{12} = 0.8,\ r_{13} = 0.6,\ r_{23} = 0.5$

$ó_1 = 10,\ ó_2 = 8\ ,\ ó_3 = 5$

Solution: (1) Regression equation of x_1 on x_2 and x_3 is given by:

$$x_1 = b_{12.3}\, x_2 + b_{13.2}\, x_3$$

If the variates x_1, x_2, and x_3 are measured as deviations from their respective means, the value of 'a' i.e., intercept will be zero.

$$b_{12.3} = \frac{\sigma 1}{\sigma 2} \times \left(\frac{r12 - r13\ r23}{(1 - r23^2)} \right)$$

$$= \frac{10}{8} \times \left(\frac{0.8 - 0.6 \times 0.5}{(1 - 0.5^2)} \right) = 0.833$$

$$b_{13.2} = \frac{\sigma 1}{\sigma 3} \times \left(\frac{r13 - r12\ r23}{(1 - r23^2)} \right)$$

$$= \frac{10}{5} \times \left(\frac{0.6 - 0.8 \times 0.5}{(1 - 0.5^2)} \right) = 0.533$$

$\therefore$ The required regression equation is:

$$x_1 = 0.833x_2 + 0.533x_3$$

(2) Regression equation of x_2 on x_1 and x_3

$x_2 = b_{12.3}\, x_1 + b_{23.1}\, x_3$

$$b_{21.3} = \frac{\sigma 2}{\sigma 1} \times \left(\frac{r21 - r13\ r23}{(1 - r13^2)} \right)$$

$$= \frac{8}{10} \times \left(\frac{0.8 - 0.5 \times 0.6}{(1 - 0.6^2)} \right) = 0.625$$

$$b_{23.1} = \frac{\sigma 2}{\sigma 3} \times \left(\frac{r23 - r21\ r31}{(1 - r13^2)} \right)$$

$$= \frac{8}{5} \times \left(\frac{0.5 - 0.8 \times 0.6}{(1 - 0.6^2)} \right) = 0.05$$

Thus, the required equation is: $x_2 = 0.625x_1 + 0.05x_3$.

15.5 SHORTCOMINGS OF MULTIPLE REGRESSION ANALYSIS

Any disadvantage of using a multiple regression model usually comes down to the data being used. Two examples of this are using incomplete data and falsely concluding that a correlation is a causation.

When reviewing the price of homes, for example, suppose the real estate agent looked at only 10 homes, seven of which were purchased by young parents. In this case, the relationship between the proximity of schools may lead her to believe that this had an effect on the sale price for all homes being sold in the community. This illustrates the pitfalls of incomplete data. Had she used a larger sample, she could have found that, out of 100 homes sold, only ten percent of the home values were related to a school's proximity. If she had used the buyers' ages as a predictor value, she could have found that younger buyers were willing to pay more for homes in the community than older buyers.

In the example of management salaries, suppose there was one outlier who had a smaller budget, less seniority and with fewer personnel to manage but was making more than anyone else. The HR manager could look at the data and conclude that this individual is being overpaid. However, this conclusion would be erroneous if he didn't take into account that this manager was in charge of the company's website and had a highly coveted skillset in network security

15.6 SUMMARY

This lesson is designed to discus about multiple regression analysis. Multiple regression is a statistical technique that can be used to analyse the relationship between a single dependent variable and several independent variables. The objective of multiple regression analysis is to use the independent variables whose values are known to predict the value of the single dependent value. This approach can be applied to analyze multivariate time series data when one of the variables is dependent on a set of other variables. We can model the dependent variable Y on the set of independent variables. Regression coefficients are estimates of the unknown population parameters and describe the relationship between a predictor variable and the response.

15.7 GLOSSARY

- **Regression:** Regression is a statistical measurement used in finance, investing and other disciplines that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables).
- **Multiple Regressions:** Multiple regression is an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable).
- **Regression coefficients:** Regression coefficients are estimates of the unknown population parameters and describe the relationship between a predictor variable and the response. In linear regression, coefficients are the values that multiply the predictor values.
- **Coefficient of Determination:** The coefficient of determination, denoted R^2 or r^2 and pronounced "R squared", is the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

15.8 SELF ASSESSMENT QUESTIONS

Fill in the blanks:-

1. _____measures the strength of the linear relationship between the dependent and the independent variable.
2. The correlation coefficient may assume any value between_____and_____.
3. While the range for r^2 is between 0 and 1, the range for r is between_____.

True/False:-

1. The dependent variable is the variable that is being described, predicted, or controlled. T/F
2. A simple linear regression model is an equation that describes the straight-line relationship between a dependent variable and an independent variable. T/F
3. If r = -1, then we can conclude that there is a perfect relationship between X and Y. T/F

15.9 LESSON END EXERCISE

1. Given the following data, find the regression equation of X_1 on X_2 and X_3:

   X_1: 12 22 32 28
   X_2: 6 12 16 22
   X_3: 4 6 12 18
   Also, predict the value of X_1 when $X_2 = 5$ and $X_3 = 7$.

   _____
   _____
   _____

2. What is the difference in interpretation of b weights in simple regression vs. multiple regression?

   _____
   _____
   _____

3. Write a regression equation with beta weights in it.

4. In a trivariate distribution:

$ó_1 = 3, ó_2 = 4 , ó_3 = 5$

$r_{12} = .4 , r_{23 = .6}, r_{13} = .7$

Determine the regression equation of X1 on X2 and X3 if the variates are measured from their means.

15.10 SUGGESTED READING

- Gupta, S.P.: *Statistical Methods,* Sultan Chand & Sons, New Delhi.
- Gupta, S.C. and V.K. Kapoor: *Fundamentals of Applied Statistics.*
- Levin, Richard and David S Rubin: *Statistics for Management,* Prentice Hall, Delhi.
- Levin and Brevson: *Business Statistics,* Pearson Education, New Delhi.
- Lawrence B. Morse: *Statistics for Business and Economics,* Harper Collins.

HYPOTHESIS TESTING
CONCEPT, TYPES, PROCEDURE OF TESTING HYPOTHESIS, TYPE I AND TYPE II ERRORS

STRUCTURE

16.1 INTRODUCTION

We have discussed the various methods of selecting a sample from a population. After selecting the sample, one would naturally be interested in drawing inferences about the population based on our observations made on the sample units. This could mean estimating the value of population parameter, testing a statistical statement means hypothesis about a parent population, comparing two or more populations and many other inferences. In this lesson we shall discuss the concept of hypothesis testing, types of hypothesis, types of errors in testing hypothesis, various steps used in testing hypothesis.

16.2 OBJECTIVES

After reading this lesson, you will be able to:

- Understand the concept and types of hypotheses.
- Define the types of errors associated with hypothesis testing.
- Know the various steps involved in testing of hypothesis.

16.3 CONCEPT OF HYPOTHESIS

Hypothesis is usually considered as the principal instrument in research. Its main function is to suggest new experiments and observations. In fact, many experiments are carried out with the deliberate object of testing hypotheses. Decision-makers often face situations wherein they are interested in testing hypotheses on the basis of available information and then take decisions on the basis of such testing. In social science, where direct knowledge of population parameter(s) is rare, hypothesis testing is the often used strategy for deciding whether a sample data offer such support for a hypothesis that generalisation can be made. Thus hypothesis testing enables us to make probability statements about population parameter(s). The hypothesis may not be proved absolutely, but in practice it is accepted if it has withstood a critical testing. Before we explain how hypotheses are tested through different tests meant for the purpose, it will be appropriate to explain clearly the meaning of a hypothesis and the related concepts for better understanding of the hypothesis testing techniques.

Ordinarily, when one talks about hypothesis, one simply means a mere assumption or

some supposition to be proved or disproved. But for a researcher hypothesis is a formal question that he intends to resolve. Thus, a hypothesis may be defined as a proposition or a set of proposition set forth as an explanation for the occurrence of some specified group of phenomena either asserted merely as a provisional conjecture to guide some investigation or accepted as highly probable in the light of established facts. Quite often a research hypothesis is a predictive statement, capable of being tested by scientific methods, that relates an independent variable to some dependent variable. For example, consider statements like the following ones: "Students who receive counselling will show a greater increase in creativity than students not receiving counselling" Or "the automobile A is performing as well as automobile B." These are hypotheses capable of being objectively verified and tested. Thus, we may conclude that a hypothesis states what we are looking for and it is a proposition which can be put to a test to determine its validity.

Hypothesis must possess the following characteristics:

(i) Hypothesis should be clear and precise. If the hypothesis is not clear and precise, the inferences drawn on its basis cannot be taken as reliable.

(ii) Hypothesis should be capable of being tested. In a swamp of untestable hypotheses, many a time the research programmes have bogged down. Some prior study may be done by researcher in order to make hypothesis a testable one. A hypothesis "is testable if other deductions can be made from it which, in turn, can be confirmed or disproved by observation."

(iii) Hypothesis should state relationship between variables, if it happens to be a relational hypothesis.

(iv) Hypothesis should be limited in scope and must be specific. A researcher must remember that narrower hypotheses are generally more testable and he should develop such hypotheses.

(v) Hypothesis should be stated as far as possible in most simple terms so that the same is easily understandable by all concerned. But one must remember that simplicity of hypothesis has nothing to do with its significance.

(vi) Hypothesis should be consistent with most known facts i.e., it must be consistent with a substantial body of established facts.

(vii) Hypothesis should be amenable to testing within a reasonable time. One should not use even an excellent hypothesis, if the same cannot be tested in reasonable time for one cannot spend a life-time collecting data to test it.

(viii) Hypothesis must explain the facts that gave rise to the need for explanation. This means that by using the hypothesis plus other known and accepted generalisations, one should be able to deduce the original problem condition. Thus, hypothesis must actually explain what it claims to explain; it should have empirical reference.

16.4 TYPES OF HYPOTHESIS

Hypothesis is just an assumption or supposition made on the basis of some evidences. It's a starting point of any investigation that translates questions of research into prediction. Suppose there are two or more variables in a study, the hypothesis predicts the relationship between them. In every research researcher's objective is to test null hypothesis and if this null hypothesis is rejected the alternative hypothesis is supposed to be accepted.

16.4.1 Null Hypothesis

It gives the statement contrary to working hypothesis. It is a negative statement and you find no relationship between dependent and independent variable here. It is denoted by 'H_0'. Null hypothesis can be simple or complex and directional or non directional. Null hypothesis testing is a formal approach to deciding between two interpretations of a statistical relationship in a sample. One interpretation is called the null hypothesis (often symbolized H_0 and read as "H-naught"). This is the idea that there is no relationship in the population and that the relationship in the sample reflects only sampling error. Informally, the null hypothesis is that the sample relationship "occurred by chance." For more clarity, a null hypothesis is a type of hypothesis used in statistics that proposes no statistical significance exists in a set of given observations. The null hypothesis attempts to show that no variation exists between variables or that a single variable is not different than its mean. It is presumed to be true until statistical evidence nullifies it for an alternative hypothesis. Further, it is the proposition that implies no effect or no relationship between phenomena or populations. Any observed difference would be due to sampling error (random chance) or experimental error. The null hypothesis is popular because it can be tested and found to be false, which then implies there is a relationship between the observed data. It may be

easier to think of it as a nullifiable hypothesis or one the researcher seeks to nullify. Also Known As: H_0, no-difference hypothesis. There are two ways to state a null hypothesis. One is to state it as a declarative sentence, and the other is to present it as a mathematical statement.

For example, say a researcher suspects that exercise is correlated to weight loss, assuming a diet remains unchanged. The average length of time to achieve a certain weight loss is an average of 6 weeks when a person works out five times a week. The researcher wants to test whether weight loss takes longer if the number of workouts is reduced to three times a week.

The first step to writing the null hypothesis is to find the (alternate) hypothesis. In a word problem like this, you're looking for what you expect as the outcome of the experiment. In this case, the hypothesis is "I expect weight loss to take longer than 6 weeks."

This can be written mathematically as: H_1: ì > 6

In this example, ì is the average.

Now, the null hypothesis is what you expect if this hypothesis does not happen. In this case, if weight loss isn't achieved in greater than 6 weeks, then it must occur at a time equal to or less than 6 weeks.

