STATISTICAL METHODS

M. A. Economics First Year

Semester – II, Paper-V

Director, I/c

Prof. V.VENKATESWARLU

MA., M.P.S., M.S.W., M.Phil., Ph.D.

CENTRE FOR DISTANCE EDUCATION

ACHARAYANAGARJUNAUNIVERSITY

NAGARJUNANAGAR – 522510

Ph:0863-2346222,2346208,

0863-2346259(Study Material)

Website: www.anucde.info

e-mail:anucdedirector@gmail.com

205EC21: STATISTICAL METHODS

MODULE 1 : SAMPLING METHODS

Concept of sampling - random and non-random sampling; Simple random; Stratified random, systematic sampling, cluster sampling and non-random sampling methods.

MODULE 2 : CORRELATION AND REGRESSION

Correlation and regression analysis and their properties; Concept of the least squares and the lines of regression and applications.

MODULE 3 : TIME SERIES ANALYSIS

Introduction - components - measurement of trend - graphic, (Free hand curve fitting) method, method of semi average, method of moving average, method of curve fitting by principle of least squares.

MODULE 4 : PROBABILITY

Deterministic and non-deterministic relationships - Terminology - Some basic concepts of set theory - Probability defined - Theorems of probability - Conditional probability - Bayes theorem and inverse probability - Joint and marginal probabilties - Review Exercises.

MODULE 5 : THEORETICAL DISTRIBUTIONS

Binomial Distribution and Poisson Distibution - Assumptions constants - Normal Distribution - properties of normal distribution, constants of normal distribution - Review Exercises.

READING LIST:

1. S.C. Gupta Fundamentals of statistics.
2. K.Chandra Sekhar, Business of statistics.
3. K.V.Sarma, Statistics made simple, Prentice Hall of India.

STATISTICAL METHODS – 205EC21

CONTENTS

1. SAMPLING

Objectives

After completion of this chapter, you should be able to:

- Understand the basic idea of sampling:
- Know the differences between census and sample:
- Identify the principles of sampling
- Understand about the sampling and non- sampling errors.

Structure

1.1 Introduction:-

The science of statistic may be broadly classified under the following two headings: (a) Descriptive

(b) Inductive

The descriptive statistics which consists in describing some characteristics of the numerical data. The inductive statistics, also known as statistical inference, may be termed as the logic of drawing statistically valid conclusions about the totality of cases or items termed as population, in any statistical investigations on the basis of examining a part of the population, termed as sample, and which is drawn from the population in scientific manners. In modern „decision making process" in different fields of human activity, including the ordinary actions of our daily life, most of our decisions and attitudes depend very much upon the inspection or examination of only a few objects or items out of the total lot. This process of studying only the sample data and then generalizing the results to the population (i.e drawing inferences about the population on the basic of sample study) involves an element of risk, the risk of making wrong decisions.

1.2 Sampling

A finite subset of the population, selected from it with the objective of investigating its properties is called a sample and the number of units in the sample is known as the sample size. Sampling is a tool which enables us to draw conclusions about the characteristics of the population after studying only those objective or items that are included in the sample.

The main objectives of the sampling theory are:

(i) To obtain the optimum results, i.e., the maximum information about the characteristics of the population with the available sources at our disposal in terms of time, money and manpower by studying the sample values only.

(ii) To obtain the best possible estimates of the population parameters. From times immemorial, people have been using it without knowing that some scientific procedure has been used in arriving at the conclusions. On inspecting the sample of a particular stuff, we arrive at a conclusion about accepting or rejecting it. For example, the consumer examines only a handful of the rice, pulsar or any commodity in a shop to assess its quality and then decides to buy it or not. The housewife usually tastes a spoonful of the cooked products to as certain if it is properly cooked and also to see if it contains proper quantity of salt or sugar. The consumer as certain the quality of the grapes by testing one or two from the seller"s basket. The intelligence of the individuals in a subject is estimated by the university by giving them a 3 -hour test. A business man orders for the products after examining only a sample from it. In fact, the entire business is done on the basis of display of a few specimen samples only.

1.3 Parameter and statistic:

The statistics constant of the population like mean (μ), variance (σ^2), skewness (β_1), kurtosis (β_2), moments (μ_r), correlation coefficient ($\int$), etc., are known as parameters. We can compute similar statistical constants for the sample drawn from the given population. Prof. R.A. fisher termed the statistical constants of the sample like mean ($\bar{x}$), variance (S^2), Skewness (b_1) kurtosis (b_2), moments (M_r), correlation coefficient(r) etc., as statistics obviously, parameters are function of the population values while statistics are functions of the sample observations.

The sample mean ($\bar{x}$) and variance (S^2) are given by:

$$\bar{x} = 1/n\ (x_1+x_2 + \ldots..+x_n) = 1/n \sum_{i=1}^{n_1} x_i$$

$$S^2 = 1/n\,[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots\ldots (x_n - \bar{x})^2] = 1/n \sum_{i=1}^{n_1} (x_i - \bar{x})^2$$

The population parameters are unknown and their estimates provided by the appropriate sample statistics are used. The sample statistics are functions of the sample observations and vary from sample.

1.4 : Principles of sampling:

The fact that the characteristics of the sample provide a fairly good idea about the population characteristics is born out by the theory of probability.

(i) Law of statistical regularity;

This law has its origin in the mathematical theory of probability. In the words of L.R.Conner, "the law of statistical regularity lays down that a group of objects chosen at random from a larges group tends to possess the characteristics of that large group (universe)". According to king" the law of statistical regularity lays down that the moderately large number of items chosen at random from a large group are almost sure on the average to possess the characteristics of the large group.

The principle of statistical regularity impresses upon the following two points:

(a) Large sample size: logically it seems that as the sample size increases the sample is more likely to reveal the true characteristics of the population and thus provide better estimates of the parameters. It is known that the reliability of the sample statistic as an estimate of the population parameters is proportional to the square root of the sample size n. But due to certain limitations in terms of time, money and man power, it is not always possible to take very large samples. Moreover, the effort and cost of drawing large samples might outline the utility of the sample study as against the complete enumeration (censes).

b) Random Selection: the sample should be selected at random from the population. by random selection we mean a selection in which each and every unit in the population has an equal chance of being selected has an equal chance of being selected in the sample

ii) Principle of Inertia of Large Numbers:

An immediate deduction from the principle of statistical regularity is the principal of inertia of large numbers which states, "other things being equal as the sample size increases, the results tend to be more reliable and accurate." This is based on the fact that the behavior of a phenomenon on mass i.e. on a large scale is generally stable. By this we mean that if individual events are observed, their behaviors may be erratic and unpredictable but when a large number of events are considered, they tend to behave in a stable pattern.

(iii) Principle of Persistence of Small Numbers:

If some of the items in a population possess markedly distinct characteristics form the remaining items, then this tendency would be revealed in the sample values also. Rather this tendency of persistence will be there even if the population size is increased or even in the case of large samples. For example, if the day"s, production of any manufacturing unit is made 4 times, the proportion of defectives in the lot remains more or less same. This means that the number of defectives in the lot will also increase more or less in the same proportion.

iv) Principle of validity:

A sampling design is termed as valid if it enables us to obtain valid tests and estimates about the population parameters. This principle is satisfied by the samples drawn by the technique of probability sampling

v) Principle of optimization:

This principal stresses the need of obtaining optimum results in terms of efficiency and cost of the sampling design with the sources available at our disposal. As has been pointed out earlier, a measure of efficiency or reliability of an estimate of the population parameter is provided by the reciprocal of the standard error of the estimate and the cost of the design is determined by the total expenses incurred in terms of money and manpower. This principles aims at:

a) obtaining a desired level of efficiency at minimum cost and
b) obtaining maximum possible efficiency with given level of cost.

1.5 Census Versus Sample Enumeration:

For any statistical enquiry in any field of human activity, whether it is in business, economics or social sciences, the basic problems is to adequate and reliable data relating to the particular phenomenon under study. There are two methods of collecting the data:

(i) The Census Method or Complete Enumeration
(ii) The Sample Method or Partial Enumeration

Census Method:

In the census method we report to 100% inspection of the population and enumerate each and every unit of the population. In the sample method we inspect only a selected representative and adequate fraction (finite subset) of the population and after analyzing the results of the sample data we draw conclusions about the characteristics of the population.

The census method seems to provide more accurate and exact information as compared to sample enumeration as the information is collected from each and every unit of the population. Moreover, it affords more extensive and detailed study. The census method has its obvious limitations and draw backs as follows:

(i) The complete enumeration of the population requires lot of time, money, manpower and administrative personnel. As such this method can be adopted only by the government and big organizations who have vast resources at their disposal.
(ii) Since the entire population is to be enumerated, the census method is usually very time consuming. If the population is sufficiently large, then it is possible that the processing and the analysis of the data might take so much time that when the results are available they are not of much use because of changed conditions.

1.6 Sample Method:

The sample method has a number of distinct advantages over the complete enumeration method. Prof. R.A. Fisher sums up the advantages of sampling techniques over complete census in just four words: Speed, Economy, Adaptability and scientific Approach. A properly designed and carefully executed sampling plan yields fairly good results, often better than those obtained by the census method. The merits of the sample method over the census method are

(1) Speed:- i.e less time, since only a part of the population is to be inspected and examined the sample method results in considerable amount of saving in time and labour. There is saving in time not only in conducting the sampling enquiry but also in the processing, editing and analyzing the data.
(2) Economy: i.e Reduced Cost of the enquiry. The sample method is much more economical than a complete census. In a sample enquiry, there is reduction in the cost of collection of the information, administration, transport, training and man hours. Al though, the labour and the expanses of obtaining information per unit are generally large in a sample enquiry than in the census method, the overall expenses of a sample survey are relatively much less.
(3) Administrative Convenience: A complete census requires a very huge administrative set up involving lot of personnel, trained investigations and above all the co -ordination

between the various operating agencies. On the other hand, the organization and administration of a sample survey is relatively much convenient as it requires less personnel staff and the field of enquiry is also limited.

(4) Reliability: In the census, the sampling errors are completely absent in the non – sampling errors are also absent the results could be 100% accurate. On the other hand a ample enquiry contains both sampling and non-sampling errors. Inspire of this weakness, a carefully designed and scientifically executed sample survey gives results which are more reliable than those obtained from a complete census.

(5) Greater Scope: It appears that there is possibility of obtaining detailed information only in a complete census where each and every unit in the population enumerated. But in practice because of our limitations in any statistical enquiry in terms of time, money and man hours and because of the fact that sampling procedure results in considerable savings in time, money and labour. It is possible to obtain more detailed and exhaustive information from the limited few unite selected in the sample.

(6) Infinite or Hypothetical Population: If the population is infinite or too large, then sampling procedure. Is the only way of estimating the parameters of a population. For instance, the number of fish in the sea or the number of wild elephants in a dense forest can be estimated only by sampling method.

(7) Destructive Testing: If the testing of units is destructive i.e if in the course of inspection the units are destroyed or affected adversely, then we are left with no others way but to resort to sampling.

1.7 Limitations of Sampling:

The sampling procedure has its limitations and problems which are enumerated below:

i) If a sample survey is not properly planned (or designed) and executed carefully, the results obtained will not be reliable and quite often might even be misleading. In this context, it may be worthwhile to quote the words of Frederick to Stephen “Samples are like medicines. They can be harmful when they are taken carelessly or without knowledge of their effects. Every good sample should have a proper lable with instruction about its use.”

Sampling design must be perfect otherwise it might lead to serious complications in the final results. The omission of a few units in a complete census may be immaterial but non-response or incomplete response from even one or two units in a small sample might have a significant effect on the final result.

ii) An efficient sampling scheme requires the services of qualified, skilled and experienced personnel, better supervision and more sophisticated equipment and statistical techniques for the planning and execution of the survey and for the collection processing and analysis of the sample data. In the absence of these, the results of the survey may not be reliable.

iii) Sometimes the sample survey might require more time, money and labour than a complete census. This will be so if the sample size is a large proportion if the population size and if complicated weighted system is used.

iv) Sampling procedure cannot be used if we want to obtain information about each unit of the population. Further, if the population is too heterogeneous, it may be impossible to use a sampling procedure.

v) Each sampling procedure has its own limitations.

1.8 : Principal steps in a Sample Survey:

The following are the principal steps in the planning and execution of the sample surveys:

1) Objectives and Scope of the Survey: As in any statistical investigation, the first step in organizing a sampling survey is to define in clear and concrete terms the objectives and

scope of the survey. This is of immense help in deciding about the type of data to be collected and also the statistical techniques to be used for the processing and analysis of the data. In the absence of the purpose of the enquiry being explicitly specified, we are bound to collect some irrelevant information which is never used subsequently and also omit some important information which will ultimately lead to fallacious conclusions and wastage of sources.

2) Defining the population to be sampled: the population from which units are to be sampled should be defined clearly without any ambiguity. For example, in field experimentation, the field should be clearly defined in term of the shape, size, etc., keeping in mind the boards line cases so that nothing is left at the discretion of the investigator.

3) The frame and sampling units: the population must be capable of division into what are called sampling units. In a sample survey, the sampling units are the units into which the population to be sampled is divided. These units are units of enumeration i.e, the units on which the observations are to be made. It may be an individual person, a household, a family, a farm, a shop, a firm, a livestock or a block in a locality.

In order to draw a sample of villages in a state, we must have a map of the districts and villages of that state; for households, we must have a list of blocks in the locality; for selecting a group of students in a college, the list of students enrolled in the college is needed. This map, list or other acceptable material which serves as a guide for the population to be covered is known as the Frame. The frame may not contain detailed information about the sampling units. What is desired is that is should have at least enough information so as to enable us to identify and locate sampling units properly for statistical investigation. An up to date and good frame is very important to obtain efficient results in a sample survey because the structure of the sample survey is determined by the frame. A frame which is routinely prepared for some purpose should be used only after careful scrutiny and examination as it is usually found to be incomplete or it contain an unknown amount of duplication. Before using such a frame it should be ascertained that it is up – to – date, and free from these defects of incompleteness and duplication, if it is not up – to – date, it should be made so before using it.

(4) Data to be collected: The decision about the type of the data to be collected should be taken keeping in view the nature, objectives and the scope of the survey, the time and finances at our disposal and the degree of accuracy aimed at in the final results. Attempt should be made to eliminate the collection of irrelevant and unnecessary data which are never used subsequently and to ensure that no important or essential information is omitted. An outlying of the tables needed for the results of the survey is quite helpful in this regard.

(5) The Questionnaire or Schedule: After deciding about the nature of statistics to be collected, the next step is the preparation of questionnaire to be filled by the respondents or schedule of enquiry to be completed by the investigators after interviewing people for collecting the requisite information. The drafting of the questionnaire is a highly specialized job and requires great skill, wisdom, case, efficiency and experience. The questionnaire or schedule should be designed and drafted with at most care and caution, keeping in view the knowledge, understanding and the general educational level of respondents.

(6) Method of Collecting Information: there are two methods commonly used for obtaining numerical data for human population:

(i) Direct Personal Investigation or Interview Method.

(ii) Mailed Questionnaire Method.

A choice between the two methods depends on the objectives of the enquiry, the expenses involved and the accuracy of the results aimed at.

(7) Non – Respondents:-Due to Certain Practical Problems, it may not be possible to obtain information from each of the sampled units.

(8) Selection of Proper Sampling Design: A judicious decision about the sampling plan to be adopted is of paramount importance in the planning and execution of a sample survey. From among the several available sampling Designs like Simple Random Sampling, Stratified Random Sampling, Systematic Sampling etc.

(9) Organization of field work: to obtain reliable results in a sample survey, the sampling errors should be minimized. For this it is essential that the field work is properly organized and the personnel engaged in the conducting and execution of the survey should be properly trained to handles the problems of Survey like locating the sample units, use of the equipment and recording of the observations, the methods of collecting the desired information dealing with non – response, etc. it is also desirable to provide for adequate and frequent supervisory check on the field work.

(10) Pilot Survey or Pre – Test: From practical point of view it is found useful to conduct a pre –test or a guiding survey known as pilot survey, on a small scale before starting the main surrey. This is done to try out the questionnaire and the field methods for obtaining the general information about the population to be sampled. The information supplied by the pilot survey helps in:

(i) Estimating the cost of the sample survey and also the time needed for the availability of the results.

(ii) Improving the organization of the field work by removing the defects or faults observed in the pilot survey.

(iii) Formulating effective methods of asking questions and also in the improvement of the questionnaire.

(iv) Training of field staff.

(v) Disclosing certain problems or troubles that may otherwise be of a serious nature in a large scale main survey.

(11) Summary and Analysis of the Data:- After the planning and execution of the sample survey, the last step is the analysis of the collected data. It basically involves the following steps:

(i) Scruting and Editing of the data

(ii) Tabulation of data

(iii) Statistical analysis

(iv) Reports, Summary and conclusions

1.9 Sampling and Non – Sampling Errors:- the inaccuracies or errors in any statistical investigation, i.e in the collection, processing, analysis, and interpretation of the data may be broadly classified as follows: (i) Sampling Errors and (ii) Non- Sampling Errors.

(i) Sampling Errors:- In a sample survey, since only a small portion of the population is studied, its results are bound to differ from the census results and thus have a certain amount of error. This error would always be there no matter that the sample is drawn at random and that it is highly representative. This error is attributed to fluctuations of sampling and is called sampling error. Sampling error is due to the fact that only a subset of the population i.e., sample has been used to estimate the population parameters and draw inferences about the population. Thus, sampling error is present only in a sample survey and is completely absent in census method.