H_0: ì d" 6

The other way to state the null hypothesis is to make no assumption about the outcome of the experiment. In this case, the null hypothesis is simply that the treatment or change will have no effect on the outcome of the experiment. For this example, it would be that reducing the number of workouts would not affect time to achieve weight loss:

H_0: ì = 6

"Hyperactivity is unrelated to eating sugar" is an example of a null hypothesis. If the hypothesis is tested and found to be false, using statistics, then a connection between hyperactivity and sugar ingestion may be indicated. A significance test is the most common statistical test used to establish confidence in a null hypothesis.

Another example of a null hypothesis would be, "Plant growth rate is unaffected by the

presence of cadmium in the soil." A researcher could test the hypothesis by measuring the rate of plant growth of plants grown in a medium lacking cadmium compared with the rate of growth of plants grown in a medium containing different amounts of cadmium. Disproving the null hypothesis would set the groundwork for further research into the effects of different concentrations of the element in soil.

Sometimes, you may be wondering why you would want to test a hypothesis just to find it false. Why not just test an alternate hypothesis and find it true? The short answer is that it's part of the scientific method. In science, "proving" something doesn't occur. Science uses math to determine the probability a statement is true or false. It turns out it's much easier to disprove a hypothesis than to ever prove one. Also, while the null hypothesis may be simply stated, there's a good chance the alternate hypothesis is incorrect.

For example, if your null hypothesis is that plant growth is unaffected by duration of sunlight, you could state the alternate hypothesis several different ways. Some of these statements might be incorrect. You could say plants are harmed by more than 12 hours of sunlight or that plants need at least 3 hours of sunlight, etc. There are clear exceptions to those alternate hypotheses, so if you test the wrong plants, you could reach the wrong conclusion. The null hypothesis is a general statement that can be used to develop an alternate hypothesis, which may or may not be correct. Testing the null hypothesis can tell you whether your results are due to the effect of manipulating the dependent variable or due to chance. Rejecting a hypothesis does not mean an experiment was "bad" or that it didn't produce results. In fact, it is often one of the first steps toward further inquiry. A significance test is used to determine the likelihood that the results supporting the null hypothesis are not due to chance. A confidence level of 95 percent or 99 percent is common. Keep in mind, even if the confidence level is high, there is still a small chance the null hypothesis is not true, perhaps because the experimenter did not account for a critical factor or because of chance.

Null Hypothesis Examples	
Question	**Null Hypothesis**
Are teens better at math than adults?	Age has no effect on mathematical ability
Does taking aspirin every day reduce the chance of having a heart attack?	Taking aspirin daily does not affect heart attack risk.
Do teens use cell phones to access the internet more than adults?	Age has no effect on how cell phones are used for internet access.
Do cats care about the colour of their food?	Cats express no food preference based on colour.
Does chewing willow bark relieve pain?	There is no difference in pain relief after chewing willow bark versus taking a placebo

16.4.2 Alternative Hypothesis

The other interpretation is called the alternative hypothesis (often symbolised as H_1/H_a). This is the idea that there is a relationship in the population. The alternate hypothesis, H_A or H_1, proposes that observations are influenced by a non random factor. In an experiment, the alternate hypothesis suggests that the experimental or independent variable has an effect on the dependent variable.

16.5 PROCEDURE OF TESTING HYPOTHESIS

To test a hypothesis means to tell (on the basis of the data the researcher has collected) whether or not the hypothesis seems to be valid. In hypothesis testing the main question is: whether to accept the null hypothesis or not to accept the null hypothesis? Procedure for hypothesis testing refers to all those steps that we undertake for making a choice between the two actions i.e., rejection and acceptance of a null hypothesis. The various steps involved in hypothesis testing are stated below:

1) Set up a hypothesis

This step consists of making a formal statement of the null hypothesis (H_0) and also of the alternative hypothesis (Ha). This means that hypotheses should be clearly stated, considering the nature of the research problem. For instance Mr. Mohan of the Civil Engineering Department wants to test the load bearing capacity of an old bridge which must be more than 10 tons, in that case he can state his hypotheses as under:

Null hypothesis H_0: ì = 10 tons

Alternative Hypothesis Ha: ì >10 tons

Take another example. The average score in an aptitude test administered at the national level is 80. To evaluate a state's education system, the average score of 100 of the state's students selected on random basis was 75. The state wants to know if there is a significant difference between the local scores and the national scores. In such a situation the hypotheses may be stated as under:

Null hypothesis H_0: ì = 80

Alternative Hypothesis Ha: ì > 80

The formulation of hypotheses is an important step which must be accomplished with due care in accordance with the object and nature of the problem under consideration. It also indicates whether we should use a one-tailed test or a two-tailed test. If Ha is of the type greater than (or of the type lesser than), we use a one-tailed test, but when Ha is of the type "whether greater or smaller" then we use a two-tailed test.

2) Set up a suitable level of significance

The hypotheses are tested on a pre-determined level of significance and as such the same should be specified. Generally, in practice, either 5% level or 1% level is adopted for the purpose. The level of significance is depending on the question. If it is given then we have to take that given value. Otherwise, if nothing is given about the level of question, then we take 5%. The factors that affect the level of significance are: (a) the magnitude of the difference between sample means; (b) the size of the samples; (c) the variability of measurements within samples; and (d) whether the hypothesis is directional or non-directional (A directional hypothesis is one which predicts the direction of the difference between, say, means). In brief, the level of significance must be adequate in the context of the purpose and nature of enquiry.

3) Test statistics

After deciding the level of significance, the next step in hypothesis testing is to determine the appropriate sampling distribution. The choice generally remains between normal distribution and the t-distribution. The rules for selecting the correct distribution are similar to those which we have stated earlier in the context of estimation. Test statistics include, t-test, z-test, F-test, chi-square test etc.

4) Computations or calculations

Another step is to select a random sample(s) and compute an appropriate value from the sample data concerning the test statistic utilising the relevant distribution. In other words, draw a sample to furnish empirical data. One has then to calculate the probability that the sample result would diverge as widely as it has from expectations, if the null hypothesis were in fact true.

5) Making Decisions

Yet another step consists in comparing the probability thus calculated with the specified value for á, the significance level. If the calculated probability is equal to or smaller than á value in case of one-tailed test (and á/2 in case of two-tailed test), then reject the null hypothesis (i.e., accept the alternative hypothesis), but if the calculated probability is greater, then accept the null hypothesis. In case we reject H_0, we run a risk of (at most the level of significance) committing an error of Type I, but if we accept H_0, then we run some risk (the size of which cannot be specified as long as the H_0 happens to be vague rather than specific) of committing an error of Type II. Hence, for selecting or rejecting a null hypothesis, we make comparison between calculated and table value. If calculated value is more than table value, we reject the null hypothesis and when calculated value is less than the table value, null hypothesis is accepted.

16.6 ERRORS IN HYPOTHESIS TESTING

In the context of testing of hypothesis, there are basically two types of errors we can make. We may reject H_0 when H_0 is true and we may accept H_0 when in fact H_0 is not true. The former is known as Type I error and the latter as Type II error.

16.6.1 Type I Error

Type I error means rejection of hypothesis which should have been accepted. Type I error is denoted by á (alpha) known as á error, also called the level of significance of test. The probability of Type I error is usually determined in advance and is understood as the level of significance of testing the hypothesis. If type I error is fixed at 5 per cent, it means that there are about 5 chances in 100 that we will reject H_0 when H_0 is true. We can control Type I error just by fixing it at a lower level. For instance, if we fix it at 1 per cent, we will

say that the maximum probability of committing Type I error would only be 0.01. But with a fixed sample size, n, when we try to reduce Type I error, the probability of committing Type II error increases. Both types of errors cannot be reduced simultaneously.

16.6.2 Type II Error

Type II error means accepting the hypothesis which should have been rejected. Type II error is denoted by â (beta) known as â error. There is a trade-off between two types of errors which means that the probability of making one type of error can only be reduced if we are willing to increase the probability of making the other type of error. To deal with this trade-off in business situations, decision-makers decide the appropriate level of Type I error by examining the costs or penalties attached to both types of errors. If Type I error involves the time and trouble of reworking a batch of chemicals that should have been accepted, whereas Type II error means taking a chance that an entire group of users of this chemical compound will be poisoned, then in such a situation one should prefer a Type I error to a Type II error. As a result one must set very high level for Type I error in one's testing technique of a given hypothesis. Hence, in the testing of hypothesis, one must make all possible effort to strike an adequate balance between Type I and Type II errors.

In a tabular form the said two errors can be presented as follows:

	Decision	
	Accept H_0	**Reject H_0**
H_0 (true)	Correct decision	Type I error (á)
H_0 (false)	Type II error (â)	Correct Decision

16.7 SUMMARY

There is an ever increasing demand for business managers with numerate ability as well as literary skills, so that they can present data which requires analysis and interpretation but, more importantly, they can quickly scan and understand analysis provided both from within the firm and by outside organisations. In the competitive and dynamic business world, those enterprises which are most likely to succeed, and indeed survive are those which are capable of maximising the use of various tools of testing hypothesis. This lesson has attempted to describe the meaning of hypothesis which helps the researcher in analysing

the data in the field of business and management.

16.8 GLOSSARY

- **Hypothesis:** It is a quantitative statement about a parent population based on a subset of the population i.e. sample.
- **Null Hypothesis (H_0):** Hypothesis which is tested for its possible rejection under the assumption is called as null hypothesis.
- **Alternative Hypothesis (H_1)**: Any hypothesis which is complementary to the null hypothesis.
- **Type I error:** Probability of rejecting the null hypothesis when it should be accepted, that is, concluding that two means are significantly different when in fact they are the same. Denoted with á.
- **Type II error:** Probability of failing to reject the null hypothesis when it should be rejected, that is, concluding that two means are not significantly different when in fact they are different. Denoted with â.

16.9 SELF ASSESSMENT QUESTIONS

Fill in the blanks:

1. The null hypothesis asserts that there is no true difference in the.....and the ..... in the particular matter under consideration.
2. Type I error is committed when the hypothesis is true but our test.....it.
3. Type II errors are made when we accept a null hypothesis which is.....
4. In .....tail test the rejection region is located in one tail.

16.10 LESSON END EXERCISE

1) Explain various steps of testing hypothesis.

2) Distinguish between null and alternative hypothesis by taking suitable examples.

3) Explain the two types of error. Which error is more serious and why. Explain?

4) Point out the difference between one tail and two tail tests.

5) Discuss the role of hypothesis in making a research design. What are the characteristics of a good hypothesis?

6) Point out the significance of null hypothesis.

16.11 SUGGESTED READING

- Levin, R.I. Robin, D.S. *Statistics for Management,* Prentice-Hall of India.New Delhi.
- Aczel, A. D. Sounderpandian, *J. Complete Business Statistics,* Mc Graw Hill Publishing. New Delhi. downloaded
- Anderson, S., W. *Statistics for Business and Economics,* Cengage Learning. New Delhi.
- Kazmeir L. J. *Business Statistics,* Tata Mc Graw Hill. New Delhi.
- Vohra, N. D. *Business Statistics,* Tata Mc Graw Hill. New Delhi.

PARAMETRIC AND NON-PARAMETRIC TESTS AND TEST OF SIGNIFICANCE FOR LARGE SAMPLES

STRUCTURE

17.1 INTRODUCTION

Hypothesis testing helps to decide on the basis of a sample data, whether an assumption about the population is likely to be true or false. Statisticians have developed several tests of hypotheses (also known as the tests of significance) for the purpose of testing of

hypotheses which can be classified as: (a) Parametric tests or standard tests of hypotheses; and (b) Non-parametric tests or distribution-free test of hypotheses. Parametric tests usually assume certain properties of the parent population from which we draw samples. Assumptions like observations come from a normal population, sample size is large, assumptions about the population parameters like mean, variance, etc., must hold good before parametric tests can be used. But there are situations when the researcher cannot or does not want to make such assumptions. In such situations we use statistical methods for testing hypotheses which are called non-parametric tests because such tests do not depend on any assumption about the parameters of the parent population. Besides, most non-parametric tests assume only nominal or ordinal data, whereas parametric tests require measurement equivalent to at least an interval scale. Non-parametric tests need more observations than parametric tests. In this lesson we take up only concept and important fundamentals of parametric tests, non-parametric tests and also test of significance for large samples. The details of each individual parametric (t-test) and non-parametric test (Chi-Square, Mann-Whitney, Kruskal Wallis Test) along with their computations will be taken up in subsequent lessons.

17.2 OBJECTIVES

After reading this lesson, you will be able to:

- Explain the concept of test of significance for large samples.
- Understand the meaning of parametric and non-parametric tests.
- Differentiate between parametric and non-parametric tests.
- Distinguish between small and large samples.

17.3 PARAMETRIC TESTS

A parametric statistical test is one that makes assumptions about the parameters (defining properties) of the population distribution(s) from which one's data are drawn. Parametric test is a test in which parameters are assumed and the population distribution is always known. These tests are common, and this makes performing research pretty straightforward without consuming much time. Also, these tests assume the parameters of the population

and the distributions of the data it came from. The parametric test is used for quantitative data with continuous variables. The data that parametric tests are used on are measured on ratio scales measurement and follow a normal distribution.

The most widely and commonly used parametric tests are t-test (for sample size less than 30), Z-test (for sample size greater than 30), ANOVA, Pearson's rank Correlation. The central tendency value that is taken into considerations is the mean of the distribution and is mostly applicable to a normal distribution for data. The disadvantage of this kind of test is that since the central tendency value is mean, the data is highly prone to be affected by outliers and thus prone to being skewed and this reduces the statistical power of this test. Continuous distributions like the data about various heights or weights of a species over time, data about temperatures are examples where parametric tests are used. Although, due to the assumptions about the data, its application is a little less versatile in real life. In some cases the population may not be normally distributed, yet the tests will be applicable on account of the fact that we mostly deal with samples and the sampling distributions closely approach normal distributions. Parametric tests are preferred for the following reasons:

1. We are rarely interested in a significance test alone; we would like to say something about the population from which the samples came, and this is best done with estimates of parameters and confidence intervals.
2. It is difficult to do flexible modeling with non-parametric tests, for example allowing for confounding factors using multiple regression.
3. Parametric tests usually have more statistical power than their non-parametric equivalents. In other words, one is more likely to detect significant differences when they truly exist.

17.4 NON-PARAMETRIC TESTS

Non-parametric tests are tests that aren't dependent on any assumptions of the data distribution or parameters to analyse them. They are also sometimes referred to as "distribution-free tests". Here, Non-parametric doesn't necessarily mean that we know nothing about the population, it means that the data is skewed or "not normally distributed".

Since, no assumptions are made about the population parameters in the and therefore; it measures with the help of the median value. A few instances of Non-parametric tests are Kruskal-Wallis, Mann-Whitney etc. Also, the non-parametric test is a type of hypothesis test that is not dependent on any underlying hypothesis. Test values are found based on the ordinal or the nominal level. The non-parametric test is usually performed when the independent variables are non-metric.

This non-parametric test is more flexible in real-life as data found in real life is not necessarily normally distributed and is mostly clumped or non-linear. Due to their simplicity and robust nature, non-parametric tests are seen as less prone to improper use and misunderstanding. They are mostly used in populations that come in ranked order, such as movie ratings and reviews, putting up ratings for restaurants, and such. But, for data with large sample size, these tests lose a lot of their statistical power. It would seem prudent to use non-parametric tests in all cases, which would save one the bother of testing for Normality.

17.4.1 Difference between Parametric and Non-Parametric Tests

The fundamental differences between parametric and non-parametric test are discussed in the following points:

1. A statistical test, in which specific assumptions are made about the population parameter, is known as the parametric test. A statistical test used in the case of non-metric independent variables is called as non-parametric test.

2. In the parametric test, the test statistic is based on distribution. On the other hand, the test statistic is arbitrary in the case of non-parametric test.

3. In the parametric test, it is assumed that the measurement of variables of interest is done on interval or ratio level. As opposed to the non-parametric test, wherein the variable of interest are measured on nominal or ordinal scale.

4. In general, the measure of central tendency in the parametric test is mean, while in the case of the non-parametric test is median.

5. In the parametric test, there is complete information about the population. Conversely, in the non-parametric test, there is no information about the population.

6. The applicability of parametric test is for variables only, whereas non-parametric test applies to both variables and attributes.

7. For measuring the degree of association between two quantitative variables, Pearson's coefficient of correlation is used in the parametric test, while spearman's rank correlation is used in the nonparametric test.

17.5 TEST OF SIGNIFICANCE FOR LARGE SAMPLES

While testing hypothesis in research there are two versions of tests available based on the size of samples. These are small samples and large samples test. The sampling theory for large samples is not applicable in small samples because when samples are small, we cannot assume that the sampling distribution is approximately normal. As such we require a new technique for handling small samples, particularly when population parameters are unknown. The division between the theories of large and small samples is therefore, a very real one, though it is not always easy to draw a precise line of demarcation. It should be noted that, the method and theory of small samples are applicable to large samples, though the reverse is not true. While dealing with small samples our main interest is not to estimate the population values as is true in large samples; but here our purpose lies in testing a given hypothesis, i.e., in ascertaining whether observed values could have arisen by sampling fluctuations from some values given in advance.

Therefore, the tests of significance used for dealing with problems relating to large samples are different from those used for small samples. This is so because the assumptions we make in case of large samples do not hold good for small samples. In case of large samples, we assume that the sampling distribution tends to be normal and the sample values are approximately close to the population values. As such we use the characteristics of normal distribution and apply what is known as z-test. When n is large, the probability of a sample value of the statistic deviating from the parameter by more than 3 times, its standard error is very small (it is 0.0027 as per the table giving area under normal curve) and as such the z-test is applied to find out the degree of reliability of a statistic in case of large samples.

17.5.1 Z TEST

Z-test is based on the normal probability distribution and is used for judging the significance

of several statistical measures, particularly the mean. The relevant test statistic, z, is worked out and compared with its probable value (to be read from table showing area under normal curve) at a specified level of significance for judging the significance of the measure concerned. This is a most frequently used test in research studies. This test is used even when binomial distribution or t-distribution is applicable on the presumption that such a distribution tends to approximate normal distribution as 'n' becomes larger. Z-test is generally used for comparing the mean of a sample to some hypothesised mean for the population in case of large sample, or when population variance is known. Z-test is also used for judging the significance of difference between means of two independent samples in case of large samples, or when population variance is known. Z-test is also used for comparing the sample proportion to a theoretical value of population proportion or for judging the difference in proportions of two independent samples when n happens to be large. Besides, this test may be used for judging the significance of median, mode, coefficient of correlation and several other measures.

I. Population normal, population infinite, sample size is large but variance of the population is known

In such a situation z-test is used for testing hypothesis of mean and the test statistic z is worked our as under:

$$Z = \frac{(\bar{X} - \mu)}{\sigma} \times \sqrt{n}$$

whereas, $\bar{X}$ = Sample mean

μ = Population mean, which is given in the question

n= sample size

σ = Population standard deviation which is known in case of Z-test.

In this case the table value is already specified.

At 5% it is 1.96 and at 1% it is 2.58.

I. Population Variances are known or the Samples Happen to be Large Samples

Here we are interested to find out the significant difference between the means of two samples in case of large samples when population variances are known.

$$Z = \frac{(\bar{X}1 - \bar{X}2)}{\sqrt{\sigma 1^2/n1 + \sigma 2^2/n2}}$$

III. To test the significance of the difference between two sample variances or standard deviation when sample size is more than 30

$$Z = \frac{(S1 - S2)}{\sqrt{\sigma 1^2/2n1 + \sigma 2^2/2n2}}$$

Example 1: A sample of 400 male students is found to have a mean height 67.47 inches. Can it be reasonably regarded as a sample from a large population with mean height 67.39 inches and standard deviation 1.30 inches? Test at 5% level of significance.

Solution: Taking the null hypothesis that the mean height of the population is equal to 67.39 inches.

Level of significance = 5%

$$Z = \frac{(\bar{X} - \mu)}{\sigma} \times \sqrt{n}$$

Given data: $\bar{X} = 67.47$ $\mu = 67.39$ $\sigma = 1.30$ n= 400

Therefore, $Z = \frac{(\bar{X} - \mu)}{\sigma} \times \sqrt{n}$

$$= \frac{(67.47 - 67.39)}{1.30} \times \sqrt{400} = 1.231 \text{ (CV)}$$

Table value at 5% = 1.96

Since the CV is less than the table value.

Hence, H_0 is accepted. We may conclude that the given sample (with mean height = 67.47") can be regarded to have been taken from a population with mean height 67.39" and standard deviation 1.30" at 5% level of significance.

Example 2: A simple random sampling survey in respect of monthly earnings of semi-skilled workers in two cities gives the following statistical information:

City	*Mean monthly earnings (Rs)*	*Standard deviation of sample data of monthly earnings*	*Size of sample*
A	695	40	200
B	710	60	175

Test the hypothesis at 5 per cent level that there is no difference between monthly earnings

of workers in the two cities.

Solution: Taking the null hypothesis that there is no difference in earnings of workers in the two cities.

Level of significance = 5%

$$\text{Test statistics} = Z - \frac{(\bar{X}1 - \bar{X}2)}{\sqrt{\frac{\sigma 1^2}{n1} + \frac{\sigma 2^2}{n2}}}$$

Given Data: The given information as:

Sample 1 (City A)	Sample 2 (City B)
Mean of X_1 = Rs 695	Mean of X_2 = Rs 710
$n_1 = 200$	$n_2 = 175$
$S_1 = 40$	$S_2 = 60$

As the sample size is large, we shall use z-test for difference in means assuming the populations to be normal and shall work out the test statistic z as under:

$$Z = \frac{(\bar{X}1 - \bar{X}2)}{\sqrt{\frac{\sigma 1^2}{n1} + \frac{\sigma 2^2}{n2}}}$$

$$= \frac{695 - 710}{\sqrt{\frac{40^2}{200} + \frac{60^2}{175}}} = -2.809 = 2.809$$

The critical value of z at 5% level is 1.96.

The calculated value of z is 2.809 which falls in the rejection region and thus we reject H_0 at

5 per cent level and conclude that earning of workers in the two cities differ significantly.

17.6 SUMMARY

To make a choice between parametric and the non-parametric test is not easy for a researcher conducting statistical analysis. For performing hypothesis, if the information about the population is completely known, by way of parameters, then the test is said to be parametric test whereas, if there is no knowledge about population and it is needed to test the hypothesis on population, then the test conducted is considered as the non-parametric test. Both parametric and non-parametric tests are integral parts of analyzing any given data. The choosing of which test to perform depends upon what kind of data we have,

what the sample size is, and how much pre-requisite knowledge about the population we have. Data having a large sample size requires a parametric test instead of non-parametric as it is more accurate. In the case of a small sample size data, the non-parametric test is preferred. No test is better than the other as both operate in different situations. As a statistician, we need to keep in mind that a non-parametric test is an alternative to a parametric test, not a substitute.

17.7 GLOSSARY

- **Parametric Test:** Parametric tests are those that make assumptions about the parameters of the population distribution from which the sample is drawn. This is often the assumption that the population data are normally distributed.
- **Non-parametric Test:** Non-parametric tests are "distribution-free" and, as such, can be used for non-Normal variables.
- **Small Samples:** If the sample size n is less than 30 (n<30), it is known as small sample. For small samples the sampling distributions are t, F and ?2 distribution. A study of sampling distributions for small samples is known as small sample theory.
- **Large Samples:** The sample size n is greater than 30 (n?30) it is known as large sample.

17.8 SELF ASSESSMENT QUESTIONS

A. Fill in the Blanks:

1. Parametric tests are based on some restrictive assumptions about the _____.
2. _____are not dependent upon the restrictive normality assumption of the population.
3. The Mann-Whitney U test is a counterpart of the_____to compares the means of two independent populations.
4. The _____is a non-parametric alternative to the t-test for related samples.
5. The Kruskal-Wallis test is the non-parametric test alternative to the _____.

17.9 LESSON END EXERCISE

1. Explain parametric tests with examples.

2. Differentiate between small and large sample tests.

3. Can we use parametric tests with non-parametric data?

4. What is main difference between parametric tests and non-parametric tests?

5. A sample of 900 members is found to have a mean of 3.47 cm. Can it be reasonably regarded as a simple sample from a large population with mean 3.23 cm. and standard deviation 2.31 cm.?