Sampling errors are primarily due to the following reasons:

(1) Faulty selection of the sample: Some of the bias is introduced by the use of defective sampling technique for the selection of a sample, e.g. purposive or judgment sampling in

which the investigator deliberately selects a representative sample to obtain certain results. This bias can be overcome by strictly adhering to a simple random sample or by selecting a sample at random subject to restrictions which while improving the accuracy are of such nature that they do not introduce bias in the results.
(2) Substitution: If difficulties arise in enumerating a particular sampling unit included in the random sample, the investigators usually substitute a convenient member of the population. This obviously leads to some bias since the characteristics possessed by the substituted unit will usually be different from those possessed by the unit originally included in the sample.
(3) Faulty demarcation of sampling units: Bias due to defective demarcation of sampling units is particularly significant in area surveys such as agricultural experiments in the field or crop cutting surveys, etc. In such surveys, while dealing with border-line cases, it depends more or less on the discretion of the investigator whether to include them in the sample or not.
(4) Error due to bias in the estimation method: Sampling method consists in estimating the parameters of the population by appropriate statistics computed from the sample.
(5) Variability of the population: Sampling error also depends on the variability or heterogeneity of the population to be sampled.
ii) Non-Sampling Errors:

Non – Sampling errors are not attributed to chance and are a consequence of certain factors which are within human control. In other words, they are due to certain causes which can be traced and may arise to any stage of the enquiry like planning and execution of the survey and collection, processing and analysis of the data. Non – sampling errors are thus present both in census surveys as well as sample surveys. Obviously, non – sampling errors will be of large magnitude in a census survey than in a sample survey because they increase with the increase in the number of units to be examined and enumerated. It is very difficult to prepare an exhaustive list of all the sources of non – sampling errors. We enumerate below some of the important factors responsible for non-sampling errors in any survey (census or sample).
(i) Faulty planning, including vague and faculty definitions of the population or the statistical units to be used, incomplete list of population – members (i.e., incomplete frame in case of sample survey).
(ii) Vague and imperfect questionnaire which might result in the incomplete or wrong information.
(iii) Defective methods of interviewing and asking questions.
(iv) Vagueness about the type of the data to be collected.
(v) Exaggerated or wrong answers to the questions which appeal to the pride or prestige or self – interest of the respondents. For example, a person may over -state his education or income or understate his age or he may give wrong statements to safeguard his self- interest.
(vi) Personal bias of the investigator
(vii) lack of trained and qualified investigators and lack of supervisions staff.
(viii) Failure of respondents memory to recall the events or happenings in the past.
(ix) Non – response and inadequate or incomplete Response – bias due to non – response results if in a house – to – house survey the respondent is not available inspire of repeated visits by the investigator or if the respondent refuses to furnish the information. Incomplete response error is introduced if the respondent is unable to furnish information on certain questions or if he is unwilling or even refuses to answer certain questions.
(x) Improper Coverage – if the objects of the survey are not precisely stated in clear cut terms, this may result in

(1) The inclusion in the survey of certain units which are to be excluded, or
(2) The exclusion of certain units which were to be included in the survey under the objectives.

For example, in a census to determine the number of individuals in the age group, say, 15 Years, to 55 years more or less serious errors may occur in deciding whom to enumerate unless particular community or area is not specified and also the time at which the age is to be specified.

(xi) Compiling Errors, i.e., wrong calculations or entries made during the processing and analysis of the data. Various operations of data processing such as editing and coding of the responses, punching of cards, tabulation and summarizing the original observations made in the survey are a potential source of error. Compilation errors are subject to control through verification, consistency checks, etc.

(xii) Publication Errors i.e., the errors committed during presentation and printing of tabulated results are basically due to two sources. The first refers to the mechanics of publication – the proofing error and the like. The other, which is of a more serious nature, lies in the failure of the survey organization to point out the limitations of the statistics.

1.10 Summary:

Sampling theory is a study of relationships existing between a population and samples drawn from the population. Sampling theory provides the tools and techniques for data collection keeping in mind the objectives to be fulfilled and nature of population. The sampling process comprises several stages. Sampling errors and biases are induced by the sample design. Non-sampling errors are other errors which can impact the final survey estimates, caused by problems in data collection, processing, or sample design. It therefore, becomes essential to draw inferences for the total population based on the analysis carried out on some of its members, information about whom can be collected easily and their selection itself can be structured in a definite way, so that this sample analysis can be relied upon and deployed for the total population.

1.11 Self – Assessment Questions

(1) Define clearly about the Population, Sample, Parameter and Statistics with an example.
(2) Explain about the various principles of Sampling.
(3) State clearly about the merits of the Sampling method over the census method.
(4) Write about the limitations of Sampling
(5) Describe about the principle steps involved in the planning and execution of the sample surveys.
(6) Distinguish between the Sampling and Non – sampling errors.

Reference Books:

1. S.C Gupta: Fundamentals of Statistics
2. K. Chandra Sekhar: Business Statistics
3. K. V Sarma: Statistics made simple, Prentice Hall of India

Lesson Writer
Dr. J. Pratapa Reddy

2. RANDOM SAMPLING METHODS

OBJECTIVES

- After completion of this chapter, you should be able to:
- Understand the concept of simple random sampling
- Understand the concept of stratified random sampling
- Know the difference between simple and stratified random sampling;
- Describe about the systematic sampling.

Structure

2.1 Introduction:

The choice of an appropriate sampling design is of paramount importance in the execution of a sample survey and is generally made keeping in view the objectives and scope of the enquiry and the type of the universe to be sampled. The sampling techniques may be broadly classified as:

(i) Purposive or judgment or subjective sampling
(ii) Probability of sampling
(iii) Mixed sampling

Based on the available information or data, we use the above sampling techniques. Some techniques are based on the concept of theory of probability.

2.2 Simple random sampling:

Simple Random Sampling (SRS) is the technique in which the sample is so drawn that each and every unit in the population has an equal and independent chance of being included in the sample. If the unit is selected in any draw is not replaced in the population before making the next draw, then it is known as simple random sampling without replacement (srswor) and if it is replaced back before making the next draw, then the sampling plan is called simple random sampling with replacement (srswr). Thus, simple random sampling with replacement always amount to sampling from an infinite population, even though the population is finite.

The important feature of srswor is that, "the probability of selecting a specified unit of the population at any given draw is equal to the probability of its being selected at the first draw". This implies that in srswor from a population of size N, the probability that any sampling unit is included in the sample is 1/N and this probability remains constant throughout the drawing.

Mathematically, if Er is the event that any specified unit is selected at the rth draw, then

$P(E_r) = 1/N, (r = 1, 2,\ldots..,n)$

Where „n" is the sample size. In particular it implies $P(E_r) = 1/N, = P(E_1)$,

i., e, the chance of selection of any specified item is same at any draw as it was in the first draw, are 1/N

since a unit can be selected in the sample at any one of the r exhaustive and mutually disjoint draws, (r = 1,2,...., n) by addition theorem of probability, the chance of any item being included in a sample of size n is:

$$\sum_{r=1}^{n} P(Er) = \sum_{r=1}^{n} (1/N) = n/N$$

Alternative Definition of srswor: If a sample of size n is drawn without replacement from a population of size N then there are ${}^{N}C_{n}$ possible samples. Simple random sampling is the technique of selecting the sample so that each of these ${}^{N}C_{n}$ samples has an equal chance or probability of being selected in the sample.

$$P = \frac{1}{N_{C_n}}$$

If sampling is done with replacement, then there are N^n possible samples of size n. In this case, simple random sampling (srswor) gives equal chance

$$P = \frac{1}{N^n}$$

For each of the N^n samples to be selected.

2.3 Selection of a simple Random sample:

Proper case must be exercised to ensure that the sample drawn is random and there fore, representative of the population. A random sample may be selected by: 1. Lottery Method

2. Use of table of Random Numbers.

1. Lottery Method: The simplest method of drawing a random sample is the lottery system. This consists in identifying each and every member or unit of the population with a distinct number which is recorded on a slip or a card. These slips should be as homogeneous as possible in shape, size, color etc. to avoid the human bias. The lot of these slips are cards is a kind of miniature of the population for sampling purposes. If the population is small, then these slips are put in a bag and thoroughly shuffled and then as many slips as units needed in the sample are drawn one by one, the slips being thoroughly shuffled after each draw. The sampling units corresponding to the numbers on the selected slips will constitute a random sample. For example, let us suppose that we want to draw a random sample of 10 individuals from a population of 100 individuals. We assign the numbers 1 to 100, one number to each individual of the population and prepare 100 identical slips bearing the numbers from 1 to 100. These slips are then placed in a bag or container and shuffled thoroughly. Finally, a sample of 10 slips is drawn out one by one. The individuals bearing the numbers on these selected slips will constitute the desired sample.

If the population to be sampled is fairly large, then we may adopt the lottery method in which all the slips or cards are placed in a metal cylinder which is thrown into a large rotating drawn working under a mechanical system. The rotation of the drawn results in through mixing or randomization of the cards. Then a sample of desired size „n" is drawn out of the container mechanically and the corresponding n sample units constitute the desired random sample.

The lottery method gives a sample which is quite independent of the properties of the population. It is one of the best and most commonly used methods of selecting random samples. It is quite frequently used in the random draw of prizes, in the tambola games and so on.

In sampling with replacement (srswr) each card drawn is replaced back in the container before making the next draw.

But in sampling without replacement (srswor) cards once drawn are not returned back. Since cards are drawn one by one, a through mixing is required before the next draw.

Use of Table of Random Numbers:

The lottery method described above is quite time consuming and cumbersome to use if the population to be sampled is sufficiently large. More over, in this method, it is not humanly possible to make all the slips or cards exactly alike and as such some bias is likely to be introduced. Statisticians have avoided this difficulty by considering the random sampling number series. Most these series are the results of actual sampling operations recorded for future use. The most practical and inexpensive method of selecting a random sample consists in the use of „Random Number tables" which have been so constructed that each of the digits 0, 1, 2…..9 appears with approximately the same frequency and independently of each other. If we have to select a sample from a population of size N (≤ 99), then the numbers can be combined two by two to give pairs from 00 to 99. Similarly if N ≤ 999 or N ≤ 9999 and so on. Since each of the digits 0, 1, 2,…. ,9 occurs with approximately the same frequency and independently of each other,

so does each of the pairs 00 to 99, triplets 000 to 999 or quadruplets 0000 to 9999 and so on.

The method of drawing a random sample comprises the following steps:

i. identify N units in the population with the numbers 1 to N.

ii. Select at random, any page of the „random number table‟ and pick up the numbers in any row, column or diagonal at random.

iii. The population units corresponding to the numbers selected in step ii, constitute the random sample.

The different sets of random numbers commonly used in practice are

1. Tippet‟s (1927) Random number tables (Tracts for computers No. 15, Cambridge University Press).
2. Fisher and Yates (1938) Tables (in statistical Tables for Biological, Agricultural and Medical Research).
3. Kendall and Babington smiths (1939) random tables.
4. Rand corporation (1955), (Free Press, Illinois) random number tables.
5. Table of Random numbers (The ISI series, Calcutta) by C.R. Rao Mitra and Mathai.

2.4: Merits and limitations of simple Random sampling:

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

2. Because of its random character, it is possible to ascertain the efficiency of the estimate by considering the standard errors of their sampling distributions. Moreover, large sample will be more representative of the population according to the principle of statistical regularity and the principle of Inertia of large numbers and thus provide better results.

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 labor.

Demerits:

1. Simple random sampling requires an up-to-date frame i.e, a compute 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 steps is quite time consuming and uneconomical particularly if the population is large. According this method can‟t be used effectively to collect most of the data in social sciences.

5. For given degree of accuracy, simple random sampling usually requires larger sample as compared to stratified random sampling.

6. Sometimes, simple random sample gives results which are highly improbabilistic in nature i. e., whose probability is very small.

2.5: Stratified Random Sampling: When the population is heterogeneous with respect to the variable or characteristic under study, then the technique of stratified random

sampling is used to obtain more efficient results. Stratification means division into layers or groups. Stratified random sampling involves following steps:

1. Stratify the given population into a number of sub-groups or sub-populations known as"strata" such that:

(a) The units within each stratum (sub-group) are as homogeneous as possible.

(b) The differences between various strata are as marked as possible, i. e. the stratum means differ as widely as possible

(c) Various strata are non- over lapping. This means each and every unit in the population belongs to one and only one stratum.

The criterion used for the stratification of the universe into various strata is known as stratifying factor. In general, geographical, sociological or economic characteristics from the basis of stratification of the given population. Some of the commonly used stratifying factors, are age, sex income, occupation, education level, geographic area, economic status, etc, stratification will be effective only if it possesses the three characteristics (a), (b), (c) enumerated above. In many fields of highly skewed distributions, stratification is a very effective and valuable tool.

Thus is stratified sampling the given population of size N is divided into, (say) „k" relatively homogeneous strata of sizes N_1, N_{21}……, N_k respectively such that N = $\sum_{i=1}^{k} n_i$

2. Draw simple random samples (without replacement) from each of the „k" strata. Let srswor of size ni be drawn from the i^{th} strata, (i=1,2,……, k) such that $\sum_{i=1}^{k} n_i$ = n_1 where n is the sample size from a population of size N.

The sample of n = $\sum_{i=1}^{k} n_i$ units is known as stratified random sample (without replacement) and the technique of drawing such a sample is known as stratified random sampling.

The basic problems in stratified random sampling are:

1. The stratification of the universe into different strata or sub-groups.

2. The determination of the sizes of the samples to be drawn from different strata. Both these points are equally important. A faculty stratification cannot be compensated even by taking large samples.

2.6: Allocation of sample size in stratified sampling: To obtain efficient results, the allocation of sample size ni, (i= 1, 2,…., k) i.e., the number of units to be selected from the i^{th} stratum, the total sample size n = n_1 + n_2 +….. + n_k being given, is done in the following ways:

i. Proportional Allocation

ii. Optimum Allocation

iii. Disproportionate Allocation

(i) Proportional Allocation: In this, the items are selected from each stratum in the same proportion as they exist in the population. The allocation of sample sizes is termed as proportional if the sample fraction, i.e., the ratio of sample size to the population size remain the same in all the strata, mathematically, the principle of proportional allocation gives:

$n_1/N_1 = n_2/N_2 = ….. n_k/N_k$

ii. Optimum Allocation: In this case the size of the samples to be drawn from the various strata is determined by the principle of optimization i.e., obtaining best results at minimum possible cost. In optimum allocation, n_i"s (i = 1, 2,…., k) are determined so that:
(i) Variance of sample estimate of the population mean is minimum (i.e., its precision is maximum) for fixed total sample size n. (Neyman"s Allocation).
(ii) Variance of the estimate is minimum for a fixed cost of the plan.
(iii) Total cost of the sampling design is minimum for fixed desired precision, i.e., total cost is minimum for a fixed value of the variance of the sample estimate.
iii. Disproportionate Allocation: In this case an equal number of items are taken from every stratum regardless of how the stratum is represented in the population. Sometimes, the proportion may vary from stratum to stratum also. In short, a stratified sample in which the number of items selected from each stratum is independent of its size is called disproportionate stratified sample.

2.7: Merits and Demerits of stratified Random Sampling:
Merits: 1. More Representative Sample: A properly constructed and executed stratified random sampling plan overcomes the drawbacks of purposive sampling and random sampling and still enjoys the virtues of both these methods by dividing the given universe into a number of homogeneous subgroups with respect to purposive characteristic and then using the technique of random sampling in drawing samples from each stratum. A stratified random sample gives adequate representation to each strata or important section of the population and eliminates the possibility of any important group of the population being completely ignored. The stratified random sampling provides a more representative sample of the population and accordingly results in less variability as compared with others sampling designs.
2. Greats Precision: As a consequence of the reduction in the variability within each stratum, stratified random sampling provides more efficient estimates as compared with simple random sampling. For instance, the sample estimate of the population mean is more efficient in both proportional and Neyman"s allocation of the samples to different strata in stratified random sampling as compared with corresponding estimate obtained in simple random sampling.
3. Administrative Convenience: The division of the population into relatively homogeneous subgroups brings administrative convenience. Unlike random samples, the stratified samples are expected to be localized geographically. This ultimately results in reduction in cost and saving in time in terms of collection of the data, interviewing the respondents and supervision of the field work.
4. Some times it is desired to achieve different degrees of accuracy for different segments of the population. Stratified random sampling is the only sampling plan which enables us to obtain the results of known precision for each of the stratum.
5. Quite often, the sampling problems differ quite significantly in different segments of the population. In such a situation, the problem can be tackled effectively through stratified sampling by regarding each segment of the population as a different strata and approaching upon them independently design sampling.
DEMERITS:
1. As already pointed out the success of stratified random sampling depends on:
    i. Effective stratification of the universe into homogeneous strata and
    ii. Appropriate size of the samples to be drawn from each of the stratum.

If stratification is faulty, the results will be biased. The error due to wrong stratification cannot be compensated even by taking large samples.
2. 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.

2.8: Systematic Sampling:

Systematic sampling is slight variation of the simple random sampling in which only the first sample unit is selected at random and the remaining units are automatically selected in a definite sequence at equal spacing from one another. This technique of drawing samples is usually recommended if the complete and up-to-date list of the sampling units, i.e., the frame is available and the units are arranged in some systematic order such as alphabetical, chronological, geographical order, etc. This requires the sampling units in the population to be such a way that each item in the population uniquely ordered identified by its order, for example the names of persons in a telephone directory, the list of voters, etc.