17.10 SUGGESTED READING

- Levin, R.I. Robin, D.S. *Statistics for Management,* Prentice-Hall of India.New Delhi.

- Aczel, A. D. Sounderpandian, J. *Complete Business Statistics,* Mc Graw Hill Publishing. New Delhi. downloaded
- Anderson, S., W. *Statistics for Business and Economics,* Cengage Learning. New Delhi.
- Kazmeir L. J. *Business Statistics,* Tata Mc Graw Hill. New Delhi.
- Vohra, N. D. Business Statistics. Tata Mc Graw Hill. New Delhi.

T-TEST: ONE SAMPLE T-TEST, INDEPENDENT SAMPLES T-TEST AND DEPENDENT SAMPLES T-TEST

STRUCTURE

18.1 INTRODUCTION

The sampling theory for large samples is not applicable in small samples because when samples are small, we cannot assume that the sampling distribution is approximately normal. As such we require a new technique for handling small samples, particularly when population

parameters are unknown. The division between the theories of large and small samples is therefore, a very real one, though it is not always easy to draw a precise line of demarcation. It should be noted that, the method and theory of small samples are applicable to large samples, though the reverse is not true. While dealing with small samples our main interest is not to estimate the population values as is true in large samples; but here our purpose lies in testing a given hypothesis, i.e., in ascertaining whether observed values could have arisen by sampling fluctuations from some values given in advance.

Sir William S. Gosset (pen name Student) developed a significance test, known as Student's t-test, based on t distribution and through it made significant contribution in the theory of sampling applicable in case of small samples. Student's t-test is used when two conditions are fulfilled viz., the sample size is 30 or less and the population variance is not known. While using t-test we assume that the population from which sample has been taken is normal or approximately normal, sample is a random sample, observations are independent, there is no measurement error and that in the case of two samples when equality of the two population means is to be tested, we assume that the population variances are equal.

18.2 OBJECTIVES

After reading this lesson, you will be able to:

- Learn the concept of testing of significance under small samples.
- Perform test of significance of mean.
- Perform test of significance of difference between two means.
- Distinguish between independent and dependent samples.

18.3 T-Test

T-test is based on t-distribution and is considered an appropriate test for judging the significance of a sample mean or for judging the significance of difference between the means of two samples in case of small sample(s) when population variance is not known (in which case we use variance of the sample as an estimate of the population variance). In case two samples are related, we use paired t-test (or what is known as difference test) for judging the significance of the mean of difference between the two related samples. It

can also be used for judging the significance of the coefficients of simple and partial correlations. The relevant test statistic, t, is calculated from the sample data and then compared with its probable value based on t-distribution (to be read from the table that gives probable values of t for different levels of significance for different degrees of freedom) at a specified level of significance for concerning degrees of freedom for accepting or rejecting the null hypothesis. It may be noted that t-test applies only in case of small sample(s) when population variance is unknown. For applying t-test, we work out the value of test statistic (i.e., 't') and then compare with the table value of t (based on 't' distribution) at certain level of significance for given degrees of freedom. If the calculated value of 't' is either equal to or exceeds the table value, we infer that the difference is significant, but if calculated value of t is less than the concerning table value of t, the difference is not treated as significant.

18.4 ONE SAMPLE T-TEST

One Sample t-test is generally applied to test the significance of the Mean of a random sample i.e. whether there is a significance difference between the sample mean and the population mean. In other words, here in this case generally the null hypothesis will be:

$H_0 = \bar{X} = \mu$

$$t = \frac{(\bar{X} - \mu)}{s} \times \sqrt{n}$$

Where $\bar{X}$ = Sample mean

μ = Population mean

n= sample size

S = sample standard deviation $= \sqrt{\frac{\Sigma(X - \bar{x})^2}{n - 1}}$

and the degrees of freedom = $(n - 1)$.

If this calculated value of t is greater than the table value, null hypothesis is rejected, means the difference is significant. But if calculated value is less than the table value, null hypothesis is accepted, difference is insignificant.

Example 1: The specimen of copper wires drawn form a large lot has the following breaking strength (in kg. weight):

578, 572, 570, 568, 572, 578, 570, 572, 596, 544

Test (using Student's t-statistic) whether the mean breaking strength of the lot may be taken to be 578 kg weight (Test at 5 per cent level of significance.

Solution: Let us take the null hypothesis that there is no significance difference between the sample mean and the population mean.

Level of significance: 5%

As the sample size is mall (since n = 10) and the population standard deviation is not known, we shall use t-test assuming normal population and shall work out the test statistic t as under:

Test statistics = $t = \frac{(\bar{X} - \mu)}{s} \times \sqrt{n}$

Give date n = 10

To find the value of mean and s we make the following computations.

S. No	X	(X-mean of X)	(X-mean of X)^2
1	578	6	36
2	572	0	0
3	570	-2	4
4	568	-4	16
5	572	0	0
6	578	6	36
7	570	-2	4
8	572	0	0
9	596	24	576
10	544	-28	784
	$\sum X = 5720$		$\sum d^2 = 1456$
	Mean = 5720/10 = 572		

$S = \sqrt{\frac{1456}{10-1}} = \sqrt{\frac{1456}{9}} = 12.72$

Therefore, $t = \frac{(\bar{X} - \mu)}{S} \times \sqrt{n} = \frac{(572 - 578)}{12.72} \times \sqrt{10} = 1.488$ (CV)

Degree of freedom = $(n - 1) = (10 - 1) = 9$

$t_{0.05} = 2.262$ (TV)

Since CV (1.488) is less than the table value (2.262). Therefore, null hypothesis is accepted.

18.5 INDEPENDENT SAMPLES T-TEST

The Independent samples t-test is used to test the significance difference between the means of two samples or when two independent samples are given.

$$\text{t- Statistics} = \frac{(\bar{X}1 - \bar{X}2)}{S} \times \sqrt{\frac{n1 \times n2}{(n1 + n2)}}$$

Where, X_1 = mean of first sample

X_2 = mean of second sample

n_1 = number of observations in the first sample

n_2 = number of observations in the second sample

S = combined standard deviation

The value of S is calculated as:

$$S = \sqrt{\frac{\Sigma(X1 - \bar{x}1)^2 + \Sigma(X2 - \bar{x}2)^2}{n1 + n2 - 2}}$$

When actual means are in fractions the deviations should be taken from assumed means. In such a case the combined standard deviation is obtained by applying the following formula:

$$S = \sqrt{\frac{\Sigma(X1 - A1)^2 + \Sigma(X2 - A2)^2}{n1 + n2 - 2}}$$

A_1 = Assumed mean of the first sample

A_2 = Assumed mean of the second sample

X_1 = Actual mean of first sample

X_2 = Actual mean of second sample

the degrees of freedom = $(n_1 + n_2 - 2)$

Example 2: Sample of sales in similar shops in two towns are taken for a new product with the following results:

Town	*Mean sales*	*Variance*	*Size of sample*
A	57	5.3	5
B	61	4.8	7

Is there any evidence of difference in sales in the two towns? Use 5 per cent level of significance for testing this difference between the means of two samples.

Solution: Taking the null hypothesis that the means of two populations do not differ.

Level of significance = 5%

$$\text{t- statistics} = \frac{(\bar{X}1 - \bar{X}2)}{S} \times \sqrt{\frac{n1 \times n2}{(n1+n2)}}$$

Given data: Mean of Ist sample, i.e. $\bar{X}_1 = 57$, variance = 5.3 $n_1 = 5$

Mean of second sample, i.e. $\bar{X}_2 = 61$, variance = 4.8 $n_2 = 7$

Since in the given question variances of the population are not known and the size of samples is small, we shall use t-test for difference in means, assuming the populations to be normal and can work out the test statistic t as under:

$$t = \frac{(\bar{X}1 - \bar{X}2)}{\sqrt{(n1-1)\,variance\ of\ ist\ sample + (n2-1) variance of\ second\ sample}} \sqrt{\frac{n1 \times n2}{(n1+n2)}}$$

$$t = \frac{57-61}{4 \times 5.3 + 6 \times 4.8} \sqrt{(5 \times 7) \div 12} = 3.053$$

Degree of freedom = $n_1 + n_2 - 2 = 5 + 7 - 2 = 10$

Table value at 5% level with 10 degree of freedom = 2.228

Here, CV > TV. Hence, null hypothesis is rejected.

18.6 DEPENDENT SAMPLES T-TEST

When we are interesting in testing the difference between the means of two dependent samples, the in that case we should apply dependent samples t-test also known as paired samples t-test or simply paired t-test.

Two samples are said to be dependent when the elements in one sample are related to those in the other in any significant or meaningful manner.

$$t = \frac{(\text{mean of d})}{s} \times \sqrt{n}$$

mean of d= mean of the differences = $\Sigma d/n$

S = standard deviation of the differences, calculated as:

$$S = \sqrt{\frac{\Sigma d^2 - (\Sigma d)^2/n}{n-1}}$$

Degree of freedom = n-1

If the CV of t exceeds for df, we say that the value of t is significant at 5% level. If $t < t_{0.05}$ the data are consistent with the hypothesis of an uncorrelated population.

Example 3: Memory capacity of 9 students was tested before and after training. State at 5 per cent level of significance whether the training was effective from the following scores.

Student	1	2	3	4	5	6	7	8	9
Before	10	15	9	3	7	12	16	17	4
After	12	17	8	5	6	11	18	20	3

Use paired t-test.

Solution: Take the score before training as X and the score after training as Y and then taking the null hypothesis that the mean of difference is zero i.e. there is no significance difference between the performance of students before and after training.

Level of significance = 5%

$$t = \frac{(\text{mean of d}) \times \sqrt{n}}{s}$$

Given data, n= 9

Calculations:

To find the value of t, we shall first have to work out the mean and standard deviation of differences as shown below:

Students	Score before training X	Score after training Y	Difference (d)	d^2
1	10	12	-2	4
2	15	17	-2	4
3	9	8	1	1
4	3	5	-2	4
5	7	6	1	1
6	12	11	1	1
7	16	18	-2	4
8	17	20	-3	9
9	4	3	1	1
n= 9			$\sum d$ = -7	$\sum d^2$ = 29

Mean of differences = Σd / n = -7/9 = -0.778

$$S = \sqrt{\frac{\Sigma d^2 - (\Sigma d)^2/n}{n-1}}$$

$$= \sqrt{\frac{29 - 49/9}{9-1}} = 1.715$$

Therefore, $t = \frac{-0.778}{1.715} \times 3 = -1.361 = 1.316$ (CV)

The table value of t at 5% level of significance is 1.860. The calculated value of t is 1.361 which is in the acceptance region and thus, we accept H0 and conclude that the difference in score before and after training is insignificant i.e., it is only due to sampling fluctuations. Hence we can infer that the training was not effective.

18.7 TWO TAILED AND ONE TAILED TESTS

In the context of hypothesis testing, these two terms are quite important and must be clearly understood. A two-tailed test rejects the null hypothesis if, say, the sample mean is significantly higher or lower than the hypothesised value of the mean of the population.

Such a test is appropriate when the null hypothesis is some specified value and the alternative hypothesis is a value not equal to the specified value of the null hypothesis. Symbolically, the two tailed test is appropriate when we have $H_0: \mu = \mu_{H_0}$ and which may mean $\mu > \mu_{H_0}$ or $\mu < \mu_{H_0}$ Thus, in a two-tailed test, there are two rejection regions, one on each tail of the curve which can be illustrated as under:

Mathematically we can state:

If the significance level is 5 per cent and the two-tailed test is to be applied, the probability of the rejection area will be 0.05 (equally spitted on both tails of the curve as 0.025) and that of the acceptance region will be 0.95 as shown in the above curve. If we take $\mu = 100$ and if our sample mean deviates significantly from 100 in either direction, then we shall reject the null hypothesis; but if the sample mean does not deviate significantly from , in that case we shall accept the null hypothesis. But there are situations when only one-tailed test is considered appropriate. A one-tailed test would be used when we are to test, say, whether the population mean is either lower than or higher than some hypothesised value.

For instance, if our then we are interested in what is known as left-tailed test (wherein there is one rejection region only on the left tail) which can be illustrated as below:

Mathematically we can state:

If our $\mu = 100$ and if our sample mean deviates significantly from100 in the lower direction, we shall reject H0, otherwise we shall accept H0 at a certain level of significance. If the significance level in the given case is kept at 5%, then the rejection region will be equal to 0.05 of area in the left

tail as has been shown in the above curve.

In case our ,we are then interested in what is known as one tailed test (right tail) and the rejection region will be on the right tail of the curve as shown below:

Mathematically we can state:

If our $\mu = 100$ and if our sample mean deviates significantly from 100 in the upward direction, we shall reject H0, otherwise we shall accept the same. If in the given case the significance level is kept at 5%, then the rejection region will be equal to 0.05 of area in the right-tail as has been shown in the above curve.

It should always be remembered that accepting H0 on the basis of sample information does not constitute the proof that H0 is true. We only mean that there is no statistical evidence to reject it, but we are certainly not saying that H0 is true (although we behave as if H0 is true).

18.7.1 Measuring the Power of a Hypothesis Test

As stated in lesson 16 we may commit Type I and Type II errors while testing a hypothesis. The probability of Type I error is denoted as α (the significance level of the test) and the probability of Type II error is referred to as β. Usually the significance level of a test is assigned in advance and once we decide it, there is nothing else we can do about β. But what can we say about β? We all know that hypothesis test cannot be a proof sometimes the test does not reject H_0 when it happens to be a false one and this way a Type II error is made. But we would certainly like that β (the probability of accepting H_0 when H_0 is not true) to be as small as possible. Alternatively, we would like that $1 - \beta$ (the probability of rejecting H0 when H0 is not true) to be as large as possible. If $1 - \beta$ is very much nearer to unity (i.e., nearer to 1.0), we can infer that the test is working quite well, meaning thereby that the test is rejecting H_0 when it is not true and if $1 - \beta$ is very much nearer to 0.0, then we infer that the test is poorly working, meaning thereby that it is not rejecting H0 when H0 is not true. Accordingly, $1 - \beta$ value is the measure of how well the test is working or what is technically described as the power of the test. In case we plot the values of $1 - \beta$ for each possible value of the population parameter (say μ, the true population mean) for which the H_0 is not true (alternatively the Ha is true), the resulting curve is known as the power curve associated with the given test. Thus, power curve of a hypothesis test is the curve that shows the conditional probability of rejecting H_0 as a function of the population parameter and size of the sample. The function defining this curve is known as the power function. In other words, the power function of a test is that function defined for

all values of the parameter(s) which yields the probability that H0 is rejected and the value of the power function at a specific parameter point is called the power of the test at that point. As the population parameter gets closer and closer to hypothesised value of the population parameter, the power of the test (i.e., 1 - β) must get closer and closer to the probability of rejecting H_0 when the population parameter is exactly equal to hypothesised value of the parameter. We know that this probability is simply the significance level of the test, and as such the power curve of a test terminates at a point that lies at a height of α (the significance level) directly over the population parameter.

Closely related to the power function, there is another function which is known as the operating characteristic function which shows the conditional probability of accepting H_0 for all values of population parameter(s) for a given sample size, whether or not the decision happens to be a correct one. If power function is represented as H and operating characteristic function as L, then we have L = 1 - H. However, one needs only one of these two functions for any decision rule in the context of testing hypotheses.

18.8 SUMMARY

In this lesson we have discussed about the t-test, two tailed and one tailed tests as well as the power of a hypotheses tests. At the start of the 20th century, William S. Gosset was working at Guinness in Ireland, trying to help brew better beer less expensively. As he had only small samples to study, he needed to find a way to make inferences about means without having to know standard deviation of the population. Writing under the pen name of "student", William Gosset solved this problem by developing what today is known as the student's t-distribution or the t-distribution. This t-distribution is very similar in appearance to the standardised normal distribution. Both the distributions are symmetrical and bell-shaped, with the mean and the median equal to zero. However, because S is used to estimate the unknown value of standard deviation of population, the values of t are more variable than those of Z. Therefore, the t distribution has more area in the nails and less in the center than does the standardized normal distribution. Moreover, a test result is statistically significant when the sample statistic is unusual enough relative to the null hypothesis that is we can reject the null hypothesis for the entire population. Keep in mind that there is no magic significance level that distinguishes between the studies that have a true effect and those that don't with 100% accuracy. The common alpha values of 0.05 and 0.01 are

simply based on tradition. For a significance level of 0.05, expect to obtain sample means in the critical region 5% of the time when the null hypothesis is true. In these cases, you won't know that the null hypothesis is true but you'll reject it because the sample mean falls in the critical region. That's why the significance level is also referred to as an error rate. This type of error doesn't imply that the experimenter did anything wrong or require any other unusual explanation. When the null hypothesis is true, it is possible to obtain these unusual sample means for no reason other than random sampling error. It's just luck of the draw. Significance levels and P values are important tools that help you quantify and control this type of error in a hypothesis test. Using these tools to decide when to reject the null hypothesis increases your chance of making the correct decision.

18.9 GLOSSARY

- **Critical Region:** Critical region is the region of values that corresponds to the rejection of the null hypothesis at some chosen probability level.
- **Level of Significance:** The level of significance is defined as the probability of rejecting a null hypothesis by the test when it is really true, which is denoted as ?.
- **Power of Hypothesis Test:** A tests power is the probability of correctly rejecting the null hypothesis when it is false; a test's power is influenced by the choice of significance level for the test, the size of the effect being measured, and the amount of data available.
- **T-test:** A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups or not.
- **Independent Samples:** Independent samples are samples that are selected randomly so that its observations do not depend on the values other observations. If the values in one sample reveal no information about those of the other sample, then the samples are independent.
- **Dependent Samples:** Dependent samples are paired measurements for one set of items. If the values in one sample affect the values in the other sample, then the samples are dependent.
- **One Tailed Test:** A one-tailed test is a statistical test in which the critical area of a

distribution is one-sided so that it is either greater than or less than a certain value, but not both

- Two Tailed Test: Two-tailed hypothesis tests are also known as non-directional and two-sided tests because you can test for effects in both directions. When you perform a two-tailed test, you split the significance level percentage between both tails of the distribution

18.10 SELF ASSESSMENT QUESTIONS

A. Fill in the blanks:

1. The standard deviation of sampling distribution is called as.....
2. The distribution formed of all possible values of a statistics is called the.....
3. The mean of sampling distribution of means is equal to the.....
4. Standard error provides an idea about the.....of sample.

B. Multiple Choice Questions:

1. A t-test is a significance test that assesses:

   a. The means of two independent groups
   b. The medians of two dependent groups
   c. The modes of two independent variables
   d. The standard deviation of three independent variables

2. To use a t-test, the dependent variable must have:

   a. Nominal or interval data
   b. Ordinal or ratio data
   c. Interval or ratio data
   d. Ordinal or interval data

3. The three types of t-tests are:

   a. One-sample t-tests
   b. Null Hypothesis t-tests

c. Independent sample t-tests
d. Paired samples t-tests

4. The T-test is not a reliable test:

a. True b. False

5. The T-test tells you about the significant difference between the two groups:

a. True b. False

18.11 LESSON END EXERCISE

1) Explain the concept of standard error. How it is useful in testing of hypothesis.

2) Explain terms: (i) Critical region (ii) Level of significance.

3) Explain the power of testing hypothesis with a suitable example.

4) What is a t-test? When it is used and for what purpose(s)? Explain by means of examples.

5) Ten students are selected at random from a school and their heights are found to be, in inches, 50, 52, 52, 53, 55, 56, 57, 58, 58 and 59. In the light of these data, discuss the suggestion that the mean height of the students of the school is 54 inches. You may use 5% level of significance.

6) The heights of six randomly chosen sailors are, in inches, 63, 65, 58, 69, 71 and 72. The heights of 10 randomly chosen soldiers are, in inches, 61, 62, 65, 66, 69, 69, 70, 71, 72 and 73. Do these figures indicate that soldiers are on an average shorter than sailors? Test at 5% level of significance.

18.12 SUGGESTED READING

- Levin, R.I. Robin, D.S. *Statistics for Management,* Prentice-Hall of India.New Delhi.
- Aczel, A. D. Sounderpandian, *J. Complete Business Statistics,* Mc Graw Hill Publishing. New Delhi. downloaded
- Anderson, S., W. Statistics for *Business and Economics,* Cengage Learning. New Delhi.
- Kazmeir L. J. *Business Statistics,* Tata Mc Graw Hill. New Delhi.
- Vohra, N. D. *Business Statistics,* Tata Mc Graw Hill. New Delhi.

CHI-SQUARE TEST

STRUCTURE

19.1 INTRODUCTION

In the preceding lessons, we used hypothesis testing procedure to analyse both numerical and categorical data. Lesson 18 presented some one-sample t-test, two samples test as well as dependent samples t-test. This lesson extends hypothesis testing to analyse differences between population proportions based on two or more samples and to test the hypothesis of independence in the joint responses to two categorical variables. The Chi-Square test is a statistical procedure used by researchers to examine the differences between categorical variables in the same population. For example, imagine that a research group is interested in whether or not education level and marital status are related for all people in the U.S. A chi-square (χ^2) statistic is a test that measures how a model compares to actual i.e. observed data. The data used in calculating a chi-square statistic must be random, raw, mutually exclusive, drawn from independent variables, and drawn from a large enough sample. For example, the results of tossing a fair coin meet these criteria. Also, the chi-square statistic compares the size of any discrepancies between the expected results and the actual results, given the size of the sample and the number of variables in the relationship. For these tests, degrees of freedom are utilized to determine if a certain null hypothesis can be rejected based on the total number of variables and samples within the experiment. As with any statistic, the larger the sample size, the more reliable the results.

There are two main kinds of chi-square tests: the test of independence, which asks a question of relationship, such as, "Is there a relationship between student sex and course choice?"; and the goodness-of-fit test, which asks something like "How well does the coin in my hand match a theoretically fair coin?" Chi-square analysis is applied to categorical variables and is especially useful when those variables are nominal (where order doesn't matter, like marital status or gender). For example, consider an imaginary coin with exactly a 50/50 chance of landing heads or tails and a real coin that you toss 100 times. If this coin is fair, then it will also have an equal probability of landing on either side, and the expected result of tossing the coin 100 times is that heads will come up 50 times and tails will come up 50 times. In this case, χ^2 can tell us how well the actual results of 100 coin flips compare to the theoretical model that a fair coin will give 50/50 results. The actual toss could come up 50/50, or 60/40, or even 90/10. The farther away the actual results of the 100 tosses is from 50/50, the less good the fit of this set of tosses is to the theoretical

expectation of 50/50, and the more likely we might conclude that this coin is not actually a fair coin.

19.2 OBJECTIVES

On successful completion of this lesson, you will be able to:

- Understand the role of non-parametric tests in hypothesis testing.
- Perform Chi-square test for hypothesis testing.
- Explain Chi-square as a test of independence.
- Understand the conditions for the applicability of Chi-square test.
- Know the various uses of this test statistics.

19.3 CHI-SQUARE TEST

The chi-square test is an important test amongst the several tests of significance developed by statisticians. Chi-square, symbolically written as χ^2 (Pronounced as Chi-square), is a statistical measure used in the context of sampling analysis for comparing a variance to a theoretical variance. 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." Thus, the chi-square test is applicable in large number of problems. The test is, in fact, a technique through the use of which it is possible for all researchers to (i) test the goodness of fit; (ii) test the significance of association between two attributes, and (iii) test the homogeneity or the significance of population variance. The test is based on χ^2-distribution. Such a distribution we encounter when we deal with collections of values that involve adding up squares. If we take each one of a collection of sample variances, divided them by the known population variance and multiply these quotients by $(n - 1)$, where n means the number of items in the sample, we shall obtain a χ^2-distribution. The χ^2-distribution is not symmetrical and all the values are positive. For making use of this distribution, one is required to know the degrees of freedom since for different degrees of freedom we have different curves. The smaller the number of degrees of freedom, the more skewed is the distribution.