Let us suppose that „N" sampling units in the population are arranged in some systematic order and serially numbered from 1 to N and we want to draw a sample of size n from it such that N = nk $\Rightarrow$ k = N/n where „k" is usually called sample interval, systematic sampling consists in selecting any unit at random from the first „k" units numbered from 1 to k and then selecting every k^{th} unit in succession subsequently. Thus, if the first unit selected at random is i^{th}, unit, then systematic sample of size n will consist of the units numbered.
i, i+k, i+2k, …., i+ (n-1)k.

The random number „i" is called the random start and its value infact, determines the whole sample. As an sample, let us suppose that we want to select 50 voters from a list of voters containing 1,000 names arranged systematically. Here n = 50 and N = 1000
$\Rightarrow$k = N/n = 1000/50 = 20
We select any number from 1 to 20 at random and the corresponding voter in the list is selected. Suppose the selected number is 6. Then, the systematic sample will consists of 50 voters in the list at serial numbers: 6, 26, 46, 66,…., 966, 986.

2.9: Merits and Demerits:
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 labor relative to simple random sampling or stratified random sampling.
2. Systematic sampling may be more efficient than simple random sampling provided the form 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.
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 of there are periodic features in the frame and the sampling interval (k) is equal to or a multiple of the period.

The relatively efficiency of the systematic sampling over stratified random sampling or simple random sampling without replacement (srswor) largely depends on the properties of the population under study. Without a knowledge of the structure of the population no hard and fast rules can be laid down and no situations can be pinpointed where the use of systematic sampling is to be recommended.

2.10 Summary

For achieving desired correct results from a sample survey, the execution of sample design is of utmost importance and hence proper selection of the sampling method becomes imperative. The probability sampling is the scientific technique which draws sample from the population has same predefined probability of inclusion of an event into the drawn sample. In the probability sampling, the samples are drawn based on the random procedure and not on any judgmental method.

In simple random sampling, we draw very homogeneous samples and the individual elements are drawn from the whole population. Stratified random sampling method is an effective sampling tool to create homogeneous class samples rather than the total. In systematic random sampling, the first sample unit is selected at random and the remaining units are automatically selected on a definite sequence at equal spacing from one another.

2.11 Self assessment questions

(1) Describe about the simple random sampling
(2) State the merits and limitations of simple random sampling.
(3) What are the methods of selection of a simple random sample? Explain.
(4) Explain about the stratified random sampling.
(5) Describe the allocation methods in stratified random sampling method.
(6) State the merits and demerits of stratified random sampling.
(7) Explain about the systematic sampling.
(8) State the merits and demerits of systematic sampling.

2.12 Reference Books:

1. S. C. Gupta: Fundamentals of Statistics
2. K. Chandra Sekhar: Business Statistics
3. K. V. Sarma: Statistics made simple, prentice Hall of India

Lesson Writer
Dr. J. Pratapa Reddy

3. NON- RANDOM SAMPLING METHODS

Objectives

After completion of this chapter, you should be able to:

- Know the various non-random sampling methods;
- Explain the differences between random and non-random sampling methods;
- Understand the applications of various non-random sampling methods.

Structure

3.1 Introduction

As against the Probability sampling, the Non- probability sampling is the procedure of selection of a sample without the use of randomization. It is based on convenience or judgment and hence is likely to be biased. The sampling variation in such a case is very uncertain and cannot be estimated.

3.2 Non-random sampling methods

The various Non-random sampling methods (including mixed) are

(i) Cluster sampling.
(ii) Quota sampling.
(iii) Purposive or subjective or judgment sampling
(iv) Heterogeneity Sampling
(v) Snowball Sampling

3.3 Difference between random and non-random sampling methods:

Probability (Random) sampling	Non-Probability (Non-Random) Sampling
Allows use of statistics, tests hypotheses	Exploratory research, generates hypotheses
Can estimate population parameters	Population parameters are not of interest
Eliminates bias	Adequacy of the sample can't be known
Must have random selection of units	Cheaper, easier, quicker to carry out

3.4: Cluster sampling:

In this case the total population is divided, depending on problem under study, into some recognizable sub-divisions which are termed as clusters and a simple random sample of these clusters and a simple random sample of these clusters is drawn. We then observe, measure and interview each and every unit in the selected clusters.

For example, if we are interested in obtaining the income or opinion data in a city, the whole city may be divided into "N' different blocks or localities which determine the clusters and simple random sample of "N" blocks is drawn. The individuals in the selected blocks determine the clusters sample.

In using cluster sampling the following points should be borne in mind:

(i) Cluster should be as small as possible consistent with the cost and limitations of the survey, and
(ii) The number of sampling units in each cluster should be approximately same.

Thus, cluster sampling is not to be recommended if we are sampling areas in the city where there are private residential houses, business and industrial complexes, apartment buildings, etc., with widely varying number of persons or households.

3.5 Quota sampling:

Quota sampling may be looked as a special form of stratified sampling. In this method, the investigator is told in advance the number of the sample units he is to examine or enumerate from the stratum assigned to him. In the language of stratified sampling, the quota of the units to examined by the investigator from the stratum assigned to him is fixed for each investigator. The sampling quotas may be fixed according to some specified characteristic such as income group, sex, occupation, political or religious affiliations, etc., The choice of the particular units or individuals for investigation is left to the Investigators themselves. They are merely given the quotas with the specific instruction to inspect or interview a specified number of units from each stratum. Quite often the investigator does not make random selection of the sample units. He usually applies his judgment and discretion in the choice of the sample and tries to get the desired information as quickly as possible. Moreover, in case of non-response from some of the selected sample units (due to certain reasons like non-availability of the respondent even after repeated calls by the investigator, or the inability or refusal of the informant to furnish the requisite information), the investigator selects some fresh units himself to complete his quota. In doing so he is likely

to include some purposive units to get the desired information.

3.6 Merits and Demerits of Quota Sampling:

Merits

(1). Quota sampling is a stratified-cum-purposive or judgment sampling and thus enjoys the benefits of both. It aims at making the best use of stratification without incurring high costs involved in following any probabilistic method of sampling. There is considerable saving in time and money as the sampled units may be so selected that they are close together.

(2). If carefully executed by skilled and experienced investigators who are aware of the limitations of judgment sampling and if proper controls or checks are imposed on the investigators, Quota sampling is likely to give quite reliable results.

Demerits:

Since quota sampling is a restricted type of judgment sampling, it suffers from all the limitations of judgment or purposive sampling,

(i) It may be baised because of the personal beliefs and prejudices of the investigator in the selection of the units and inspecting them.

(ii) It may involve the bias due to the substitution of the sampled units from where there is no response.

(iii) Since it is not based on random sampling, the sampling error cannot be estimated.

Inspite of all these short comings, the technique of quota sampling is generally adopted in market surveys, political surveys of opinion poll where it is very difficult, rather impossible, to identify the strata in advance.

3.7 Purposive or subjective or judgment sampling:

In this method, a desired number of sample units is selected deliberately or purposely depending upon the object of the enquiry so that only the important items representing the true characteristics of the population are included in the sample.

An obvious and serious draw back of this sampling scheme is that it is highly subjective in nature, since the selection of the sample depends entirely on the personal convenience, beliefs, biases and prejudices of the investigator. For example, if in socio-economic survey it is desired to study the standard of living of the people in New Delhi and if the investigator wants to show that the standard has gone down, then he may include individuals in the samples only from the low income stratum of the society and exclude the people from the posh colonies like south Extension, Greater kailash, Jor Bagh,Chanakyapuri and so on. This method cannot be worked out for large samples and is expected to give good results in small samples only provided the selection of the sample is representative. This can be achieved if the investigator is thoroughly skilled and experienced in the field of enquiry and knows the limitations of such a selection. Further, since this scheme does not involve the principle of probability, estimation of the sampling error depends upon the hypothesis which are rarely met in practice.

The selection of the sample based on the theory of probability is also known as random selection and sometimes the probability sampling is also called random sampling.

3.8 Heterogeneity Sampling and snowball sampling:

We sample for heterogeneity when we want to include all opinions or views, and we aren't concerned about representing these views proportionately. Another term for this is sampling for *diversity*. In many brainstorming or nominal group processes (including concept mapping), we would use some form of heterogeneity sampling because our primary interest is in getting broad spectrum of ideas, not identifying the "average" or "modal instance" ones. In effect, what we would like to be sampling is not people, but ideas. We imagine that there is a universe of all possible ideas relevant to some topic and that we want to sample this population, not the population of people who have the ideas. Clearly, in order to get all of the ideas, and especially the "outlier" or unusual ones, we have to include a broad and diverse range of participants. Heterogeneity sampling is, in this sense, almost the opposite of modal instance sampling.

In snowball sampling, you begin by identifying someone who meets the criteria for inclusion in your study. You then ask them to recommend others who they may know who also meet the criteria. Although this method would hardly lead to representative samples, there are times when it may be the best method available. Snowball sampling is especially useful when you are trying to reach populations that are inaccessible or hard to find. For instance, if you are studying the homeless, you are not likely to be able to find good lists of homeless people within a specific geographical area. However, if you go to that area and identify one or two, you may find that they know very well who the other homeless people in their vicinity are and how can found them.

3.9 Summary

The difference between non-probabilistic and probabilistic sampling is that non-probabilistic sampling does not involve *random* selection and probability sampling does. Non-probabilistic samples don't depend upon the rationale of probability theory. At least with a probabilistic sample, we know the odds or probability that we have represented the population well. We are able to estimate confidence intervals for the statistic. With non-probabilistic samples, we may or may not represent the population well and it will often be hard for us to know how well we have done so. In general researchers prefer probabilistic or random sampling methods to non probabilistic ones, and consider them to be more accurate and rigorous. However, in applied social research there may be circumstances where it is not feasible, practical or theoretically sensible to do random sampling. Here, we consider a wide range of non-probabilistic alternatives.

3.10 Self assessment questions

(1) Distinguish between random and non-random sampling methods.
(2) Describe about the cluster sampling.
(3) Explain the Quota sampling with merits and demerits.
(4) Describe about the purposive sampling.
(5) Write brief note on Heterogeneity Sampling and snowball and snowball sampling.

3.11 Reference books:

1. S.C. Gupta: Fundamentals of statistics.
2. K. Chandra Sekhar: Business Statistics
3. K. V. Sarma: Statistics made simple, Prentice Hall of India.

Lesson Writer
Dr.J. Pratapa Reddy.

4 Correlation

Objectives:

- After completion of this chapter you should be able to
- Know the meaning of correlation;
- Explain the types of correlation;
- Understand the methods to measure correlation

Structure:

4.1 Introduction:

In univariate distributions only i. e, the distributions involving only one variable and also saw how the various measures of central tendency, dispersion, skewness and Kurtosis can be used for the purposes of comparison and analysis. We may, however, come across certain series where each item of the series may assume the values of two or more variables. If we measure the heights and weights of n individuals, we obtain a series in which each unit (individual) of the series assumes two values –one relating to heights and the other relating to weights. Such distribution, in which each unit of the series assumes two values is called a bivarate distribution. In series, the units on which different measurements are taken may be of almost any nature such as different individuals, times, places, etc. for example we may have:

(1) The series of marks of individuals in two subjects in an examination.

(ii) The series of sales revenue and advertising expenditure of different companies in a particular year.

(iii) The series of ages of husbands and wives in a sample of selected married couples and so on.

Thus in a bivariate distribution we are given a set of pairs of observations, one value of each paid being the values of each of the two variables.

In a bivariate distribution, we may be interested to find if there is any relationship between the two variables under study. The correlation is statistical tool which studies the relationship between two variables and correlation analysis involves various methods and techniques used for studying and measuring the extent of the relationship between the two variables. According to the correlation; cowden

"When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation.

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

4.2 Types of Correlation: The various types of correlation are.

(a) **Positive and Negative Correlation**:

If the values of the two variables deviate in the same direction i. e, if the increase in the values of variable results, on an average, in corresponding increase in the values of other variable or if a decrease in the values of the variable results, on an average, in corresponding decrease in the values of the other variable, correlation is said to be positive or direct.

Examples of series of positive correlation are:

(i) Heights and weights.

(ii) The family income and expenditure on luxary items.

(iii) Price and supply of a commodity and soon on the other hand, correlation is said to be negative or inverse if the variables deviate in the opposite direction i. e if the increase (decrease) in the values of one variable results, on the averages, in a corresponding decrease (increase) in the values of the other variables.

Examples of negative correlation are series relating to:

(i) Price and demand of a commodity.

(ii) Volume and pressure of perfect gas.

(iii) Sales of woolen garments and the day temperature and so on.

(b) Linear and Non-Linear correlation: The correlation between two variables is said to be linear if corresponding to a unit charge in one variable, there is a constant change in other variable over the entire range of the values. For example, let us consider the following data:

x	1	2	3	4	5
y	5	7	9	11	13

Thus for a unit change in the value of x, there is a constant change viz, 2 in the corresponding values of y. mathematically, above data can be expressed by the relation. $Y = 2x + 3$.
In general, two variables x and y are said to be linearly related, if there exists a relationship of the for n

$y = a + bx \quad -(1)$

(1) is the equation straight line with slope „b" and which makes an intercept „a" on the y – axis.

If the values of the two variables are plotted as the points in the xy-plane we shall get a straight line. Thus can be easily checked for the example given above. Such phenomena occur frequently in physical sciences but in economics and social sciences, we very rarely come across the data which give a straight line graph. The relation between the two variables is said to be non-linear or curvilinear if corresponding to a unit change in one variable, the other variable does not change at constant rate but fluctuates rate. In such cases if the data are plotted on the xy-plane, we do not get a straight line curve .Mathematically speaking, the correlation is said to be non- linear if the slope of the plotted curve is not constant. Such phenomena are common in the data relating to economics and social sciences.
Since the techniques for the analysis and measurement of non-linear relation are quite complicated and tedious as compared to the methods of studying and measuring linear relationship we generally assume that the relationship between the two variables under study is linear.

4.3 Scatter diagram method:

Scatter diagram is one of the simplest ways of diagrammatic representation of bivariate distribution and provides us one of simplest tools of as certaining the correlation between two variables. Suppose we have given n pairs of values (x_1, y_1), (x2, y2)…… (x_n, y_n) of two variables x and y. For example, if the variables x and y belong the height and weight respectively, then the pairs (x_1, y_1), (x_2, y_2)…… (x_n, y_n) may represent heights and weights (in pairs) of n individuals. Those n points may be plotted as dots (.) on the x- axis and y- axis in xy plane. The diagram of dots so obtained is known as "Scatter diagram". From scatter diagram we can form fairly good, through rough, idea about the relationship between the two variables.

The following points may be born in mind in interpreting the scatter diagram regarding the correlation between the two variables:

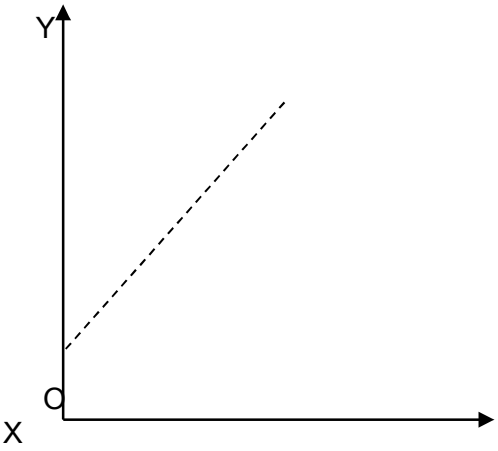
1. If the points are very dense i.e very close to each other, a fairly good amount of correlation may be expected between the two variables. On the other hand, if the points are widely scattered, a poor correlation may be expected between them.

(ii) If the points on the scatter diagram reveal any trend (either upward or down ward), the variables are said to be correlated and if no trend is revealed, the variables are uncorrelated.

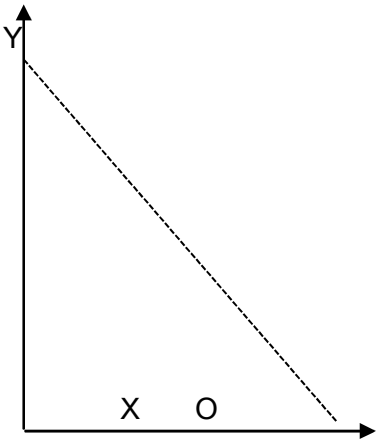
(iii) If there is an upward trend rising from lower left hand corner and going upward to the upper right hand corner, the correlation is positive since this reveals that the values of the two variables move in the same direction. If, on the other hand, the points depict a downward trend from the upper left hand corner to the lower left hand corner, the correlation is negative since in this case the values of the two variables move in the opposite direction.

(iv) In particular, if all points lie on a straight line star tip from left bottom and gong up towards the right top, the correlation is perfect and positive, and if all the points lie on straight line starting from left top and coming down to right bottom, the correlation is perfect & negative the following diagrams of the scattered data depict different forms of correlation.