19.3.1 Chi-Square as a Non-Parametric Test

Chi-square is an important non-parametric test and as such no rigid assumptions are necessary in respect of the type of population. We require only the degrees of freedom (implicitly of course the size of the sample) for using this test. As a non-parametric test, chi-square can be used (i) as a test of goodness of fit and (ii) as a test of independence.

19.3.2 As a Test of Goodness of Fit

As a test of goodness of fit 2 test enables us to see how well does the assumed theoretical distribution (such as Binomial distribution, Poisson distribution or Normal distribution) fit to the observed data? When some theoretical distribution is fitted to the given data, we are always interested in knowing as to how well this distribution fits with the observed data. The chi-square test can give answer to this. If the calculated value of 2 is less than the table value at a certain level of significance, the fit is considered to be a good one which means that the divergence between the observed and expected frequencies is attributable to fluctuations of sampling. But if the calculated value of 2 is greater than its table value, the fit is not considered to be a good one.

19.3.3 As a Test of Independence

As a test of independence, χ^2 (Chi-square) test enables us to explain whether or not two attributes are associated. For instance, we may be interested in knowing whether a new medicine is effective in controlling fever or not, χ^2 test will helps us in deciding this issue. In such a situation, we proceed with the null hypothesis that the two attributes (viz., 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 table value at a certain level of significance for given degrees of freedom, we conclude that null hypothesis stands 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 table 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). It may, however, be stated here 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. In order that we may apply the chi-square test either as a test of goodness of fit or as a test to judge the significance of association between attributes, it is necessary that the observed as well as theoretical or expected frequencies must be grouped in the same way and the theoretical distribution must be adjusted to give the same total frequency as we find in case of observed distribution. χ^2 is then calculated as follows:

$$\chi^2 = \sum \frac{(Oij - Eij)^2}{Eij}$$

where,

O_{ij} = observed frequency of the cell in *i*th row and *j*th column.
E_{ij} = expected frequency of the cell in *i*th row and *j*th column

If two distributions (observed and theoretical) are exactly alike, $\chi^2 = 0$; but generally due to sampling errors, χ^2 is not equal to zero and as such we must know the sampling distribution of χ^2so that we may find the probability of an observed χ^2 being given by a random sample from the hypothetical universe. Instead of working out the probabilities, we can use ready table which gives probabilities for given values of χ^2. Whether or not a calculated value of χ^2 is significant can be ascertained by looking at the tabulated values of χ^2 for given degrees of freedom at a certain level of significance. If the calculated value of χ^2 is equal to or exceeds the table value, the difference between the observed and expected frequencies is taken as significant, but if the table value is more than the calculated value of χ^2, then the difference is considered as insignificant i.e., considered to have arisen as a result of chance and as such can be ignored.

As already stated, degrees of freedom play an important part in using the chi-square distribution and the test based on it, one must correctly determine the degrees of freedom. If there are 10 frequency classes and there is one independent constraint, then there are (10 - 1) = 9 degrees of freedom. Thus, if 'n' is the number of groups and one constraint is placed by making the totals of observed and expected frequencies equal, the d.f would be equal to (n - 1). In the case of a contingency table (i.e., a table with 2 columns and 2 rows or a table with two columns and more than two rows or a table with two rows but more

than two columns or a table with more than two rows and more than two columns), the d.f is worked out as follows: d.f = (c - 1) (*r* - 1) where '*r*' means the number of columns and 'r' means the number of rows.

19.3.4 Conditions for the Application of 2 Test

The following conditions should be satisfied before 2 test can be applied:

1. Observations recorded and used are collected on a random basis.
2. All the items in the sample must be independent.
3. No group should contain very few items, say less than 10. In case where the frequencies are less than 10, regrouping is done by combining the frequencies of adjoining groups so that the new frequencies become greater than 10. Some statisticians take this number as 5, but 10 is regarded as better by most of the statisticians.
4. The overall number of items must also be reasonably large. It should normally be at least 50, howsoever small the number of groups may be.
5. The constraints must be linear. Constraints which involve linear equations in the cell
6. Frequencies of a contingency table (i.e., equations containing no squares or higher powers of the frequencies) are known as linear constraints.

19.4 USES OF CHI-SQUARE TEST

Chi-square is a statistical test that examines the differences between categorical variables from a random sample in order to determine whether the expected and observed results are well-fitting. Here are some of the uses of the Chi-Square test:

1. The Chi-square test can be used to see if your data follows a well-known theoretical probability distribution like the Normal or Poisson distribution or binomial distribution.
2. The Chi-square test allows you to assess your trained regression model's goodness of fit on the training, validation, and test data sets.
3. A chi-square test is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the

variables you are studying.

4. The chi-squared distribution is also used in hypothesis testing and to a lesser extent for confidence intervals for population variance when the underlying distribution is normal.

5. Although test is conducted in terms of frequencies it can be best viewed conceptually as a test about proportions.

6. χ^2 test is used in testing hypothesis and is not useful for estimation.

7. Chi-square test can be applied to complex contingency table with several classes.

8. Chi-square test has a very useful property i.e., 'the additive property'. If a number of sample studies are conducted in the same field, the results can be pooled together. This means that χ^2 - values can be added.

19.4.1 Important Characteristics of χ^2 Test

1. This test (as a non-parametric test) is based on frequencies and not on the parameters like mean and standard deviation.

2. The test is used for testing the hypothesis and is not useful for estimation.

3. This test possesses the additive property as has already been explained.

4. This test can also be applied to a complex contingency table with several classes and as such is a very useful test in research work.

5. This test is an important non-parametric test as no rigid assumptions are necessary in regard to the type of population, no need of parameter values and relatively less mathematical details are involved.

19.4.2 Caution in Using χ^2 Test

The chi-square test is no doubt a most frequently used test, but its correct application is equally an uphill task. It should be borne in mind that the test is to be applied only when the individual observations of sample are independent which means that the occurrence of one individual observation (event) has no effect upon the occurrence of any other observation (event) in the sample under consideration.

Small theoretical frequencies, if these occur in certain groups, should be dealt with under special care. The other possible reasons concerning the improper application or misuse of this test can be:

1. Neglect of frequencies of non-occurrence.

2. Failure to equalise the sum of observed and the sum of the expected frequencies.

3. Wrong determination of the degrees of freedom.

4. Wrong computations and the like. The researcher while applying this test must remain careful about all these things and must thoroughly understand the rationale of this important test before using it and drawing inferences in respect of this hypothesis.

19.5 STEPS INVOLVED IN CHI-SQUARE TEST

1. First of all calculate the expected frequencies on the basis of given hypothesis or on the basis of null hypothesis. Usually in case of a 2 × 2 or any contingency table, the expected frequency for any given cell is worked out as under:

$$\text{Expected frequency of any cell} = \frac{(\textbf{Row total for the row of that cell}) \times (\textbf{Column total for the column of that cell})}{(\textbf{Grand total}) / \textbf{N}}$$

1. Obtain the difference between observed and expected frequencies and find out the squares of such differences i.e., calculate $(O_{ij} - E_{ij})^2$.

2. Divide the quantity $(O_{ij} - E_{ij})^2$ obtained as stated above by the corresponding expected frequency to get $(Oij - Eij)^2/E_{ij}$ and this should be done for all the cell frequencies or the group frequencies.

3. Find the summation of $(O_{ij} - E_{ij})^2/E_{ij}$ values or what we call $c^2 = \sum \frac{(Oij - Eij)^2}{Eij}$. This is the required c^2 value. The c^2 value obtained as such should be compared with relevant table value of c^2 and then inference be drawn as stated above.

19.6 COMPUTATION OF CHI-SQUARE

Example 1: A die is thrown 132 times with following results:

Number turned up	1	2	3	4	5	6
Frequency	16	20	25	14	29	28

Is the die unbiased?

Solution: Let us take the hypothesis that the die is unbiased. If that is so, the probability of obtaining any one of the six numbers is 1/6 and as such the expected frequency of any one number coming upward is 132 ×1/6 = 22. Now we can write the observed frequencies along with expected frequencies and work out the value of χ^2 as follows:

Table 1: Calculation of Chi-square value and expected frequencies

No. turned up	Observed frequency (O)	Expected frequency (E)	(O –E)	$(O-E)^2$	$(O-E)^2/E$
1	16	22	-6	36	36/22
2	20	22	-2	4	4/22
3	25	22	3	9	9/22
4	14	22	-8	64	64/22
5	29	22	7	49	49/22
6	28	22	6	36	36/22

$$\therefore \chi^2 = \sum \frac{(Oij-Eij)^2}{Eij} = 9$$

Hence, the calculated value of $\chi^2 = 9$.

∵ Degrees of freedom in the given problem is

(n – 1) = (6 – 1) = 5.

The table value of χ^2 for 5 degrees of freedom at 5 per cent level of significance is 11.071. Comparing calculated and table values of χ^2, we find that calculated value is less than the table value and as such could have arisen due to fluctuations of sampling. The result, thus, supports the hypothesis and it can be concluded that the die is unbiased.

Example 2: Find the value of χ^2 for the following information:

Class	A	B	C	D	E
Observed frequency	8	29	44	15	4
Theoretical (or expected) frequency	7	24	38	24	7

Solution: Since some of the frequencies less than 10, we shall first re-group the given data as follows and then will work out the value of χ^2

Table 2

Class	Observed frequencies (O)	Expected Frequencies (E)	(O-E)	$(O-E)^2/E$
A+ B	37	31	6	36/31
C	44	38	6	36/38
D+E	19	31	-12	144/31

$$\therefore \chi^2 = \sum \frac{(Oij - Eij)^2}{Eij} = 6.76$$

Example 3: The table given below shows the data obtained during outbreak of smallpox:

	Attacked	**Not attacked**	**Total**
Vaccinated	31	469	500
Not vaccinated	185	1315	1500
Total	**216**	**1784**	**2000**

Test the effectiveness of vaccination in preventing the attack from smallpox. Test your result with the help of χ^2 at 5 per cent level of significance.

Solution: Let us take the hypothesis that vaccination is not effective in preventing the attack from smallpox i.e., vaccination and attack are independent. On the basis of this hypothesis, the expected frequency corresponding to the number of persons vaccinated and attacked would be:

$$\text{Expectation of (AB)} = \frac{\mathbf{A \times B}}{\mathbf{N}}$$

When A represents vaccination and B represent attack.

$\therefore$ (A) = 500

(B) = 216

N = 2000

Expectation of (AB) = 54

Table 2: Calculation of Chi-Square

Classes	O	E	O-E	$(O-E)^2$	$(O-E)^2/E$
(AB)	31	54	-23	529	529/54 = 9.796
(Aβ)	469	446	+23	529	529/446= 1.186
(αB)	158	162	+23	529	529/162= 3.265
(αβ)	1315	1338	-23	529	529/1338= 0.395

$$\therefore \chi^2 = \sum \frac{(Oij - Eij)^2}{Eij} = 14.642$$

Degrees of freedom in this case = $(r\text{-}1)\,(c\text{-}1) = (2\text{-}1)\,(2\text{-}1) = 1$

The table value of χ^2 for 1 degree of freedom at 5 per cent level of significance is 3.841. The calculated value of χ^2 is much higher than this table value and hence the result of the experiment does not support the hypothesis. We can, thus, conclude that vaccination is effective in preventing the attack from small pox.

19.7 LIMITATIONS OF THE TESTS OF HYPOTHESES

We have described above some important test often used for testing hypotheses on the basis of which important decisions may be based. But there are several limitations of the said tests which should always be borne in mind by a researcher. Important limitations are as follows:

(i) The tests should not be used in a mechanical fashion. It should be kept in view that testing is not decision-making itself; the tests are only useful aids for decision-making. Hence "proper interpretation of statistical evidence is important to intelligent decisions."

(ii) Test does not explain the reasons as to why do the difference exist, say between the means of the two samples. They simply indicate whether the difference is due to fluctuations of sampling or because of other reasons but the tests do not tell us as to which is/are the other reason(s) causing the difference.

(iii) Results of significance tests are based on probabilities and as such cannot be expressed with full certainty. When a test shows that a difference is statistically significant, then it simply suggests that the difference is probably not due to chance.