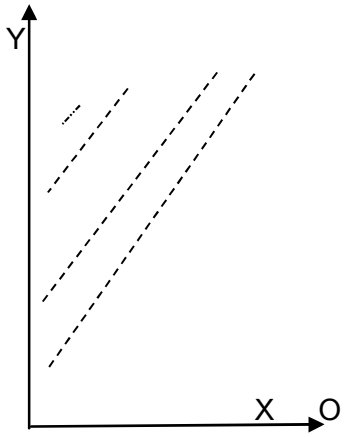
Perfect Positive Correlation



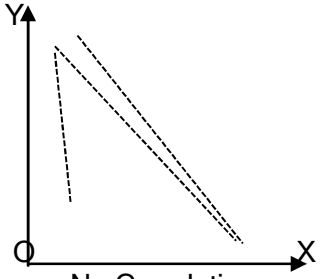
Perfect Negative Correlation



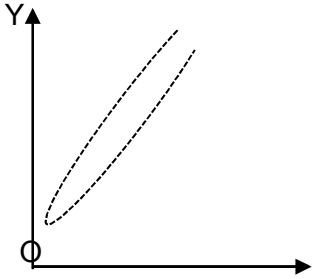
Low degree of Positive Correlation



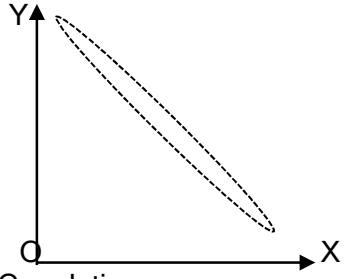
Low Degree of Negative Correlation



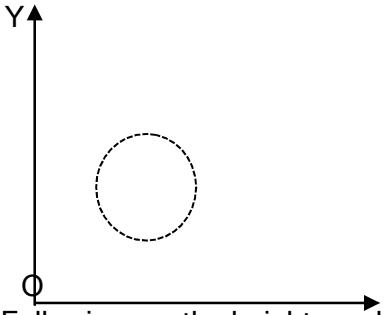
High degree of positive correlation



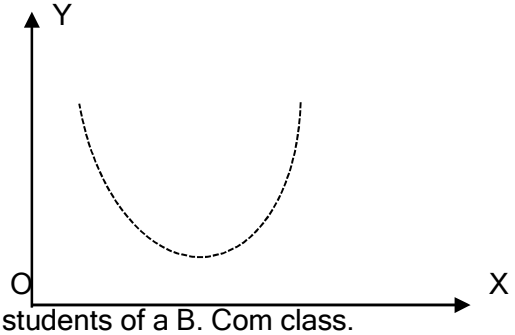
High degree of negative correlation



No Correlation



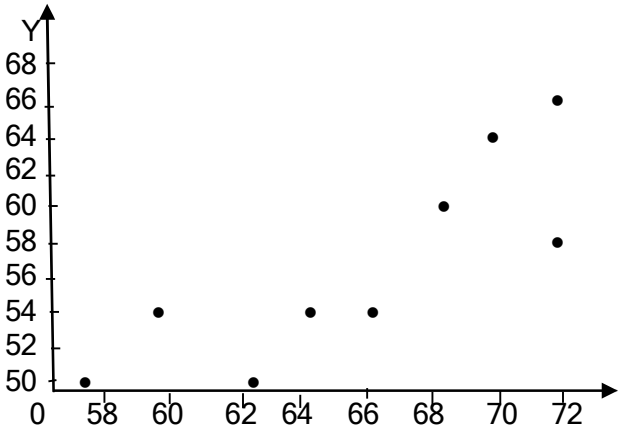
No Correlation



→ Following are the heights and weights of 10 students of a B. Com class.

Height (in inches) X:	62	72	68	58	65	70	66	63	60	72
Weight (in kgs) Y:	50	65	63	50	54	60	61	55	54	65

Draw a scatter diagram and indicate whether the correlation is positive of negative.
Solution: The scatter diagram of the above data is shown below



Since the points are dense i. e close to each other we may expect a high degree of correlation between the series of heights and weights.

1. 4 Karl Pearson's Coefficient of Correlation: (Co-Variance method)

Karl Pearson"s measure, known as Pearson"s an Correlation coefficient between two variables (series) X and Y. usually denoted by r (x, y) of r_{xy} of simply r, is a numerical measure of linear relationship between them and is defined as the ratio of the covariance between x and y, written as cov (x, y) to the product of the standard deviations of x and y symbolically,

$$R = \frac{\text{cov}(x, y)}{\sigma x \sigma y} \quad \text{.....4. 1}$$

If (x_1, y_1), (x_2, y_2) ..... (x_n, y_n) are n pairs of observations of the variables x and y in a bivarite distribution, then

Cov (x, y) = $\frac{1}{n}\sum (x - \bar{x})(y - \bar{y})$; $\sigma x = \sqrt{\frac{1}{n}(\varepsilon(x - \overline{x})^2}$

$\sigma y = \sqrt{\frac{1}{n}(\varepsilon(y - \overline{y})^2}$ 4.2

Summation being taken over n pairs of observations. Substituting in (4.1), we get.

$$r = \frac{\frac{1}{n}\sum (x - \bar{x})(y - \overline{y})}{\sqrt{\frac{1}{n}\sum (x - \bar{x})^2 . \frac{1}{n}\sum (y - \overline{y})^2}} = \frac{\sum (x - \bar{x})(y - \overline{y})}{\sqrt{\sum (x - \bar{x})^2 . \sum (y - \bar{y})^2}} \text{ - 4.3}$$

the formula (4.3) can also be written as.

$$r = \frac{\sum dxdy}{\sqrt{\sum dx^2 . \sum dy^2}} \quad \text{.....4.3 (a)}$$

Where dx and dy denote the deviations of x and y values from their arithmetic means $\bar{x}$ and $\bar{y}$ respectively i.e.,

dx = x - $\bar{x}$, dy = y - $\bar{y}$ - 4.3 (b)

simplifying (4.2) we get

cov (x, y) = $\frac{1}{n}\sum (x - \bar{x})(y - \bar{y}) = \frac{1}{n}\sum (xy - \bar{x}y - x\bar{y} + \bar{x}\bar{y})$

$= \frac{1}{n}\sum xy - \bar{x}.\frac{1}{n}\sum y - \bar{y}.\frac{1}{n}\sum x + \frac{1}{n}.n\bar{x}\bar{y}$

Since $\bar{x}$ and $\bar{y}$ are constants and since $\sum cx = c\sum x$ and $\sum c = mc$, where c is a constant.

$\Rightarrow$ Cov (x, y) = $\frac{1}{n}\sum xy - \bar{x}\bar{y} - \bar{y}\bar{x} + \bar{x}\bar{y}$

$\Rightarrow$ cov (x, y) = $\frac{1}{n}\sum xy - \bar{x}\bar{y}$ (4.4)

$\Rightarrow$ cov (x, y) = $\frac{1}{n}\sum xy - \left(\frac{\sum x}{n}\right)\left(\frac{\sum y}{n}\right)$

$\Rightarrow$ cov (x, y) = $\frac{1}{n^2}\left[n\sum xy - (\sum x)(\sum y)\right]$ 4. 4 (a)

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

$\frac{1}{n}\sum x^2 - \left(\frac{\sum x}{n}\right)^2 = \frac{1}{n}\left(n\sum x^2 - (\sum x)_2\right)$

Similarly we have $\sigma y^2 = \frac{1}{n^2}\left[n\sum y^2 - (\sum y)^2\right]$

Substituting (4.4a), (4. 4b) and (4.4c) and (4.1) we get.

$$r = \frac{\frac{1}{n^2}\left[n\sum xy - (\sum x)(\sum y)\right]}{\sqrt{\frac{1}{n^2}\left\{n\sum x^2 - (\sum x)^2\right\}}\sqrt{\frac{1}{n^2}\left\{n\sum y^2 - (\sum y)^2\right\}}}$$

$$= \frac{n\sum xy - (\sum x).(\sum y)}{\sqrt{\left[(n\sum x^2 - (\sum x)^2\sqrt{n\sum y^2 - (\sum y)^2}\right]}} = -4.5$$

4. 5 Properties of Karl Pearson's Coefficient of Correlation. Property I:

Pearsonian Correlation Coefficient can not exceed 1 numerically. In other words it lies between

-1 and +1. Symbolically,

$-1 \le r \le 1$

If r = +1implies perfect positive correlation between the variables and r = -1implies perfect

negative correlation between the variables.

Property II : Correlation Coefficient is independent of the change of origin and scale.

Mathematically if x and y are the given variables and they are transformed to the new variables

u and v by the change of origin & scale viz,

$u = \frac{x - A}{h}$ and $\frac{y - B}{k}$; h > 0, k > 0.

Where A, B, h and k are constants, h > 0, k > 0, then the correlation coefficient between x and y is same as the correlation coefficient between u and v i. e,

r = (x, y) = r (u, v)

$\Rightarrow r_{xy} = r_{uv.}$

Property III: Two independent variables are un correlated but the converse is not true.

Peoperty IV: r (ax + b, cy + d) = (axc)/|axc| . r (x, y)

Where $|axc|$ is the modulus value of a x c.

Property V: If the variables x and y are connected by linear equation ax + by + c = 0, then the correlation coefficient between x and y is (+1) if the signals of a and b are different and (-1) if the signs of a and b are alike.

If ax + by + c = 0, then r = r (x, y) = +1 if a and b are of opposite signs -1 if a and b are of same sign.

4.6 Solved Problems:

(1) Calculate the coefficient of correlation for the ages of husbands and wives:

Age of Husband (years)	23	27	28	29	30	31	33	35	36	39
Age of wife (years)	18	22	23	24	25	26	28	29	30	32

Solution: calculations for correlation coefficient

X	Y	U = x -31	V = y - 25	U^2	V^2	u v
23	18	-8	-7	64	49	56
27	22	-4	-3	16	9	12
28	23	-3	-2	9	4	6
29	24	-2	-1	4	1	2
30	25	-1	0	1	0	0
31	26	0	1	0	1	0
33	28	2	3	4	9	6
35	29	4	4	16	16	16
36	30	5	5	25	25	25
39	32	8	7	64	49	56
∑x =311	∑y = 257	∑U = 1	∑v = 7	$\sum u^2 = 203$	$\sum v^2 = 163$	∑u v = 179

Karl Pearson"s correlation coefficient between U and V is given by

$$r_{uv} = \frac{n\left(\sum Uv\right) - (\sum u)(\sum y)}{\sqrt{\left[n\sum u^2 - \left(\sum u\right)^2\right]\left(n\sum v^2 - \left(\sum v^2\right)\right)}} = \frac{10\times 179 - 1\times 7}{\sqrt{\left(10\times 203 - (1)^2\right)\left[10\times 63 - (7)^2\right]}}$$

$$\frac{1790 - 7}{\sqrt{(2030 - (1))(1630 - 49)}} = \frac{1783}{\sqrt{2029\times 1581}} = \frac{1783}{45.04\times 39.76}$$

$$\frac{1783}{1790.79} = 0.9956$$

Since Karl Pearson"s Correlation coefficient (r) is independent of change of origin, we get

$r_{xy} = r_{uv} = 0.9956$

2. Find Karl Pearson"s coefficient of Correlation between sales and expenses of the following ten firms

Firm	1	2	3	4	5	6	7	8	9	10
Sales (000 units)	50	50	55	60	65	65	65	60	60	50
Expenses (000 rupees	11	13	14	16	16	15	15	14	13	13

Solution: Let sales of firm be denoted by x and expenses be denoted by y. It may be noted that we can take out factor 5 common in x series. Hence, it will be convenient to change the scale also in x. taking 65 and 13 as working means for x and y respectively,

Let u = (x -65) 5; v = y -13

Calculations for correlation coefficient

Firms	x	y	U = x - 65/5	V = y -13	U^2	V^2	u v
1	50	11	-3	-2	9	4	6
2	50	13	-3	0	9	0	0
3	55	14	-2	1	4	1	-2
4	60	16	-1	3	1	9	-3
5	65	16	0	3	0	9	0
6	65	15	0	2	0	4	0
7	65	15	0	2	0	4	0
8	60	14	-1	1	1	1	-1
9	60	13	-1	0	1	0	0
10	50	13	-3	0	9	0	0
	Σx= 580	Σy = 140	Σu = -14	Σv = 10	Σu^2 = 34	Σv^2 = 32	$\Sigma u\, v$ = 0

Karl Pearson correlation coefficient between u and v is given by,

$$r_{uv} = \frac{n\sum uv - \left(\sum u\right)\left(\sum v\right)}{\sqrt{\left(n\sum u^2 - \left(\sum u^2\right)\right)\left(n\sum v^2 - \left(\sum v\right)^2\right)}} = \frac{10\times 0 - (-14)\times 10}{\sqrt{\left(10\times 34 - (14)^2\right)\left(10\times 32 - (10)^2\right.}}$$

$$= \frac{140}{\sqrt{144\times 220}} = \frac{140}{\sqrt{31680}} = \frac{140}{177.99} = 0.78666$$

Since correlation coefficient is independent of change of origin and scales we finally have

$r_{xy} = r_{uv} = 0.7866$

3. Find the coefficient of correlation between the heights of brothers and sisters from the following data.

Heights of brothers (in (m) (x)	65	66	67	68	69	70	71
Height of sisters (in (m) (y)	67	68	66	69	72	72	69

Solution: let the heights of brothers be denoted by x and sisters by y, then

$\bar{x} = \frac{65+66+67+68+69+70+71}{7} = 68$, and

$\bar{y} = \frac{67+68+66+69+72+72+69}{7} = 69.$

Let us prepare the following table:

x	dx (x -68)	$Dx^2(x-68)^2$	y	dy = y -69	$dy^2 = (y-69)^2$	dxdy
65	-3	9	67	-2	4	6
66	-2	4	68	-1	1	2
67	-1	1	66	-3	9	3
68	0	0	69	0	0	0
69	1	1	72	3	9	3
70	2	4	72	3	9	6
71	3	9	69	0	0	0
	Σ dx = 0	$\Sigma dx^2 = 28$		$\Sigma dy = 0$	$\Sigma dy^2 = 32$	Σ dxdy = 20

Now, $r = \frac{\sum dxdy}{\sqrt{\sum dx^2}\ \sqrt{\sum dy^2}} = \frac{20}{\sqrt{28\times 32}} = \frac{5}{7.5} = 0.67$ (approx)

4. Calculate the coefficient of correlation for the following pairs of values of x and y.

X:	17	19	21	26	20	28	26	27
Y:	23	27	25	26	27	25	30	33

Solution: let the assumed means for x and y be 23 and 27 respectively, dx = x -23, dy = (y-27), we have table,

X	Y	dx = (x - 23)	dy = (y -27)	dx^2	dy2	dxdy
17	23	-6	-4	36	16	24
19	27	-4	0	16	0	0
21	25	-2	-2	4	4	4
26	26	3	-1	9	1	3
20	27	-3	0	9	0	0
28	25	5	-2	25	4	-10
26	30	3	3	9	9	9
27	33	4	6	16	36	24
n = 8	n = 8	$\Sigma dx = 0$	$\Sigma dy = 0$	$\Sigma dx^2 = 124$	$\Sigma dy^2 = 70$	$\Sigma dxdy = 48$

Now, $r = \frac{\sum dxdy - \frac{\sum dx \times \sum dy}{n}}{\sqrt{\sum dx^2}\ \frac{\sum dx^2}{n} \times \sqrt{\sum dy^2 - \frac{\sum dy^2}{n}}}$

$= \frac{48 - 0}{124 - 0\ \sqrt{7 - 0}}$

$= \frac{48}{\sqrt{124\times 70}}$

$= \frac{48}{93.166} = 0.515$

Hence r = 0.515

4.7 Summary:

Business managers need the information about the relation among the various parameters like, demand, expenditure, cash flows, prices etc, to take the decisions about their future operations. In day to day life also, we would like to find out whether any price increase we would change our behavior towards buying certain household items. Correlation gives us an idea to find the relationship between the two variables. Graphical method is an easy method, which provide to know the relation between the two variable, graphically. Carl Pearson‟s coefficient of correlation is a mathematical measure to find the linear relationship between the two variables. It is independent of units and also useful to find coefficient of determination.

4.8 Self assessment Questions:

(1) Calculate product moment coefficient of correlation for the following data of sales (x) and expenses (y) in lakhs of rupees of 10 firms.

X	46	33	41	38	36	45	34	37	50	40
Y	12	13	24	16	15	14	21	17	19	19

Ans: r_{xy} = -0.0213.

(2) Compute Karl Pearson‟s coefficient of correlation in the following data relating to over head expenses and cost of production.

Overheads (in 000Rs)	80	90	100	110	120	130	140	150	160
Cost (in 000 Rs)	15	15	16	19	17	18	16	18	19

Ans: r = 0.6928.

(3) In order to find the correlation coefficient between two variables X and Y from 12 pairs of observations, the following calculations were made.

$\Sigma x = 30$, $\Sigma y = 5$, $\Sigma x^2 = 670$, $\Sigma y^2 = 285$ $\Sigma x y = 33.4$

On subsequent verification it was found that the pair (x = 11, y = 4) was copied wrongly, the correct value being (x = 10, y = 14) find correct value of r.

Ans: 0.78

(4) Calculate the coefficient of correlation between X and Y series from the following data.

	X series	Y series
Number of pairs of observation	15	15
Arithmetic mean	25	8
Standard deviation	3.01	3.03
Sum of squares of deviations form mean	136	138

Sum of product of the deviations of X and Y series from their respective means = 122

Ans: r = 0.89

(5) From the following information relating to stock exchange quotations for two shares A and B, ascertain by using Pearson‟s coefficient of correlation how shares A and B are correlated, in their prices?

Price share (A) Rs	160	164	172	182	166	170	178
Price Share (B) Rs.	292	280	260	234	266	254	230

(6) Find the correlation between sales and advertising expenditure from the following data:

Sales (Rs. Lakh)	65	66	67	67	68	69	70	22
Adv. Expenditure	67	68	65	68	72	72	69	71

1. Explain various types of correlation
2. Define scatter diagram
3. State the properties of karl Pearson"s coefficient of correlation.

4.9 Reference Books

1. S. C Gupta: Fundamentals of Statistics
2. K. Chandra Sekhar: Business Statistics
3. K. V. Sarma: Statistics made simple, prentice Hall of India.