(iv) Statistical inferences based on the significance tests cannot be said to be entirely correct evidences concerning the truth of the hypotheses. This is specially so in case

of small samples where the probability of drawing erring inferences happens to be generally higher. For greater reliability, the size of samples be sufficiently enlarged.

All these limitations suggest that in problems of statistical significance, the inference techniques (or the tests) must be combined with adequate knowledge of the subject-matter along with the ability of good judgement.

19.8 SUMMARY

When samples are small, i.e., less than 30 the large sample results do not hold good for small samples, i.e., the assumption of approximate normality of the distribution is not true; in fact another distribution (exact distribution) of the test statistics is to be used and the result modified accordingly. Generally, the student's t-test, Z-test, Chi-square test and F-test are exact tests or small test of interest. In the present lesson we have studied discussed about Chi-square test. The Chi-square (χ^2) test represents a useful method of comparing experimentally obtained results with those to be expected theoretically on some hypothesis. Thus Chi-square is a measure of actual divergence of the observed and expected frequencies. It is very obvious that the importance of such a measure would be very great in sampling studies where we have invariably to study the divergence between theory and fact. Chi-square as we have seen is a measure of divergence between the expected and observed frequencies and as such if there is no difference between expected and observed frequencies the value of Chi-square is 0. If there is a difference between the observed and the expected frequencies then the value of Chi square would be more than 0. That is, the larger the Chi-square the greater the probability of a real divergence of experimentally observed from expected results. If the calculated value of chi-square is very small as compared to its table value it indicates that the divergence between actual and expected frequencies is very little and consequently the fit is good. If, on the other hand, the calculated value of chi-square is very big as compared to its table value it indicates that the divergence between expected and observed frequencies is very great and consequently the fit is poor. To evaluate Chi-square, we enter Table E with the computed value of chi- square and the appropriate number of degrees of freedom. The number of df = $(r - 1)(c - 1)$ in which r is the number of rows and c the number of columns in which the data are tabulated. : Thus in 2 x 2 table degrees of freedom are (2 - 1) (2 - 1) or 1. Similarly in 3 x 3 table, degrees of freedom are (3 - 1) (3 - 1) or 4 and in 3 x 4 table the degrees of freedom are (3 - 1) (4 - 1) or 6. The

calculated values of χ^2 (Chi-square) are compared with the table values, to conclude whether the difference between expected and observed frequencies is due to the sampling fluctuations and as such significant or whether the difference is due to some other reason and as such significant.

19.9 GLOSSARY

- **Confidence interval:** In statistics, a confidence interval (CI) is a type of interval estimate, computed from the statistics of the observed data that might contain the true value of an unknown population parameter. Most commonly, the 95% confidence level is used
- **Chi-square test:** It is a test used to measure the difference between observed and expected frequencies.
- **Chi-Square distribution**: In probability theory and statistics, the chi-squared distribution (also chi-square or χ^2-distribution) with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables
- **Expected Frequency**: The expected frequency is a probability count that appears in contingency table calculations including the chi-square test. An expected frequency is computed by multiplying the probability that an event occurs by the total number of possible times that the event could occur.
- **Observed Frequency:** Observed Frequencies are counts made from experimental data. In other words, you actually observe the data happening and take measurements. For example, you roll a die ten times and then count how many times each number is rolled.

19.10 SELF ASSESSMENT QUESTIONS/CHECK YOUR PROGRESS

Fill in the blanks:

1. The chi-square test is one of the simplest and most widely used .....tests.
2. The distribution of chi-square depends on the.....
3. The chi-square test should not be applied if n is less than.....

4. The greater the discrepancy between the observed and expected frequencies.....the value of chi-square.

5. The additive property of Chi-square test is.....

6. The overall number of items in case of chi-square test must also be reasonably.....

19.11 LESSON END EXERCISE

1. What are the basic conditions for the application of Chi-square test?

2. The following table shows the result of inoculation against cholera in a certain state:

	Not attacked	**Attacked**	**Total**
Inoculated	267	37	304
Not Inoculated	757	155	912
Total	**1024**	**192**	**1216**

3. The following results were obtained when two sets of items were subjected to two different treatments X and Y, to enhance their tensile strength.

(i) Treatment X was applied on 400 items and 80 were found to have gained in strength,
(ii) Treatment Y was applied on 400 items and 20 were found to have gained in strength,
(i) Is treatment Y superior to treatment X?

4. Point out the important limitations of tests of hypotheses. What precaution the researcher must take while drawing inferences as per the results of the said tests?

5. Briefly describe the important parametric tests used in context of testing hypotheses. How such tests differ from non-parametric tests? Explain.

6. Clearly explain how you will test the equality of variances of two normal populations.

19.12 SUGGESTED READING

- Levin, R.I. Robin, D.S. *Statistics for Management,* Prentice-Hall of India.New Delhi.
- Aczel, A. D. Sounderpandian, *J. Complete Business Statistics,* Mc Graw Hill Publishing. New Delhi. downloaded
- Anderson, S., W. *Statistics for Business and Economics,* Cengage Learning. New Delhi.
- Kazmeir L. J. *Business Statistics,* Tata Mc Graw Hill. New Delhi.
- Vohra, N. D. *Business Statistics,* Tata Mc Graw Hill. New Delhi.

MANN WHITNEY TEST AND KRUSKAL WALLIS TEST, ADVANTAGES AND DISADVANTAGES OF NON-PARAMETRIC TESTS

STRUCTURE

20.1 INTRODUCTION

Most of the statistical tests that we have discussed so far had the following two features in common:

(i) The form of the frequency function of the parent population from which the samples have been drawn is assumed to be known.

(ii) They are concerned with testing statistical hypothesis about the parameters of this frequency function or estimating its parameters.

For example, almost all the exact sample test of significance are based on the fundamental assumption that the parent population is normal and are concerned with testing or estimating the means and variances of these populations. Such tests, which deal with the population parameters, are known as parametric tests. Thus, a parametric statistical test is a test whose model specifies certain conditions about the parameters of the population from which the samples are drawn.

On the other hand, a non-parametric test is a test that does not depend on the particular form of the basic frequency function from which the samples are drawn. In other words, non-parametric test does not make any assumption regarding the form of the population.

However, certain assumptions associated with non-parametric tests are:

i. Sample observations are independent.

ii. The variable under study is continuous.

iii. Population distribution function is continuous.

iv. Lower order moments exist.

It is well known that these assumptions are fewer and much weaker than those associated with parametric tests. Also, in recent times statisticians have developed useful techniques that do not make restrictive assumptions about the shape of population distributions. These are known as distribution free or more commonly, non-parametric tests. The hypotheses of a non-parametric test are concerned with something other than the value of a parameter. A large number of these tests exist, nut this lesson will examine only few of the better known and more widely used ones.

1. A rank sum test, often called the Mann-whitney U test, which can be used to determine whether two independent samples have been drawn from the same population. It uses more information than the sign test.

2. Another, rank sum test, the kruskal wallis test, which generalises the analysis of variance, that enable us to dispense with the assumption that the populations are normally distributed.

20.2 OBJECTIVES

After successful completion of this lesson, students will be able to:

- Understand the role of non-parametric tests in hypothesis testing.
- Perform Mann Whitney test for hypothesis testing.
- Understand the Kruskal Wallis test or H-test.

20.5 MANN WHITNEY TEST

In a statistical test, two kinds of assertions are involved viz., an assertion directly related to the purpose of investigation and other assertions to make a probability statement. The former is an assertion to be tested and is technically called a hypothesis, whereas the set of all other assertions is called the model. When we apply a test (to test the hypothesis) without a model, it is known as distribution-free test, or the nonparametric test. Non-parametric tests do not make an assumption about the parameters of the population and thus do not make use of the parameters of the distribution. In other words, under non-parametric or distribution-free tests we do not assume that a particular distribution is applicable, or that a certain value is attached to a parameter of the population. Tests of hypotheses with 'order statistics' or 'nonparametric statistics' or 'distribution-free' statistics are known as nonparametric or distribution-free tests. The following distribution-free tests are important and generally used:

(i) Test of a hypothesis concerning some single value for the given data (such as one-sample sign test).

(ii) Test of a hypothesis concerning no difference among two or more sets of data (such as two-sample sign test, Fisher-Irwin test, Rank sum test, etc.).

(iii) Test of a hypothesis of a relationship between variables (such as Rank correlation, Kendall's coefficient of concordance and other tests for dependence.

(iv) Test of a hypothesis concerning variation in the given data i.e., test analogous to ANOVA viz., Kruskal-Wallis test.

(v) Tests of randomness of a sample based on the theory of runs viz., one sample runs test.

(vi) Test of hypothesis to determine if categorical data shows dependency or if two classifications are independent viz., the chi-square test.

Rank sum tests are a whole family of test, but we shall describe only two such tests commonly used viz., the U test and the H test. U test is popularly known as Wilcoxon-Mann-Whitney test, whereas H test is also known as Kruskal-Wallis test.

20.5.1 Wilcoxon-Mann-Whitney Test/Rank Sum Test/U-Test

This is a very popular test amongst the rank sum tests. This test is used to determine whether two independent samples have been drawn from the same population. It uses more information than the sign test or the Fisher-Irwin test. This test applies under very general conditions and requires only that the populations sampled are continuous. However, in practice even the violation of this assumption does not affect the results very much. To perform this test, we first of all rank the data jointly, taking them as belonging to a single sample in either an increasing or decreasing order of magnitude. We usually adopt low to high ranking process which means we assign rank 1 to an item with lowest value, rank 2 to the next higher item and so on. In case there are ties, then we would assign each of the tied observation the mean of the ranks which they jointly occupy. For example, if sixth, seventh and eighth values are identical, we would assign each the rank $(6 + 7 + 8)/3 = 7$. After this we find the sum of the ranks assigned to the values of the first sample (and call it R1) and also the sum of the ranks assigned to the values of the second sample (and call it R2). Then we work out the test statistic i.e., U, which is a measurement of the difference between the ranked observations of the two samples as under:

$$U = n_1 . n_2 + \frac{n1(n1+1)}{2} - R_1$$

where n_1, and n_2 are the sample sizes and R1 is the sum of ranks assigned to the values of

the first sample. (In practice, whichever rank sum can be conveniently obtained can be taken as R1, since it is immaterial which sample is called the first sample.) In applying U-test we take the null hypothesis that the two samples come from identical populations. If this hypothesis is true, it seems reasonable to suppose that the means of the ranks assigned to the values of the two samples should be more or less the same. Under the alternative hypothesis, the means of the two populations are not equal and if this is so, then most of the smaller ranks will go to the values of one sample while most of the higher ranks will go to those of the other sample. If the null hypothesis that the $n_1 + n_2$ observations came from identical populations is true, the said 'U' statistic has a sampling distribution with

$$\text{Mean} = \mu_U = \frac{\mathbf{n1 \cdot n2}}{\mathbf{2}}$$

and Standard deviation (or the standard error)

$$= \sigma_U = \sqrt{\boldsymbol{n1n2\,(n1 + n2 + 1)/12}}$$

If n_1 and n_2 are sufficiently large (i.e., both greater than 8), the sampling distribution of U can be approximated closely with normal distribution and the limits of the acceptance region can be determined in the usual way at a given level of significance. But if either n_1 or n_2 is so small that the normal curve approximation to the sampling distribution of U cannot be used, then exact tests may be based on special tables such as one given in the, appendix, showing selected values of Wilcoxon's (unpaired) distribution.

Example 1: The values in one sample are 53, 38, 69, 57, 46, 39, 73, 48, 73, 74, 60 and 78. In another sample they are 44, 40, 61, 52, 32, 44, 70, 41, 67, 72, 53 and 72. Test at the 10% level the hypothesis that they come from populations with the same mean. Apply U-test.