Lesson Writer
Prof. M. Koteswara Rao

5. Rank Correlation

Objectives:

- After Completion of this chapter, you should be able to
- Understand the concept of Rank Correlation.
- Explain how to allocate ranks to the data.

Structure:

5.1 Introduction:-

Rank Correlation Method:-

Sometimes the Statistical series in which the variables under consideration are not capable of quantitative measurement but can be arranged in Serial order. This happens when we are dealing with measured quantitatively but can be arranged serially, In Such Situations Karl Person's Coefficient of correlation cannot be used as such. Charles Edward Spearman, a British Psychologist, developed a formula in 1904. Which consists in obtaining the correlation coefficient between the ranks of n individuals in the two attributes under Study.

Suppose we want to find if two characteristics. A, Say intelligence and B, Say, beauty are related of not. Both the characteristics are incapable all of quantitative measurements but we can arrange a group of n individuals in order of merit (ranks) w. r. t proficiency in the two characterstics.

Let the random variables X and Y denote the ranks of the individuals in the characterstics A and B respectively. If we assume that there is no tie i.e, if no two individuals get the some rank in a characterstics the, obviously, X and Y assume numerical values ranging from 1 ton.

The Pearson Ian correlation coefficient between the ranks X and Y is called the rank correlation coefficient between the characterstics A and B for that group of individuals.

5.2 Spearman's Rank Correlation Coefficient:

The coefficient of rank correlation is given by formula

$$\int = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Where d^2 is the Square of the difference of corresponding ranks, and n is the number of pairs of observations.

Type I: When ranks are given:

Step I: Calculate the difference of ranks of X from the ranks of Y and write it under the column headed by D.

Step II: Square the difference D and write it under the column headed by D^2.

Step III: Apply the formula.

$$\int = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Where n is the total number of pairs of observations.

Type II: When the ranks are not given:

In this case we are given only the data. We assign the ranks to both the series X and Y by giving the rank I to highest values in both the series and so on.

Step I: Assign ranks to each item of both the series, if they are not given.

Step II: Calculate the difference of ranks of X from the ranks of Y and write it under the column header by d.

Step III: Square the difference d and write it under the column d^r.

Step IV: Apply the formula

$$\int = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Where n is the total number of pairs of observations.

Type III: When equal ranks are given to more than two variables (or) attributes (or) TIE CASE.

If two are more individuals are placed together in any classification w.r.t an attribute i.e, if in case of variable data, there are more than one item with the same rank in either or both the series (i.e, Tie Rank). Then the Spearman's Rank correlation coefficient formula given above does not give correlation coefficient of Tie Rank. The problem is solved by assigning average rank to each of these individuals who are put in tie. For example suppose an item is repeated at rank 5, then the common rank assigned to 5 and 6 the is (5+6)/ 2 = 5.5. Which is average of 5& 6, the ranks, which these items would have been assigned if they were different. The next rank assigned will be 7. But if an item repeated thrice at rank 2, then the common rank is assigned to each value will be (2+3+4)/3 = 3, which is the arithmetic mean of 2,3,4. The next ranks to be assigned would be 5. In order to find the Rank correlation coefficient of repeated ranks of tie ranks, an Adjustment of correlation factor is added to the Spearman's Rank Correlation formula, i.e

$$R=1 - \frac{6\sum d^2}{n(n^2-1)}$$

Correlation factors "in the formula (A), add the factor $\frac{m(m^2-1)}{12}$ to Σd^2, where m is the number of times on item is repeated. This correlation factor is to be added for each repeated value in both the series.

The modified formula for Tie Rank Correlation Coefficient is given by.

$$P = \frac{1 - 6[\sum D^2 + \frac{1}{12}\sum m_1^3 - m_1 + \frac{1}{12}(m_2^3 - m_2) + ,...]}{n(n^2-1)}$$

Where m_1, m_2….are the numbers of times the value is repeated.

5.3 Merits and Limitations of Rank Correlation Merits:

1. It is simpler to understand and easy to calculate as compared to Karl Pearson's Method.
2. It is useful for qualitative data such as beauty, honesty, efficiency etc.
3. It is a useful method when the actual data is not given but only ranks are given.

Limitations:

1. It cannot be used for grouped frequency distribution
2. It is not as accurate as Karl Pearson's Coefficient of correlation.
3. It cannot be used in a continuous series.
4. When the number of items is more than 30 and if the ranks are not known, this method consumes more time and therefore cannot conveniently be used.

5.4 Solved Problems:

(1) For the following data calculate the coefficient of Rank Correlation.

X:	80	91	99	71	61	81	70	59
Y:	123	135	154	110	105	134	121	106

Solution:

X	Rank X	Y	Rank Y	Rank difference d	Squire rank difference (d^2)
80	4	123	4	0	0
91	2	135	2	0	0
99	1	154	1	0	0
71	5	110	5	-1	1
61	7	105	8	-1	0
81	3	134	3	0	0
70	6	121	5	0	1
59	8	106	7	1	1
N = 8					$\Sigma (D^2) = 4$

$$P = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 4}{8(64-1)} = 1 - \frac{3}{63} = \frac{60}{63} = \frac{20}{21} = 0.952$$

2. From the data given below calculate the coefficient of rank correlation between X and Y.

X:	78	89	97	69	59	79	68	57
Y:	125	137	156	112	107	136	123	108

Solution: Here n = 8

X	Y	Rank in X R1	Rank in Y R2	Rank Difference D = R1 – R2	d^2
78	125	4	4	0	0
89	137	2	2	0	0
97	156	1	1	0	0
69	112	5	6	-1	1
59	107	7	8	-1	1
79	136	3	3	0	0
68	123	6	5	1	1
57	108	8	7	1	1
				Sd = 0	$\Sigma d^2 = 4$

Coefficient of rank correlation $P = 1 - \frac{6\sum d^2}{n(n^2-1)}$

$$= 1 - \frac{6\times 4}{8(64-1)} = 0.952$$

3. Calculate rank correlation coefficient between two series X and Y given below

X:	70	65	71	62	58	69	78	64
Y:	91	76	65	83	90	64	55	48

Also comment on result

Solution: We prepare the following table for finding.

Rank correlation. Here n = 8

Rank X	Rank Y	D = x - y	d^2
3	1	2	4
5	4	1	1
2	5	-3	9
7	3	4	16
8	2	6	36
4	6	1	4
1	7	-2	36
6	8	-6	4
		2	
n = 8			$\Sigma d^2 = 110$

Rank correlation Coefficient P = 1 - $\frac{6\sum d^2}{n(n^2-1)}$

=1- $\frac{6\times 110}{8(64-1)}$ = -0.3095

$\rightarrow$The correlation between X and Y is negative

4. Ten competitors in a beauty contest are ranked by three judges in the following orders.

Ordered.

1st judge	1	6	5	10	3	2	4	9	7	8
2nd judge	3	5	8	4	7	10	2	1	6	9
3rd judge	6	4	9	8	1	2	3	10	5	7

Use correlation coefficient to determine which pair of judges has the nearest approach to common taste in beauty.

Solution: let R_1, R_2, R_3 respectively be the ranks of first, second and third judge. Let rij be the rank correlation coefficient b/w the rank given by i^{th} & j^{th} judges I ± j i = 1,2,3, j = 1,2,3 let $d_{ij} = R_i - R_j$ be the difference of ranks of an individual given by i^{th} & j^{th} judge.

R_1	R_2	R_3	$d_{12}= R_1-R_2$	$d_{13} = R_1-R_3$	$d_{23}=R_2-R_3$	d^2_{12}	d^2_{13}	d^2_{23}
1	3	6	-2	-5	-3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	-3	-4	-1	9	16	1
10	4	8	6	2	-4	36	4	16
3	7	1	-4	2	6	16	4	36
2	10	2	18	0	8	64	0	64
4	2	3	-2	1	-1	4	1	1
9	1	10	8	-1	9	64	1	81
7	6	5	1	2	7	1	4	1
8	9	7	-1	1	2	1	1	4
			$\Sigma D_{12} = 0$	$\Sigma D_{13} = 0$	$\Sigma D_{23} = 0$	200	60	214

Here n = 10

∫ 12 = 1- $\frac{6\sum d^2}{n(n^2-1)}$ = $\frac{1-6\times 200}{10\times 99}$ = - $\frac{7}{33}$ = -0.2121

$\int 13 = 1 - \frac{6\sum d^2}{n(n^2-1)} = \frac{1-6\times 60}{10\times 99} = -\frac{7}{14} = 0.6363$

$\int 23 = 1 - \frac{6\sum d^2}{n(n^2-1)} = \frac{1-6\times 214}{10\times 99} = -\frac{49}{165} = -0.2970$

Since R_{13} is maximum. So the pair of first and third judge has the nearest approach to the common taste of beauty.

5. From the following table, calculate the rank correlation coefficient.

X:	48	33	40	9	16	16	65	24	16	57
Y:	13	13	24	6	15	4	20	9	6	19

Table for calculations

X	Y	Rank X R_1	Rank Y R_2	$D^1 = R_1 - R_2$	D^2
48	13	8	5.5	2.5	6.25
33	13	6	5.5	0.5	0.25
40	24	7	10	-3	9
9	6	1	2.5	-1.5	2.25
16	15	3	7	-4	16
16	4	3	1	2	4
65	20	10	9	1	1
24	9	5	4	1	1
16	6	3	2.5	0.5	0.25
57	19	9	8	1	1
					$\Sigma D^2 = 41$

In the X – series, we notice that the value 16 is repeated thrice. The common rank is given to these values is 3. Which is the average of 2, 3, 4. The ranks which these values would have assumed if they were different. The next rank given to the next value is 5. Here m1 = 3

The correlation factor for X - series is

$$\frac{m_1(m_1^2-1)}{12} = \frac{3(3^2-1)}{12} = 2$$

In the Y series the value 6 & 13 repeated twice. The value 6 occurs twice and its average rank is 2.5 {(2+3)/2)} the next value 9 is assigned the rank 4. Here $M_2 = 2$. Again the value 13 is repeated twice at different places. The rank is 5.5. The next value15 is assigned rank 7. Here $M_3 = 2$.

The correlation in Y - series is

$$\frac{m_2(m_2^2-1)}{12} + \frac{m_3(m_3^2-1)}{12} = \frac{2(2^2-1)}{12} + \frac{2(2^2-1)}{12} = 0.5 + 0.5 = 1$$

$$\int = \frac{1-6\left[\sum d^2 + \frac{m(m^2-1)}{12} + \frac{m_2(m_2^2-1)}{12} + \frac{m_3(m_3^2-1)}{12}\right]}{n(n^2-1)}$$

$$\int = \frac{1 - 6(41 + 2 + 0.5 + 0.5)}{10(10^2 - 1)} = 1.0.267 = 0.733$$

5.5 Summary:-

Rank correlation coefficient is also one of the methods of correlation. It is applied in the problems in which data cannot be measured quantitatively but quantative assessment is possible such as honest, beauty, smoking etc. We can computed the spearman's rank correlation coefficient, when the ranks are given, the ranks are not given and tie cases. Spearman's rank correlation is based on the ranks of the values of each variable. The spearman's rank correlation coefficient many also assumes the values between - 1 and + 1.

5.6 Self Assessment Questions.

1. Calculate Spearman's rank correlation coefficient between advertisement cost and sales from the following data.

Adv. Cost (Rs.1000)	39	65	62	90	82	75	25	98	36	78
Sales (Rs.100,000	47	53	58	86	62	68	60	91	51	84

Answer: 0.82

2. Calculate the rank correlation coefficient for the following table of marks of students in two subjects.

First Subject:	80	64	54	49	48	35	32	29	20	18	15	10
Second subject	36	38	39	41	27	43	45	52	51	42	40	52

Answer: - 0. 685

3. Ten competitors in a beauty contest are ranked by three judges in following order.

First judge	1	5	4	8	9	6	10	7	3	2
Second judge	4	8	7	6	5	9	10	3	2	1
Third judge	6	4	8	1	5	10	9	2	3	4

Use the rank correlation to given which pair of judges have the nearest approach to common taste in beauty.

Answer: Second and third judges are in nearest approach to the sense of beauty.

4. The following are the marks obtained by group of students in two papers. Calculate the rank coefficient of correlation.

Economics:	78	36	98	25	75	82	92	62	65
39									
Statistics:	84	51	91	69	68	62	86	58	35	49

Answer: $\int$ = 0.6121

5. The following data relate to the marks obtained by 10 students of class is statistics and costing.

Marks in Statistics:	30	38	28	27	28	23	30	33	28	35
Marks in costing:	29	27	22	29	20	29	18	21	27	22

Obtain the rank correlation coefficient.

Ans: $\int$ = - 0.3515.

6. Given following aptitude and I .Q Scores for a group of students. Find the coefficient of rank correlation.

Aptitude Score:	57	58	59	59	60	61	60	64
IQ Score:	97	108	95	106	120	126	113	110

Answer: ∫ = 0.7024

7. Quotations of index number of equity share process of a certain joint stock company and of prices of preference shares below:

Years:	1991	1992	1993	1994	1995	1996	1997
Equity shares:	97.5	99.4	98.6	96.2	95.1	98.4	97.1
Preference Shares:	75.1	75.9	77.1	78.2	79.0	74.8	76.2

Use methods of rank correlation to determine the relationship between equity and preferences share prices.

Answer: ∫ = - 0.6071

8. Define Spearman's rank correlation coefficient.
9. State merits and limitations of Rank correlation coefficient**.**

5.7 Reference Books

2. S. C. Gupta: Fundamentals of Statistics
3. K. Chandra Sekhar: Business Statistics.
4. K. V. Sarma: Statistics made simple, Prentice Hall of India.

Lesson Writer
Prof. M. Koteswara Rao

6. Regression

Objectives:

After completion of this chapter, you should be able to

- Understand the concept of Regression
- Explain the lines of regression
- Know the differences between correlation and Regression.

Structure:

6.1 Introduction:

The literal or dictionary meaning of the word „Regression" is „Stepping back or returning to the average value". The term was first used by British biometrician Sir Francis Galton. He studied the relationship between the heights of about one thousand fathers and sons and published the results in a paper. „Regression towards Mediocrity in Hereditary Stature". The interesting features of his study were:

(i) The tall fathers have tall sons and short fathers have short sons.

(ii) The average height of the sons of group of tall fathers is less than that of the fathers and the average height of the sons of a group of short fathers is more than that of the fathers.

He concluded that if the average height of a certain group of fathers is „o" cms. Above (below) the general average height than average height of their sons will be (a x r) cms. Above (below) the general average height where r is the correlation coefficients between the heights of given group of fathers and their sons.

The word regression as used in Statistics has much wider perspective without any reference to biometry. Regression analysis, in the general sense, means the estimation or prediction of the unknown value of one variable from the known value of other variable. It is one of the very important statistical tools which is extensively used in almost all sciences, natural, social and physical. It is specially used in business and economics to study the relationship between two or more variables that are related causally and for estimation of demand and supply curves, cost functions, production and consumption functions, etc.

Prediction or Estimation is one of the major problems in almost all spheres of human activity. The estimation or prediction of future production, consumption, prices, investments, sales, profiles, income etc. are of paramount importance to a businessman or economist. Population estimates and population projections are indispensable for efficient planning of an economy. The pharmaceutical concerns are interested in studying or estimating the effect of new drugs on patients. Regression analysis is one of the very scientific techniques for making such predictions. In the words of M. M. Blair "Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data".

In regression analysis there are two types of variables. The variable whose value is influenced or is to be predicted is called dependent variable and the variable which influences the values r is used for prediction, is called independent variable in regression analysis independent variable is also known as regress or predictor or explanatory while the dependent variable is also known or regressed or explained variable.

6.2 Lines of Regression

Line of Regression is the line which gives the best estimate of one variable for any given value of the other variable. Incase of two variables X and Y, we shall have two lines of regression; one of Y on X and the other of X on Y.

Definition:- Line of regression of Y on X is the line which gives the best estimate for the value of Y for any specified value of X.

Similarly, the line of regression of X and Y is the line which gives the best estimate for the value of X for any specified value of Y.

The term best fit is interpreted in accordance with the principle of least squares which consists in minimizing the sum of the squares of the residuals or the error of estimates i.e the deviations between given observed values of the variable and their corresponding estimated values as given by the line of best fit. We may minimize the sum of squares of the errors parallel to Y – axis of parallel to X – axis, the former, gives

the equation of the line of regression of Y on X and latter, Viz, minimize the sum of squares of the errors parallel to X – axis gives the equation of the line of regression of X and Y.

A line of regression is the line which gives the best estimate of one variable X for any given value of the other variable Y.

I. **Line of regression of X on Y**. It is the line which gives the best estimate for the value of X for a specified value of Y.

It is given by x $\bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$

Where $\bar{x}, \bar{y}$, are means of X series & Y series respectively $\sigma_x \sigma_y$ are S. D of X and Y series respectively and r is the correlation coefficient between X and Y.

It can also be put in the form

X = a + by

Where a is intercept of the line and b is the slope of line X on Y.

II. **Line of Regress of Y on X** . It is the line which gives the best estimate for the values of Y for any specified values of Y for any specified values of X.

Regression Equation for Y on X is given by

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

It can also be put in the form

Y = a + bx.

Where a is intercept of the line and b is slope of the line Y on X.

6.3 Regression Coefficient and Prosperities:-

Let us consider the line of regression of Y on X, viz,

Y = a +bx

The coefficient „b" which is the slope of the line of regression of Y on X is called the coefficients of regression of Y on X. It represents the increment in the value of the dependent variable Y for a unit change in the value of the independent variable X. In other words, it represents the rate of change of Y w. r. t. X. For notational conveniences, the slope „b" i.e coefficient of regression of Y on X is written as by X.