Solution: First of all we assign ranks to all observations, adopting low to high ranking process on the presumption that all given items belong to a single sample. By doing so we get the following:

Table 1

Size of sample item in ascending order	Rank	Name of related sample:[A for sample one and B for sample two]
32	1	B
38	2	A

39	3	A
40	4	B
41	5	B
44	6.5	B
44	6.5	B
46	8	A
48	9	A
52	10	B
53	11.5	B
53	11.5	A
57	13	A
60	14	A
61	15	B
67	16	B
69	17	A
70	18	B
72	19.5	B
72	19.5	B
73	21.5	A
73	21.5	A
74	23	A
78	24	A

9 + 11.5 + 13 + 14 + 17 + 21.5 + 21.5 + 23 + 24 = 167.5 and similarly we find that the sum of ranks assigned to sample two items or R_2 = 1 + 4 + 5 + 6.5 + 6.5 + 10 + 11.5 + 15 + 16 + 18 + 19.5 + 19.5 = 132.5 and we have n_1 = 12 and n_2 = 12

Hence, test statistic U = n1. n2 + (n1(n1+1))/2 - R1

= 12. 12 + (12(12+1))/2 - 167.5

= 12. 12+ (12(12+1))/2 - 167.5

= 144 + 12 *13/ 2 - 167.5

= 144 + 78- 167.5 = 54.5

Since in the given problem n1 and n_2 both are greater than 8, so the sampling distribution of U approximates closely with normal curve. Keeping this in view, we work out the mean

and standard deviation taking the null hypothesis that the two samples come from identical populations as under:

$$\text{Mean} = \mu_U = \frac{\mathbf{n1 \cdot n2}}{\mathbf{2}}$$

$$= 12 \times \frac{\mathbf{12}}{\mathbf{2}} = 72$$

$$= \sigma_U = \sqrt{\boldsymbol{n1n2\ (n1 + n2 + 1)/12}}$$

$$= \sqrt{\mathbf{12 \times 12\ (12 + 12 + 1)/12}} = 17.32$$

As the alternative hypothesis is that the means of the two populations are not equal, a two-tailed test is appropriate. Accordingly the limits of acceptance region, keeping in view 10% level of significance as given, can be worked out as under: As the z value for 0.45 of the area under the normal curve is 1.64, we have the following limits of acceptance region:

$$\text{Upper limit} = \mu_U + 1.64\ \sigma_U = 72 + 1.6417.32 = 100.40$$
$$\text{Lower limit} = \mu_U - 1.64\ \sigma_U = 72 - 1.64$$

As the observed value of U is 54.5 which is in the acceptance region, we accept the null hypothesis and conclude that the two samples come from identical populations (or that the two populations have the same mean) at 10% level. We can as well calculate the U statistic as under using R_2 value:

$$U = n_1 . n_2 + \frac{\boldsymbol{n1(n1+1)}}{\mathbf{2}} - R_2$$

$$= 12.\ 12 + \frac{\mathbf{12(12+1)}}{\mathbf{2}} - 132.5$$

$$= 12.\ 12 + \frac{\mathbf{12(12+1)}}{\mathbf{2}} - 132.5$$

$$= 144 + 12 * 13/\ 2 - 132.5$$

$$= 144 + 78 - 132.5 = 89.5$$

The value of U also lies in the acceptance region and as such our conclusion remains the same, even if we adopt this alternative way of finding U.

20.6 KRUSKAL WALLIS TEST/ H-Test

This test is conducted in a way similar to the U test described above. This test is used to test the null hypothesis that 'k' independent random samples come from identical universes against the alternative hypothesis that the means of these universes are not equal. This test

is analogous to the one-way analysis of variance, but unlike the latter it does not require the assumption that the samples come from normal populations or the universes having the same standard deviation. In this test, like the U test, the data are ranked jointly from low to high or high to low as if they constituted a single sample. The test statistic is H for this test which is worked out as under:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{K} \frac{R1^2}{ni} - 3(n+1)$$

where $n = n_1 + n_2 + ... + nk$ are the numbers in each of k samples, $N = n_1 + n_2 + n_3 + ... + n_k$ and $R_1, R_2, ...R_i$ being the sum of the ranks assigned to ni observations in the ith sample.

If the null hypothesis is true that there is no difference between the sample means and each sample has at least five items, then the sampling distribution of H can be approximated with a chi-square distribution with $(k - 1)$ degrees of freedom. As such we can reject the null hypothesis at a given level of significance if H value calculated, as stated above, exceeds the concerned table value of chi-square.

20.6.1 Assumptions for the Krystal Wallis Test

1. One independent variable with two or more levels (independent groups). The test is more commonly used when you have three or more levels. For two levels, consider using the Mann Whitney U Test instead.
2. Ordinal scale, Ratio Scale or Interval scale dependent variables.
3. Your observations should be independent. In other words, there should be no relationship between the members in each group or between groups. For more information on this point, see: Assumption of Independence.
4. All groups should have the same shape distributions. Most software (i.e. SPSS, Minitab) will test for this condition as part of the test.

Example 2: A shoe company wants to know if three groups of workers have different salaries:

Women: 23K, 41K, 54K, 66K, 78K.
Men: 45K, 55K, 60K, 70K, 72K
Minorities: 18K, 30K, 34K, 40K, 44K.
Use Kruskal Wallis test.

Solution 2:

Step 1: Sort the data for all groups/samples into ascending order in one combined set.

- 20K
- 23K
- 30K
- 34K
- 40K
- 41K
- 44K
- 45K
- 54K
- 55K
- 60K
- 66K
- 70K
- 72K
- 90K

Step 2: Assign ranks to the sorted data points. Give tied values the average rank.

Value	*Ranks*
20K	1
23K	2
30K	3
34K	4
40K	5
41K	6
44K	7
45K	8
54K	9
55K	10
60K	11
66K	12
70K	13
72K	14
90K	15

Step 3: Add up the different ranks for each group/sample.

Women: 23K, 41K, 54K, 66K, 90K = 2 + 6 + 9 + 12 + 15 = 44.
Men: 45K, 55K, 60K, 70K, 72K = 8 + 10 + 11 + 13 + 14 = 56.
Minorities: 20K, 30K, 34K, 40K, 44K = 1 + 3 + 4 + 5 + 7 = 20

Step 4: Calculate the H statistic:

$$H = \left[\frac{12}{n(n+1)} \sum_{j=1}^{c} \frac{T_j^2}{n_j} \right] - 3(n+1)$$

Where:

- n = sum of sample sizes for all samples,
- c = number of samples,
- Tj = sum of ranks in the jth sample,
- nj = size of the jth sample.

$$H = \left[\frac{12}{15(15+1)} \left[\frac{44^2}{5} + \frac{56^2}{5} + \frac{20^2}{5} \right] \right] - 3(15+1)$$

H = 6.72 9 (calculated value)

Step 5: Find the critical/ table value of chi-square value, with c-1 degrees of freedom. For 3 - 1 degrees of freedom and an alpha level of .05, the critical chi square value is 5.9915.

If the critical chi-square value (table value) is less than the H statistic, reject the null hypothesis that the medians are equal.

If the chi-square value is not less than the H statistic, there is not enough evidence to suggest that the medians are unequal.

In this case, (table value) 5.9915 is less than 6.72 (calculated value), so we can reject the null hypothesis.

20.3 NON PARAMETRIC TESTS

Parametric and No-parametric statistical tests are distinguished on (i) the basis of the scaling of the data and (ii) the assumptions regarding the sampling distribution of sample statistic. The use of parametric tests:

(i) Require the level of measurement attained on the collected data in the form of an interval scale or ratio scale.

(ii) Involve hypothesis testing of specified parameter values, and

(iii) Require assumptions about the population distribution, in particular, assumption of normality and whether standard deviation of sampling/population distribution is known or not.

If these assumptions are not justified then these tests would not yield accurate conclusions about population parameters. In such Circumstances, it is necessary to use few other hypothesis testing procedures that do not require these conditions to be met. These procedures are referred to as non-parametric tests. Non-parametric tests (i) do not depend on the form of the underlying population distribution from which the samples were drawn; and (ii) use data that are of insufficient strength, i.e., data are categorical scaled or ranks scaled.

20.3.1 Advantages of Non-Parametric tests

1) Non-parametric methods can be used to analyse categorical (nominal scaling) data, rank (ordinal scaling) data and interval (ration scaling) data.

   Non-parametric methods are generally easy to apply and quick to compute when sample size is small.

2) Non-parametric methods require few assumptions but are very useful when the scale of measurement is weaker than required for parametric methods. Hence, these methods are widely used and yield more general broad-based conclusions.

3) Non-parametric methods provide an approximate solution to an exact problem, whereas parametric methods provide an exact solution to an approximate problem.

4) Non-parametric methods provide solution to problems that do not require to make the assumption that a population is distributed normally or any specific shape.

20.3.2 Disadvantages of Non-Parametric tests

1) Non-parametric methods should not be used when all the assumptions of the parametric methods can be met. However, they are equally powerful when

assumptions are met, when assumptions are not met these may be more powerful.

2) Non-parametric methods require more manual computational time when sample size gets larger.

3) Table values for non-parametric statistics are not as readily available as parametric methods.

4) Non-parametric tests are usually not as widely used and not well known as parametric tests.

20.7 SUMMARY

Non-parametric test is a test that does not depend on the particular form of the basic frequency function from which the samples are drawn. In other words, a non-parametric test does not make any assumption regarding the form of the population. In this lesson, we have studied Non-parametric tests and their advantages and disadvantages.

20.8 GLOSSARY

- **Non-parametric test**: Nonparametric statistics refer to a statistical method in which the data is not required to fit a normal distribution. Nonparametric statistics uses data that is often ordinal, meaning it does not rely on numbers, but rather a ranking or order of sorts
- **Parameter:** A numerical or other measurable factor forming one of a set that defines a system or sets the conditions of its operation, a limit.
- **Statistics:** A statistic is a characteristic of a sample. Generally, a statistic is used to estimate the value of a population parameter
- **Mann-Whitney test:** Mann-Whitney U test is the non-parametric alternative test to the independent sample t-test. It is a non-parametric test that is used to compare two sample means.
- **Kruskal Wallis test:** The Kruskal-Wallis test by ranks, Kruskal-Wallis H test (named after William Kruskal and W. Allen Wallis), or one-way ANOVA on ranks is a non-parametric method for testing whether samples originate from the same

distribution. It is used for comparing two or more independent samples of equal or different sample sizes.

20.9 SELF ASSESSMENT QUESTIONS

Multiple Choice Questions:

1. Non parametric technique equivalent to one-way ANOVA is known as:

   a. Mann-Whitney test b. Wilcoxon signed Rank test

   c. Kruskal Wallis Test d. Friedman's ANOVA

2. Which of the following tests would be an example of a non parametric method?

   a. Z test b. T test

   c. Sign test d. All of these

3. A collection of statistical methods that generally requires very few, if any assumptions about the population distribution is known as:

   a. Parametric methods b. Non- Parametric methods

   c. Semi Parametric methods d. None of the above

4. Statistical Power is measured as a probability that equals:

   a. $1 - \beta$ b. $1 + \beta$

   c. B d. $1/ \beta$

5. The probability of rejecting the null hypothesis when it is true is called as:

   a. Level of confidence b. Level of significance

   c. Power of a test d. None of these

20.10 LESSON END EXERCISE

1. What are non-parametric tests? In what ways are they different from parametric tests?

2. Point out the advantages and limitations of non-parametric tests?

3. Explain Mann-Whiteny U-test with the help of an example.

5. Test the hypothesis of no difference between the ages of male and female employees of a certain company using the Mann-Whitney U test for the sample data. Use the 1% level of significance.

   Males 31 25 38 33 42 40 44 26 43 35
   Females 44 30 34 47 35 32 35 47 48 34

6. The following table shows sample retail prices for three brands of shoes. Use the kruskal-wallis test to determine whether there is any difference among the retail prices of the brands throughout the country. Use the 0.01 level of signifinace.

   Brand A $89 90 92 81 76 88 85 95 97 86 100
   Brand B $78 93 81 87 89 71 90 96 82 85
   Brand C $80 88 86 85 79 80 84 85 90 92

20.11 SUGGESTED READING

- Levin, R.I. Robin, D.S. *Statistics for Management,* Prentice-Hall of India.New Delhi.
- Aczel, A. D. Sounderpandian, *J. Complete Business Statistics,* Mc Graw Hill Publishing. New Delhi. downloaded
- Anderson, S., W. *Statistics for Business and Economics,* Cengage Learning. New Delhi.
- Kazmeir L. J. *Business Statistics,* Tata Mc Graw Hill. New Delhi.
- Vohra, N. D. *Business Statistics,* Tata Mc Graw Hill. New Delhi.