Similarly in the regression equation of X on Y, viz,

X = A + B y.

The coefficient B represents the change in the value dependent variable X for a unit change in the value of independent variable Y and is called the coefficient of regression of X on Y. For notational convenience, it is written as b x y.

b_{yx} = coefficient of regression of Y on X

b_{xy} = coefficient of regression of X on Y.

The coefficient of regression of Y on X is given by

$$b_{yx} = \frac{Cov(x, y)}{\sigma_x^2} = \frac{r\sigma_y}{\sigma x}$$

The coefficient of regression of X on Y is given by

$$b_{xy} = \frac{Cov(x, y)}{\sigma_y^2} = \frac{r\sigma_x}{\sigma_y}.$$

Properties of Regression Coefficients:-

1. The coefficient of correlation is the geometric mean of the coefficients of regression i.e. r = $\sqrt{b_{xy}\, b_{yx}}$.
2. If one of regression coefficients is greater than unity, then the other is less than unity.
3. Arithmetic mean of regression coefficient is greater than the correlation coefficient.
4. Regression coefficients are independent of change of origin but not of scale.
5. Both regression coefficient will have the same sign, i.e either both are positive of both are negative.
6. The sign of correlation is same as that of regression coefficients i.e r > 0 if b_{xy} >0; and r < o if regression coefficient are negative.

6.4 Differences between Regression and Correlation:

Correlation	Regression
1. It is a relationship between two ore more variables.	1. Regression means stepping back of returning to average value.
2. Correlation coefficient r between x and y is a measure of direction and degree of linear relationship between x and y.	2. b_{xy} and b_{yx} aer mathematical measures expressing the average relationship between the two variables.
3. It is symmetric in x and y i.e $r_{xy} = r_{yx}$	3. The regression coefficient b_{xy} and b_{yx} are not symmetri in x and y i.e $b_{xy} \neq b_{yx}$
4. The correlation coefficient does not reflect upon the nature of variable (independent of dependent variable)	4. Regression coefficient reflects on the nature of variable i.e, which is dependent variable and which is independent variable. In other words, it estimates the value of dependent variable for any given value of independent variable.
5. It does not imply cause and effect relationship between the variables under study.	5. It indicates the cause and effect relationship between variables. The variable corresponding to cause is taken as independent variables, whereas corresponding to effect is taken as dependent variable.
6. It is relative measure and is independent of the units of measurement.	6. Regression Coefficient are absolute measure of finding out the relationship between two or more variables.
7. It indicates the degree of association.	7. It is used to forecast the nature of dependent variable when the value of independent variable is known.
8. It has limited application as it is confined to the study of line as relationship between two variables.	8. It has wider applications as it

	also studies nonlinear relationship between the variables.

6.5 Uses of Regression Analysis:-

* The cause and effect relations are indicated from the study of regression analysis.
*It establishes the rate of change in one variable in terms of the changes in another variable.
*It is useful in economic analysis as regression equation can determine an increase in the cost of living index for a particular increase in general price level.
*It helps in prediction and thus it can estimate the values of unknown quantities.
*It helps in determining the coefficient of correlation as

$$r = \sqrt{b_{yx} \times b_{xy}}$$

*In enable us to study the nature of relationship between the variables.
*It can be useful to all natural, social and physical sciences, where the dta are in functional relationship.

6.6 Solved Problems:

1. Calculate the correlation and find the two lines of regression from the following data.

X	57	58	59	59	60	61	62	64
Y	67	68	65	68	72	72	69	71

Find the estimate of Y when X = 66.

Solution: We prepare the following table for calculations of byx and bxy by using deviations from actual means.

Table for calculations

x	X = x - $\bar{x}$	x^2	Y	Y = y - $\bar{y}$	y^2	x y
57	-3	9	67	-2	4	6
58	-2	4	68	-1	1	2
59	-1	1	65	-4	16	4
59	-1	1	68	-1	1	1
60	0	0	72	3	9	0
61	1	1	72	3	9	3
62	2	4	69	0	0	0
64	4	16	71	2	4	8
$\Sigma x = 480$	$\Sigma x = 0$	$\Sigma x^2 = 36m$	$\Sigma y = 552$	$\Sigma y = 0$	$\Sigma y^2 = 44$	$\Sigma xy = 24$

Here N = 8

Mean $\bar{x} = \frac{\Sigma x}{N} = \frac{480}{8} = 60.$

Mean $\bar{y} = \frac{\Sigma y}{N} = \frac{552}{8} = 69$

Also we know that $b_{xy} = \frac{\Sigma xy}{\Sigma x^2} = \frac{24}{36} = 0.67$

$$b_{yx} = \frac{\sum xy}{\sum y2} = \frac{24}{44} = 0.545$$

Regression equation of Y on X :

$y - \bar{y} = b_{yx} (x - \bar{x})$

$\Rightarrow$ 4-69 = 0.545 (x – 60) or y = 0.545x +69
-0.545x60

Y = 0.545 x + 27.3

When x = 66, y = 0.545 x 66 + 27.3
= 35.97 + 27.3 = 63.27

Hence Y = 63.27 l/m^x = 66

Regression equation of x on y : $x - \bar{x} = b_{xy} (y - \bar{y})$

$\Rightarrow$ x – 60= 0.67(y – 69)

$\Rightarrow$x = 0.67 y + 13.77

2. Find the two lines of regressions from following fata.

X	158	160	163	165	167	170	172	175	177	181
Y	163	158	167	170	160	180	170	175	172	175

Estimate y, when x = 164.

Solution:- let a = 170, b =175 be the assumed means for x and y series so that dx = x – 170, dy= y – 175

We have the following table for calculations.

x	Dx=x - 170	dx^2	Y	dy = Y - 175	dy^2	dx dy
158	-12	144	163	-12	144	144
160	-10	100	158	-17	289	170
163	-7	49	167	-8	64	56
165	-5	25	170	-5	25	25
167	-3	9	160	-15	225	45
170	0	0	180	5	25	0
172	2	4	170	-5	25	-10
175	5	25	175	0	0	0
177	7	49	172	-3	9	-21
181	11	121	175	0	0	0
Σx =1688	Σdx = -12	Σdx^2 = 526	Σy = 1690	Σdy = -60	Σdy^2 =806	$\Sigma dxdy$= 409

Here N = 10, $\bar{x} = \frac{\sum x}{N} = \frac{1688}{10} = 168.8$

$$\bar{y} = \frac{\sum y}{N} = \frac{1699}{10} = 169$$

$$b_{xy} = \frac{\sum dxdy.(\sum dx).(\sum dy)/N}{\sum dy^2 - (\sum dy)^2 / N} = \frac{409 - (-12) \times (-60)/10}{806 - (-60)^2 /10}$$

$$= \frac{409 - 72}{806 - 360} = \frac{337}{446} = 0.756$$

$b_{yx} = \frac{\sum dxdy - (\sum dx)(\sum dy)/N}{\sum dx^2 - (\sum dx)^2/N} = \frac{409 - 72}{526 - (-12)^2/10}$

$= \frac{337}{526 - 14.4} = \frac{337}{511.6} = 0.659$

Regression equation of x on y : $x - \bar{x} = b_{xy}(y - \bar{y})$

$X - 168.8 = 0.756\ (y - 169)$

Or $x = 0.7564 + 168 - (0.756)\ (169) = 0.7564 + 168 - 127.4$

$X = 0.7564 + 40.236$

Regression equation of y on x: $y - \bar{y} = b_{yx}(x - \bar{x})$

$Y - 169 = 0.659\ (x - 168.8)$

$Y = 0.659\ x + 169 - (0.659)\ (168.8)$

$Y = 0.659x + 57.761$

When x = 164, $y = 0.659 + 164 + 57.761$

$= 108.076 + 57.761$

$Y = 165.837$

3. The following table gives the age of cars of a certain make and actual maintences costs. Obtain the regression equtation. For costs related to age. Also estimate the maintenance cost for a ten years old car.

Age of car (years)	2	4	6	8
Maintenance cost (Rs) Hun	10	20	25	30

Solution let the x denote the age of car an y denote maintenance cost. Also we know that the maintenancecost dependence on the age of car, so we are to find the regression equation of y on x. We prepare the following table. Let 5 and 20 respectively be assumed means for x and y series.

Table for calculations

x	$X = x - \bar{x} = x - 5$	x^2	y	$Y = y - \bar{y} = y - 20$	y^2	x y
2	-3	9	10	-10	100	30
4	-1	1	20	0	0	0
6	1	1	25	5	25	5
8	3	9	30	10	100	30
$\Sigma x = 20$	$\Sigma x = 0$	$\Sigma x^2 = 20$	$\Sigma y = 85$	$\Sigma y = 5$	$\Sigma y^2 = 225$	$\Sigma_{xy} = 65$

Here $\bar{x} = \frac{20}{4} = 5.\ \bar{y} = \frac{85}{4} = 21.25$

Also $b_{yx} = \frac{\sum xy - (\sum x)(\sum y)/N}{\sum d^2 - \sum x^2/N} = \frac{65 - (0 \times 5)/4}{20 - 0/4} = 3.25$

The regression line y on x :

$Y - \bar{y} = b_{yx}(x - \bar{x})$

$Y - 21.25 = 3.25\ (x - 5)$

$Y = 3.25x - 5\ (3.25) + 21.25$

$Y = 5 + 3.25\ x$ ---(1)

The maintenance cost for a ten years old car:
Putting x – 10 in the equation (1) we get
Y = 37.5 (or) Rs. 37.501 –

4. From the data given below find:

(a) the two regression coefficients (b) the two regress in (c) the cofficient of correlation b/w marks in Economics and Statistics.
(d) the most likely marks in statistics when marks in Economics are 30.

Marks in Economics	25	28	35	35	31	36	29	38	34	32
Marks in Statistics	43	46	49	41	36	32	31	30	33	39

Solution:- Let us denote the marks in Economics by the variable x and the marks in Statistics by the variable y
Calculations for Regression Equations

X	Y	Dx = x - $\bar{x}$ = 2-32	Dy = y - $\bar{y}$ = y – 38	dx^2	dy^2	dx dy
25	43	-7	5	49	25	-35
28	46	-4	8	16	64	-32
35	49	3	11	9	121	33
32	41	0	3	0	9	0
31	36	-1	-2	1	4	2
36	32	4	-6	16	36	-24
29	31	-3	-7	9	49	21
38	30	6	-8	36	64	-48
34	33	2	-5	4	25	-10
32	39	0	1	0	1	0
Σx = 320	Σy = 380	Σdx = 0	Σdy =0	Σdx^2 =140	Σdy^2 =398	Σdxdy = -93

Here $\bar{x} = \frac{\Sigma x}{n} = \frac{320.}{10} = 32$, and $\bar{y} = \frac{\Sigma y}{n} = \frac{380}{10} = 38$

(a) Regression Coefficients

Coefficient of regression of y on x

$$= b_{yx} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{\Sigma dxdy}{\Sigma dx^2} = \frac{-93}{140} = -0.6643$$

Coefficient of regression x on y

$$b_{xy} = \frac{\Sigma(x-\bar{x})(\Sigma y-\bar{y})}{\Sigma(y-\bar{y})^2} = \frac{\Sigma dxdy}{\Sigma dy^2} = -\frac{93}{398} = -0.2337$$

(b) Regression Equations:

Equation of the line of regression of X on y is

$x - \bar{x} = b_{xy}(y - \bar{y})$

$\Rightarrow x - 32 = -0.2337(y - 38)$

$= -0.2337y + 8.8806$

$X = -0.2337y + 8.8806 + 32$

Equation of line of regression of y on x is

$y - \bar{y} = b_{yx}(x - \bar{x})$

$\Rightarrow y - 38 = 0.6443(x - 32)$

$y = -0.6643x + 21.2576 + 38$

X = -0.2337 y + 40.8806 y = 0.6643x + 59.2576

----*

(c) Correlation Coefficient - we have

$r^2 = b_{yx}. b_{xy} = (-0.6643) \times (-0.2337) = 0.1552$

$\Rightarrow r = \pm \sqrt{0.1552} = \pm 0.394$

Since both regression coefficient are negative, r must be negative, tense we get r = -0.394

(d) In order to estimate the most litely marks in Statistics (y) when marks in Economics (x) are 30,

We shall use line of regression of y on x , vig, the equatin(*). Taking x = 30 in (*), the required estimate is given by

Y = -0.6643 x30+39.2576

= -190929 + 59.2576 = 39.3286

Hence, the most likely marks in statistics when marks in Economics are 30, are 39.3286 $\cong$ 39.

5. From the following data, obtain the two regression equations.

Sales	:	91	97	108	121	67	124	51	73	111	57
Purchases	:	71	75	69	97	70	91	39	61	80	47

Solution:- Let us denote the sales by variable x and the purchases by variable y

Calculations for Regression Equations

X	Y	$dx = x - \bar{x}$	$dy = y - \bar{y}$	dx^2	dy^2	dxdy
91	71	1	1	1	1	1
97	75	7	5	49	25	35
108	69	18	-1	324	1	-18
121	97	31	27	961	729	837
67	70	-23	0	529	0	0
124	91	34	21	1156	441	714
51	39	-39	-31	1521	961	1209
73	61	-17	-9	289	81	153
111	80	21	10	441	100	210
57	47	-33	-23	1081	529	759
$\Sigma x = 900$	$\Sigma y = 700$	$\Sigma dx = 0$	$\Sigma dy = 0$	$\Sigma dx^2 = 6360$	$\Sigma dy^2 = 2868$	3900

We have $\bar{x} = \frac{\Sigma x}{N} = \frac{900}{10} = 90$, and $\bar{y} = \frac{\Sigma y}{N} = \frac{700}{10} = 70$

$b_{yx} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{\Sigma dxdy}{\Sigma dx^2} = \frac{3900}{6360} = 0.6132$

$b_{xy} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{y})^2} = \frac{\Sigma dxdy}{\Sigma dy^2} = \frac{3900}{2868} = 1.361$

Regression Equations.

Equation of line of regression yon x is equation of line of regression of x on y is

$Y - \bar{y} = b_{yx}(x - \bar{x})$ $x - \bar{x} = b_{xy}(y - \bar{y})$

$\Rightarrow y - 70 = 0.6132(x - 90)$ $\Rightarrow x - 90 = 1.361(y - 70)$

$\Rightarrow y = 0.6132x - 55.188 + 70.000$ $= 1361y - 95.27$

$\Rightarrow$y = 0.6132x + 14.812 $\Rightarrow$ x = 1.3614 – 95.27+90

$\Rightarrow$ x = 1.361 y- 5.27

We have

$r^2 = b_{yx}. b_{xy} = 0.6132 \times 1.361 = 0.8346$

$r = \pm\sqrt{0.8346} = \pm 0.9135$

But since, both the regression coefficients are positive, or must be positive.

Hence r = 0.9135

6. In a partially destroyed refold of the following data are available variance of x = 25

Regression equation of x on y : 5x –y =22

Regression equation y on X: 64x -45y = 24

Find (a) mean values of x and y (b) coefficient of correlation between x and y.

(c) Standard deviation of y

Solution:- (a) The Mean values of x and y lie on the regression lines and are obtained by solving the given regression equations.

$5\bar{x} - \bar{y} = 22$ – (1)

$64\bar{x} - 45\bar{y} = 24$.-(2)

Multiplying the equations we get

$225\bar{x} - 45\bar{y} = 990$ -(3)

Subtracting from (3), we get.

$161\bar{x} = 966 \Rightarrow \bar{x} = 6$

Putting $\bar{x}$ = b in (1) , we get

$30-\bar{y} = 22 \Rightarrow \bar{y} = 8$

Hence $\bar{x} = 6$, $\bar{y} = 8$

(b) The regression equation of y on x is

64 x – 45y = 24 or y = $\frac{64}{45}$ x - $\frac{24}{45}$

Y = $\frac{-8}{15}$ + $\frac{64}{45}$ x

b_{yx} = 64/45

Again regression equation of x on y is

5x –y = 22 or x = $\frac{22}{5}$ + 1/5y

b_{xy} = 1/5

but r = $\sqrt{b_{xy} \times b_{yx}}$ = $\sqrt{\frac{64}{45} \times \frac{1}{5}}$ = 8/15

Hence, the coefficient of correlation r = 8/15

(c) Now, it is given variance of x = $\sigma_x^2 = 25 \Rightarrow$

$\sigma_x = 5$

Also r = 8/15, b_{yx} 64/45

But we know that

$b_{yx} = \frac{r.\sigma_x}{\sigma_x}$

$64/45 = 8/15 \times \frac{\sigma_y}{5}$

$\Rightarrow \sigma_y = 40/3 = 13.33$

Hence the standard deviation of y = 13.33

6.7 Summary:

Prediction of estimation of future is an important area for the benefit of planning and looking deeper into business. Regression analysis is one of the scientific technique for making such a prediction. In the regression analysis the two types, variables are dependent variable and independent variable. Regression is a powerful tool is now used extensively in natural, social or physical sciences. The slope of the regression line is called regression coefficient.

6.8 Self Assessment Questions:-

(1) You are given the data relating to purchases and sales. Obtain the two regression equations by method of least squares and estimate the likely sales when the purchases equal 100.

Purchases:	62	72	98	76	81	56	76	92	88	49
Sales	112	124	131	117	132	96	120	136	97	85

Ans: purchase: x sales y: x = 0.6515 y -10.0775

Y = 0.78257 + 56.325 ; 134.5625

(2) You are given the following data:

	X	Y
Arithmetic mean	36	85
Standard deviation	11	8

Correlation coefficient between x and y = 0.66

(i) Find two regression equations. (ii) Estimate value of x when y = 75.

Ans: (i) y = 0.48x +67.72 ; x = 0.9075 y – 41.1375, (ii) 26.925.

(3) Calculate the two regression equations of x and y and y on x from the data given below taking deviations from actual means of x and y.

Price (Rs):	10	12	13	12	16	15
Amount demanded:	40	38	43	45	37	43

Estimate the likely demand when the price is Rs. 20.

Ans: x = -0.12y + 17.92, y = -0.25x + 44.25, when

X = 20, y = 49.25

(4) Calculate correlation and regression equations from following data.

X:	2	4	6	8	10	12	14
y:	4	2	5	10	4	11	12

Find the estimate of y when x = 13.

Answer: y = 0.73 x + 1.02 ; when x = 13, y = 10.5 , b_{xy} = 0.85, b_{yx} = 0.73, r = 0.79.

(5) Using the following data, obtain the two regression equations.

X:	14	19	24	21	26	22	15	20	19
Y:	31	36	48	39	50	45	33	41	39

Answer: x = 0.557 y -2.28; y = 1.608 x + 7.84.

(6) The following data give the correlation coefficient means and standard deviations of rainfall and yield of paddy in a certain tract:

	Yield per hectare in kgs.	Annual rainfall income
Mean	973.5	18.3
	38.4	2.0

Coefficient of correlation = 0.58

Estimate the most likely yield of paddy when the annual rainfall is 22cm, other factors being assumed to remain the same.

(Hint: Regression of y on x is: y – 973.5 = 11.36 (x – 18.3)

Put x = 22 to estimate y = 1014.7)

Answer:- 1014.7

7. Define Regression State the purpose of the regression
8. Explain about the lines of Regression.
9. Define Regression coefficient. State the properties of Regression coefficient.
10. Distinguish between Correlation and Regression.
11. State the uses of Regression Analysis.

6.9 Reference Books

1. S. C. Gupta: Fundamentals of Statistics.
2. K. Chandra Sekhar: Business Statistics.
3. K. V. Sarma: Statistics made simple, Prentice Hall of India.

Lesson Writer
Prof. M. Koteswara Rao

7. Time Series

Objectives

After completion of this chapter, you should be able to:

- Understand the basic idea of Time Series;
- Know the different components of Time Series data;
- Fit a linear or non linear trend to a time series data;
- Understand about Graphic method.

Structure

7.1 Introduction

One of the major managerial responsibilities is the design and implementation of policies for the achievement of the long – term and short – term goals of the business firm. Previous performances must be studied so as to forecast future business activity. Given a projection of the pattern and the level of future business activity, the desirability of alternative actions can then be investigated. By the time series analysis one can identify the regularity of occurrence of any specific feature over a sufficiently long period of time and this enables on to predict the probable future variations. The actual performance can be compared with the predicted performance and the causes for variations can be analysed. By analyzing a time series, we can identify patterns and tendencies that help explain variation in past sales, shipments, rainfall, or any other variable of interest. Likewise, this understanding contributes to our ability to forecast future values of the variable.

Managers use forecasting techniques to make strategic decisions about selling, buying etc every day. Time series is an important tool that can be used to predict the future. The future is always uncertain, but with the help of past data, an assessment of the future can be made. This is precisely why time series analysis is very important in the fields of economics, sales and production. It also helpful in making predictions about population, national income, capital information, etc.

The main objective in analyzing time series is to understand, interpret, and evaluate changes in economic phenomena in the hope of anticipating the course of future events correctly.

7.2 Definitions:

Arrangement of statistical data in a systematic manner in accordance with occurrence of time is called time series. Most of the series relating to Economics, Business and Commerce, e.g., the series relating to prices, investment, sales and profits, bank deposits etc., are all time series spread over a long period of time. Mathematically, a time series is defined by the functional relationship between the two variables as $y = f(t)$.

According to Ya –lun Chou, “A time series may be defined as a collection of readings belonging to different time period of some economic variable or composite variables. Such as production of steel, per capita income, gross national product, price of tobacco or index of industrial production.”

According to Patterson “A time series consist of statistical data which are collected, recorded or observed over successive increments.”

According to croxton and cowden, "A time series consists of data arrayed chronologically."

Mathematically, a time series is defined by the functional relationship between the two variables as y = f (t).

7.3 Utility of Time Series Analysis:

The utility of time series analysis are

(i) Forecasting: It helps in forecasting. The analysis of past conditions are the basis of forecasting the future behavior of the variable under study. This helps in making future plans of action. Various five-year plans of our country are based on the analysis of past data.

(ii) Analysis: It helps in the analysis of past behavior of a variable. Analysis of past data discloses the effect of different factors on the variable under study. With the help of such analysis the future behavior of the variable under study can be predicted.

(iii) Approximation: It gives approximate indicators.

(iv) Comparison: It helps in making comparative studies. Once the data is arranged in a systematic order, the comparison between one time period and another is facilitated. It provides a scientific basis for making comparisons by studying and isolating the effects of various components of a time series.

(v) Evaluation: It helps in the evaluation of current achievements. The review and evaluation of progress made on the basis of a plan are done on the basis of time series data. The progress of plans is judged by the yearly rates of growth in Gross National Product.

7.4 Components of Time series:- If the values of a variable are observed at different time periods of the time, then the values so obtained will some variations. These variations are due to the fact that the value of the variables is affected not only a single factor but die to the cumulative effect of a multiplicity of factors. For example the price of a particular product depends on its demand, raw materials, investment and so on. The various forces affecting the values of a variable in a time series may be broadly classified in to four categories called the components of time series, which are:

(I) Secular Trend (or) Long – term Movement or simply Trend (T)

(II) Seasonal variations (S)

(III) Cyclical Variations (C)

(IV) Irregular variations (I) or Random Variations (R)

(I) Secular Trend: - Trend is a general tendency of the time series data to increase or decrease or stagnates during a long period of time. This type of the tendency can be observed usually in most of the series relating to Business and Economics. Generally, the upward tendency is usually observed in the time series data relating to Population, prices, income, money circulation, etc. while the down ward tendency would be seen in the time series data relating to deaths, epidemics, etc.

If should be clearly understood that trend is the 'general, smooth, long – term average tendency'. It is not necessary that the increase or decline should be in the same direction throughout the given period. However, the overall tendency may be upward, downward or stable. Also, the 'long period of time' is a relative term and cannot define exactly. It would depend on nature of the data. In some times the period is very small while some times it is very long, which is depend purely on the variable. Trend may be classified as

(a) Linear trend (b) Non – linear trend

If the time series values are plotted on the graph cluster more or less round a straight line, the trend exhibited by the time series is known as Linear trend otherwise if it exhibits curve then it is called Curvi - linear or non – linear trend.

(II) Seasonal variations (S):-

These variations in the time series data are to the rhythmic forces which operate in a regular and periodic manner over a span of less than a year, that is during a period of 12 months and have the same or almost same pattern year after year. So, seasonal variations in a time series data will be there if the data are noted as quarterly, monthly, weekly, daily etc. The seasonal variations may be attributes to the following two cases.

(a) Natural forces:- The various seasons or weather conditions and climatic changes play an important role in seasonal movements. For instance, the sales of umbrella pick up very fast in rainy season, the demand for air conditioners goes up in summer season, the sales of woolens go up in winter.

(b) Man-made variations:- These variations in the time series within a period of twelve months are due to habits, fashions, customs and conventions of the people in the society. For instance, the sales of jewelers and ornaments go up in marriages, sales and profits of a departmental store go up during the festivals period like Dasara, Christmas, Diwali etc.

(III) Cyclical Variations (c):- The oscillatory movements in a time series with period of oscillation greater than one year are termed as Cyclical variations. These variations in a time series are due to ups and down occurring after a period of more than one year. The cyclical variations, though more or less regular, are not intervals of time. One complete period of oscillation is

called a cycle. These oscillatory movements in any business activity are the outcome of the so-called 'Business Cycles' which are the four –phased cycles comprising prosperity (boom), recession, depression and recovery from time to time. Generally, one complete period is assumed as 4 years to 9 years.

(IV) Irregular variations (I) or Random Variations (R):-

These variations do not exhibit any definite pattern and there is no regular period or time of their occurrence. These are accidental changes which are purely random, unforeseen and unpredictable. These variations do not exhibit any definite pattern and there is no regular period of time of their occurrence. Some of the such factors which are influence the data are like floods, wars, famines, earthquakes etc.

7.5 Analysis of Time series: -

The analysis of time series have:

(i) Determining or Identifying the various factors which are shows the variation in the time series data

(ii) Isolating, studying, analyzing and measuring each component independently.

Time series analysis is of great importance to business man and economists.

Some of its uses are

(a) It enables us to study the past behavior of the variable under consideration.

(b) It helps to compare the actual performance with the expected once and analyze the causes of such variation.

(c) It helps to compare the changes in the values of different variable at different times or places.

(d) The separation and study of the various components is paramount importance to business man in the planning of future operations and in the formulation of executive and policy decisions.

7.6 Mathematical Models for Time Series

(i) Additive model (Or) Decomposition by Additive Hypothesis:- According to the additive model. The time series can be expressed as:

$$Y = T + S + C + I \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Or more precisely $Y_t = T_t + S_t + C_t + I_t \ldots\ldots\ldots\ldots\ldots\ldots (2)$

Where Y (Y_t) is the time series value at time t, and T_t, S_t, C_t and I_t represent the trend, Seasonal, cyclical and random variations at time t. In this model S = S_t, C = C_t and I = I_t are absolute quantities which can take positive and negative values so that:

$\Sigma S = \Sigma S_t = 0$, for any year,

$\Sigma C = \Sigma C_t = 0$, for any cycle,

And $\Sigma I = \Sigma I_t = 0$, in long – term period.

The additive model assumes that all the four components of the time series-operate independently of each other so that none of these components has any effect on the remaining three.

(ii) Multiplicative model (or) Decomposition by Multiplicative Hypothesis. Keeping the above points, in view, most of the economic and business time series are characterized by the following classical multiplicative model:

$$Y = T \times S \times C \times I \text{ ---------- (3)}$$

Or more precisely $Y_t = T_t \times S_t \times C_t \times I_t$ ----------(4)

This model assumes that the four components of the time series are due to different causes but they are not necessarily independent and they can affect each other. In this model S, C and I are not viewed as the absolute amounts but rather as relative variations.

(iii) Mixed model: In addition to the additive and multiplicative models, discussed above the components in the time series may be combined in a large number of other ways. The different models, defined under different assumptions, will yield different results. Some of the mixed models resulting from different combinations of additive and multiplicative models are given below:

$$Y = TSC + I \quad \text{- -------(5)}$$

$$Y = TC + SI \quad \text{---------(6)}$$

$$Y = T + SCI \quad \text{---------(7)}$$

$$Y = T + S + CI \quad \text{- --------(8)}$$

If trend component (T) is known, then using multiplicative model, it can be isolated from the given time Series to give

$$S \times C \times I = \frac{Y}{T} = \frac{\text{Original Values}}{\text{Trend Values}}$$

Thus, for the annual data, for which the seasonal component S is not there, we have

$$Y = T \times C \times I \Rightarrow C \times I = \frac{Y}{T}.$$

7.7 Free Hand Curve Method:

Free hand Curve denotes that it is a non – mathematical curve. It is not based on any equation or formula. It can be drawn according to the data, if necessary by using drawing instruments. Free hand curve can be constructed in the following way.

- Take a graph and select the axis and suitable scale.
- Point out the Original data on the graph papers.
- Observe the direction of points carefully.
- Take a transparent rules and draw a smooth line through the central points. Care is to be taken while drawing the line.

Conditions:-

- The curve should be smooth. It can be a straight line or a line with gradual curves.
- The trend line should be drawn on the graph paper in such a way that the area above and below the trend line is equal.
- Vertical deviations above and below the trend should be equal.
- The sum of squares of the vertical deviations of the observations from the trend should be as small as possible.

7.8 Merits and demerits of graphic method.

Merits:-

- It is the simplest method.
- It is easy to draw in a short time
- It is flexible because no rigid formulae are used.
- Quick estimations can be made by using free hand curve method.
- It gives better expression of secular movements than other methods
- A perfectly drawn curve gives more details of the features of the data and also changes.

De – merits:-

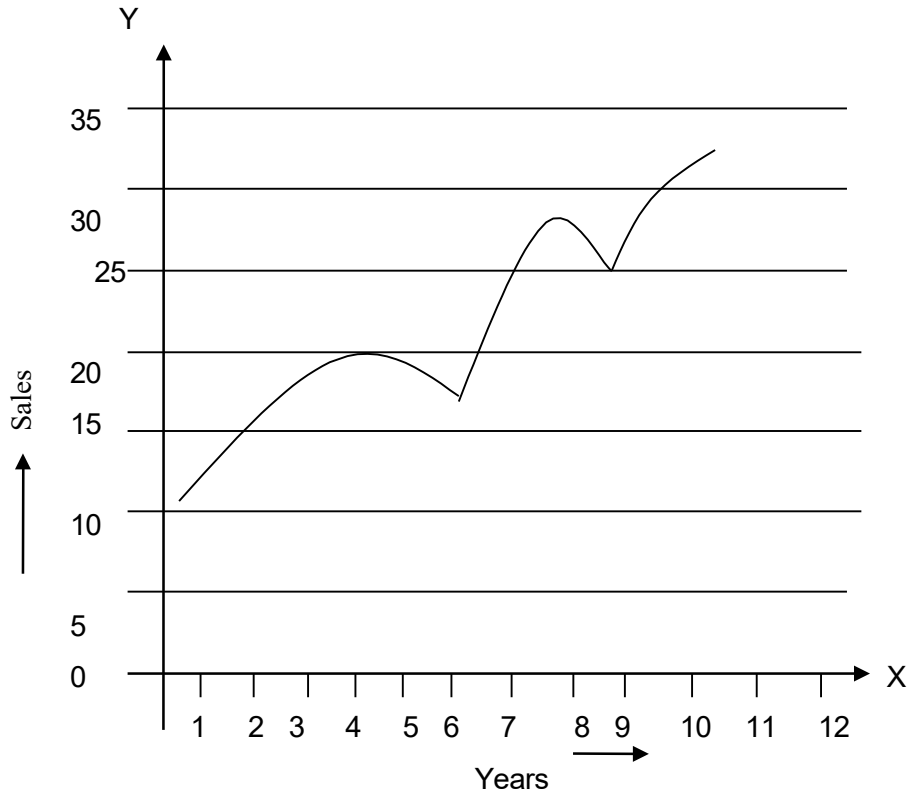
- It is highly subjective because it is very much influenced by the judgment of investigator
- No two persons can draw similar lines because there is no rigid base to draw the line.
- It is not much useful for estimation of future trends.
- Free hand curve cannot assure accuracy, as it has no mathematical formula to construct.

7. 9 Solved Problems:

1) Draw time series graph and fit the trend by free hand method

Year	Sales(Rs. 000)
1994	12
1995	15
1996	18
1997	20
1998	18
1999	22
2000	24
2001	27
2002	25
2003	29
2004	32

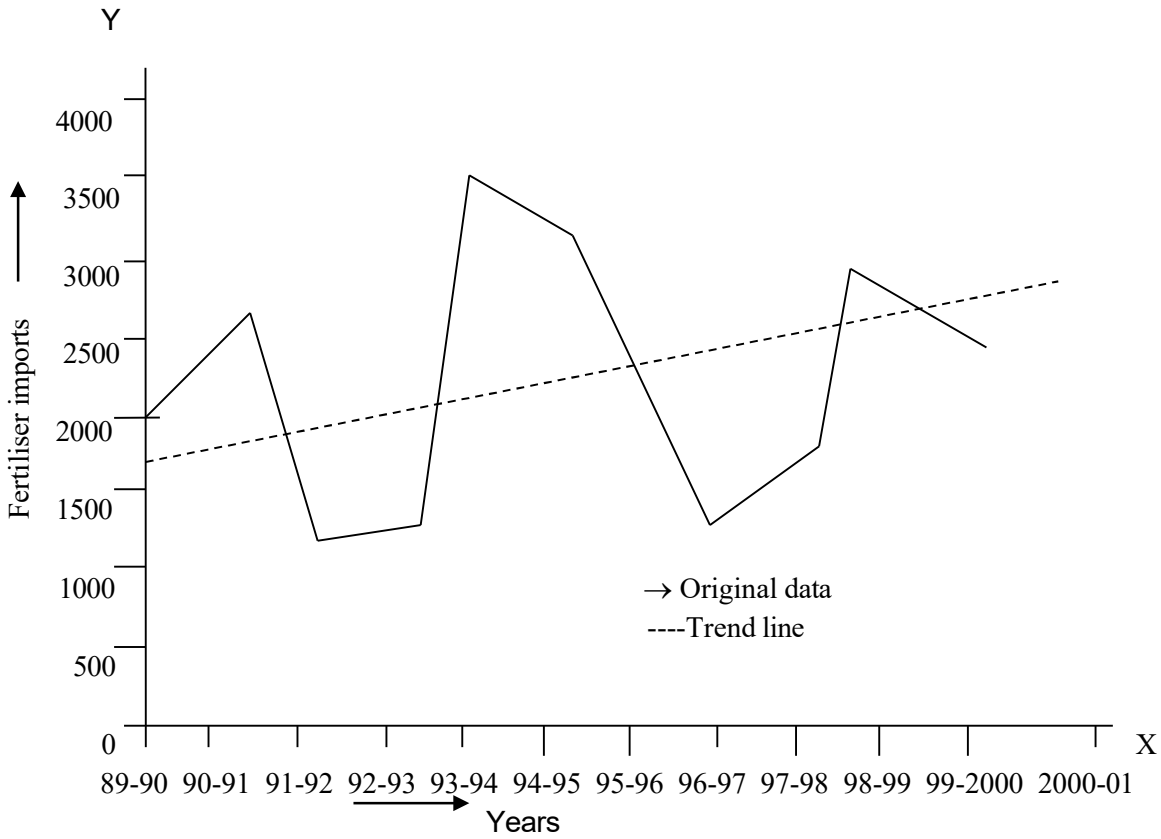
Solution: The Graph for the above data is



2) Determine the trend of the following time series data by graphical method.

Years	Fert. Imports
1989 – 90	2005
1990 – 91	2759
1991 – 92	2041
1992 – 93	1132
1993 – 94	1355
1994 – 95	3624
1995 – 96	3399
1996 – 97	2310
1997 – 98	984
1998 – 99	1608
1999 – 2000	3114
2000 – 2001	2758

Solution:- The Graph for the above data is



7.10 Summary:-

Time series data are data that have been gathered at regular intervals over a period of time. It is generally believed that time series data are composed of four components – trend, seasonality, cyclical effects and random variations. Trend is the long - term general direction of the time series data. Seasonal effects are patterns or cycles of data behavior that occur over time periods of less than one year. Cyclical effects are the business and economic cycles that occur over periods of more than one year. Irregular or random fluctuations are unaccounted for variations that occur over short or long periods of time. Trend can be analyzed using the Curve fitting method and semi average method.

7.11 Self – Assessment Questions:-

1. What is meant by Time series? Quote some of its definitions.
2. What are the components of time series? Explain.
3. Explain the utility of Time Series.
4. Fit a trend line by the Free hand curve method to the following data.

Year	2010	2011	2012	2013	2014	2015
Production (in000 tones)	100	110	90	106	112	115

5. Explain about the Free hand curve method.
6. State the merits and demerits of Graphic method.
7. Use free hand curve method to obtain trend curve to the following data.

Year	2002	2003	2004	2005	2006	2007	2008	2009
Value	65	87	74	80	68	79	62	60

7.12 Reference Books

4. S.C. Gupta, Fundamentals of Statistics, Himalaya Publishing House.
5. Digambar Patri., D.N. Patri, Quantitative Techniques, Kalyani publications.
6. P.N. Arora and S. Arora, Statistics for Management: S.Chand & Comp.Ltd.
7. G.V. Shenoy, Uma K. Srivastava, S.C. Sharma.: Business Statistics
8. B.M. Agarwal,: Business Statistics
9. Gupta S.P.: Statistical Methods

Lesson Writer

Dr. J. Pratapa Reddy

8. Time Series – Trend Methods

Structure

8.1 Introduction

Moving averages method and semi – averages method both are depend on arithmetic mean. Free hand curve method does not help to measure the trend. This drawback can be overcome with the averages method. These methods do not depend on the personal bias or judgment of the investigator. Moving averages method is a method for computing trend values in a time series which eliminates the short term and random fluctuations from the time series by means of moving average. Semi – averages method is a simple to apply, the fact that only two plotted points are used in its construction leads to the general feeling that it is un representative. In the moving averages method, if the period of the averages are chosen appropriately, it will show the true nature of the trend, whether linear or non – linear. In fact, the moving averages method is a logical extension of the semi – averages method.

8.2 Moving Averages Method:

Under this method trend line is fitted on the basis of moving averages. They are computed by using the technique of arithmetic mean. To determine trend, a series of averages are computed for overlapping years in such a way that there are as many averages as there are items in the series, except at the two extreme points. Averages as a series for every data are calculated which are moving from the top. Hence this technique is called moving averages.

Arithmetic averages are calculated for group of years. Each group may consist of 3, 4, 5, 6, 7 or 9 years. According to the period selected for calculation moving averages are calculated. For three yearly moving averages, three items are grouped into one, for 4 yearly moving average 4 items make a group etc. Grouping will be done from the first item of the series with second item and takes second, third and forth items under 3 yearly moving averages.

Moving averages is applied to eliminate the fluctuations and to determine general trend of the series. If the moving average is able to cover the periodicity in the series, then the technique, moving averages can eliminate all regular and irregular fluctuations. To use the technique more effectively in time series analysis, the following factors are to be considered:-

1. If original data gives straight line, then the trend line of moving averages reproduce the original line.
2. If the original series gives a curve, which is concave, the moving average curve will be below it.
3. If the original series gives a convex curve the moving average curve will be above it.
4. Moving averages can never eliminate completely the erratic movements.

Before computing moving averages it is necessary to determine the period of moving averages. Moving averages are calculated for odd years such as 3, 5, 7, 9 and only two even years 4 and 6. The period is determined on the basis of periodicity of the cycle in the data. On the basis of period grouping is done to compute arithmetic mean. Each group consists of equal number of items. The greater the number of items included in the groups for computing moving averages, the greater will be the blanks in the moving averages on both the ends of the series.

It is the arithmetic mean of each group. Therefore the formula of arithmetic Mean.

$$\overline{X} = \frac{X_1 + X_2 + X_3 + X_4 + ... + X_n}{N}$$

symbolically it can be represented as :-

$$\text{3 year moving averages} = \frac{X_1 + X_2 + X_3}{3}, \frac{X_2 + X_3 + X_4}{3}, \frac{X_3 + X_4 + X_5}{3} .....$$

$$\text{5 yearly M. V} = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}, \frac{X_2 + X_3 + X_4 + X_5 + X_6}{5}, \frac{X_3 + X_4 + X_5 + X_6 + X_7}{5}$$

Moving averages are computed in a tabular form for easy location of averages to plot on a graph paper. In the respective columns the values are recorded. For a group of odd numbers such as three yearly or five yearly moving averages, the totals and the averages are placed against the middle item. If the number of items is even in the groups, the totals and averages are placed against the middle points in such a way that in case of four items, two items are above the total and averages and two items are below the total and averages. Finally the averages are placed against the mid item of the respective groups. This process of averaging is called centering.

8.3 Merits of Moving Averages Method:-

1. Moving averages is the simplest method of all mathematical techniques of fitting a trend line.
2. As it is basically arithmetic mean, it is simple to understand and easy to apply to any type of series.
3. It is objective in nature. This method gives same trend line whosoever applies the technique. Investigator's bias or judgment does not influence the trend values.
4. The method is flexible. If some more values are added or deleted to the series, entire calculations are not changed. It only extends the line.
5. Moving averages follow the general pattern of movement of values and the shape of trend curve is determined by the moving averages. The investigator need not make decisions in this regard.

6. For irregular series, moving averages is the most effective method to determine trend values and fit the trend.

7. Moving averages can be used to determine seasonal, cyclical and irregular variations.

8.4 Limitations of Moving Averages Method:-

The technique of moving average has the following limitations, which restrict its popularity and applications.

1. The Prominent criticism against moving averages is that the trend values cannot be calculated for all the years. If the period of averages increases, the number of trend values decreases. In case of seven yearly moving averages, six years do not have trend values. If it is five years moving averages, four years are left.

2. There is no mathematical formula to determine the period of moving averages. This entirely depends on the judgment of the investigator. If the period does not coincide with the cycle the technique cannot fulfill the purpose of determining suitable trend values.

3. Moving averages is less useful for forecasting or further analysis because it is not based on mathematical equation value obtained by this method is not expressed by mathematical relationship.

4. In theory if the moving average coincide with the period of the cycle in the data, variations, are eliminated. But in practice, duration of cycle cannot be equal for any business or economic series of values. Thus moving average cannot eliminate fluctuations.

5. If the trend is not a straight line then moving average trend line lies above or below the trend line.

6. Under this method, trend values are affected by extreme values. It is sensitive to the freakish movements in the data.

8.5 Semi Average Method:

Under this method, the trend line is fitted to a time series basing upon the average values (usually arithmetic mean) of its halves called semi-averages. For this, the entire series is divided into two halves, leaving aside the values of the middle period, if there are odd number of periods in the series. The average value of the respective first half portion of the series is represented by $\overline{X_1}$ and that of the second half portion by $\overline{X_2}$. These two averages are placed against the mid-point of the respective halves of the series. Then, a trend line is drawn basing upon these two semi- averages $\overline{X_1}$ and $\overline{X_2}$ in a straight linear manner. The trend values for

the different times are determined by locating the points on this line. The trend values for any easier and later period.

Alternatively, the trend values for the different years can also be obtained by adjusting the average change between the two semi averages (i.e., $\overline{X_c} = \left(\frac{X_2 - X_1}{N} \right)$ to any of the semi average value in accordance with the times of deviation from the time of any of the semi-averages. Hence, the trend value for any year is computed by

$$T = \overline{X_1} \text{ (or) } \overline{X_2} + \left(\frac{X_2 - X_1}{N} \right).x$$

Where, T = trend value to be computed for any year

$\overline{X_1}$ = Semi average of the first half of the series.

$\overline{X_2}$ = Semi average of the second half of the series.

N = time difference between $\overline{X_1}$ and $\overline{X_2}$

And x= Time deviation from $\overline{X_1}$ and $\overline{X_2}$ as the case may be.

8.6 Merits of Semi – Averages method:

The merits of Semi – Averages method are:

(i) It is a simple and easy method

(ii) The trend figures are objective in the sense that any two persons will get the same trend line from a set of figures.

(iii) The line can be extended to obtain future (or) past estimates.

(iv) It is easy to understand as compared with the moving averages method or least square methods.

8.7 Limitations of semi- average method:

(i) This method assumes the presence of linear trend which would not be true in many cases.

(ii) The trend values detained by this method and the predicted values for future are not precise and reliable.

(iii) In this method, trend is affected appreciably by higher or extreme values. The use of AM is also questioned.

8.8 Solved Problems

1. What is a trend in a time series? The following table gives the annual sales (in Rs. 'ooo) of a commodity?

Year	Sales	Year	Sales
1990	710	1996	644
1991	705	1997	783
1992	680	1998	781
1993	687	1999	805
1994	757	2000	805
1995	629		

Determine the trend by calculating the 5 yearly moving average.

Solution:- Calculation of trend by 5 yearly moving average

Year	Sales	5-Yearly moving Total	5- Yearly moving average
1990	710		
1991	705		
1992	680	3539	$\frac{3539}{5} = 707.8$
1993	687	3458	$\frac{3458}{5} = 691.6$
1994	757	3397	$\frac{3397}{5} = 679.4$
1995	629	3500	$\frac{3500}{5} = 700$

Year	Sales	5-Yearly moving Total	5- Yearly moving average
1996	644	3594	$\frac{3594}{5} = 718.8$
1997	783	3642	$\frac{3642}{5} = 728.4$
1998	781	3885	$\frac{3885}{5} = 777$
1999	805		
2000	872		

2. Calculate 5-yearly moving averages for the following data.

Year	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Value (‘ooo Rs)	123	140	110	98	105	133	95	105	150	135

Solution:- Computation of 5-yearly moving averages.

Year	Value (‘000 Rs)	5-Yearly moving Total (‘000 Rs.)	5- Yearly moving average (‘000 Rs.)
1991	123	-	-
1992	140	-	-
1993	110	575	115
1994	98	585	117
1995	105	540	108
1996	133	535	107
1997	95	587	117.4
1998	105	618	123.6
1999	150	-	-
2000	135	-	-

3. From the following data, calculate the trend values using 4- yearly moving average:

Year	1989	1990	1991	1992	1993	1994	1995	1996	1997
Value	506	620	1036	673	588	696	1116	738	663

Solution:-

Year	Value	4-Yearly moving Total	2- Yearly moving Total of column	4-yearly centred moving total
1989	506			
1990	620	2835		
1991	1036	2917	5752	$\frac{5752}{8} = 719$
1992	673	2993	5910	$\frac{5910}{8} = 738.75$
1993	588	3073	6066	$\frac{6066}{8} = 758.25$
1994	696	3178	6211	$\frac{6211}{8} = 776.375$
1995	1116	3213	6351	$\frac{6351}{8} = 793.8$
1996	738			
1997	663			

4. For the following data, verify that the 5-yearly weighted moving average with weights 1, 2, 2, 2,1 respectively is equivalent to 4-years centered moving average:

Year	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Sales (in lakhs)	5	3	7	6	4	8	9	10	8	9	9

Solution: 5- yearly weighted moving average would be as under:

Year	sales	5-Yearly moving Total weights 1, 2, 2, 2,1 respectively	5- Yearly weighted moving average.
1989	5	-	-
1990	3	-	-
1991	7	41	$\frac{41}{8} = 5.125$
1992	6	45	$\frac{45}{8} = 5.625$
1993	4	52	$\frac{52}{8} = 6.500$
1994	8	58	$\frac{58}{8} = 7.250$
1995	9	66	$\frac{66}{8} = 8.250$
1996	10	71	$\frac{71}{8} = 8.875$
1997	8	72	$\frac{72}{8} = 9.000$
1998	9	-	-
1999	9	-	-

4- yearly centered moving average would be as under:

Year	sales	4-Yearly moving Total	2- Moving Total of column	4-yearly centred moving total
1989	5	-		-
1990	3	21		
1991	7	20	41	$\frac{41}{8} = 5.125$
1992	6	25	45	$\frac{45}{8} = 5.625$

1993	4	27	52	$\frac{52}{8} = 6.500$
1994	8	31	58	$\frac{58}{8} = 7.250$
1995	9	35	66	$\frac{66}{8} = 8.250$
1996	10	36	71	$\frac{71}{8} = 8.875$
1997	8	36	72	$\frac{72}{8} = 9.000$
1998	9			-
1999	9			-

5. From the following data, fit a trend line, and determine the trend values by the method of semi average. Also, forecast the trend value for the year 2002.

Year	1999	2000	2001	2002	2003	2004
Output in '000 units	20	16	24	30	28	32

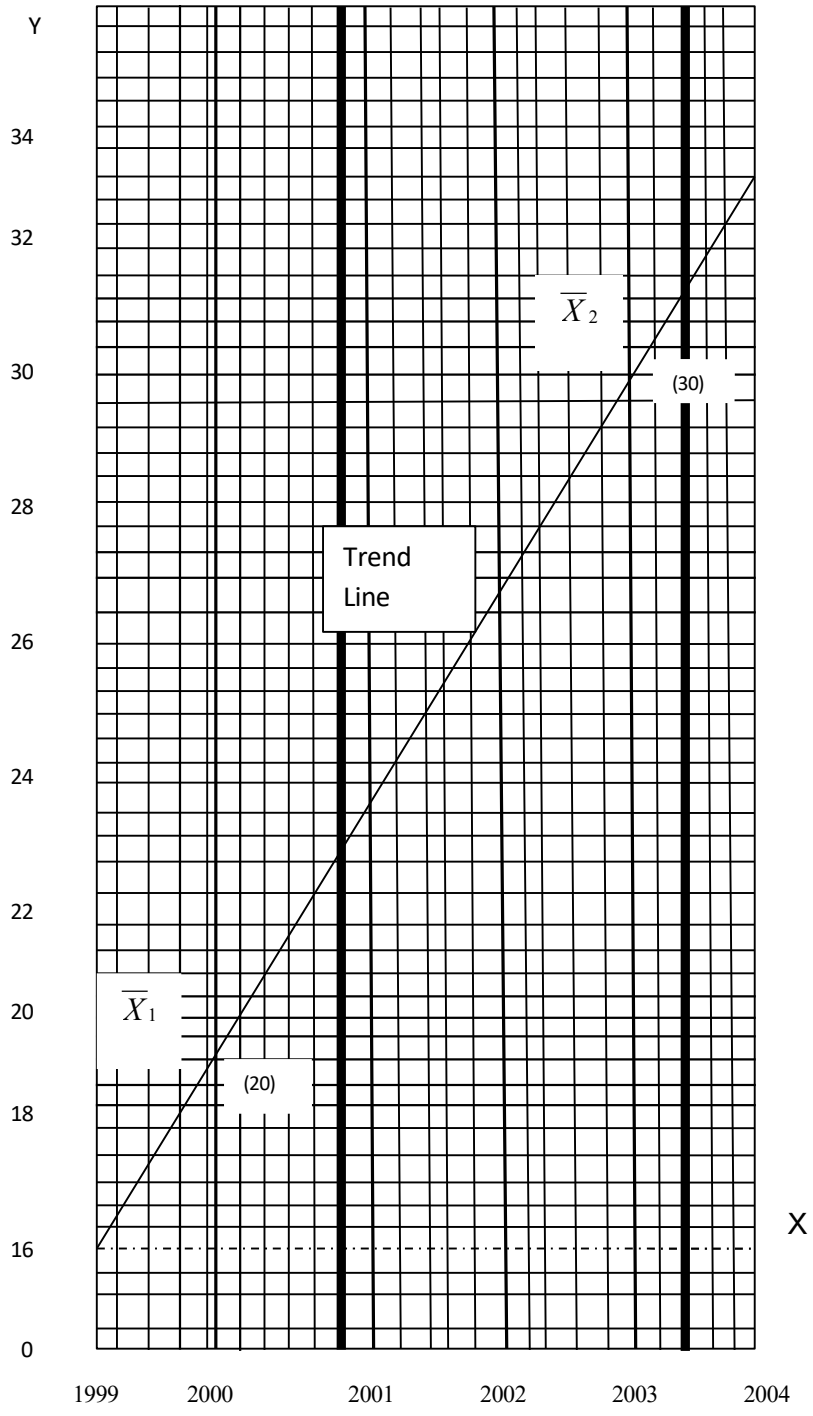
Solution:
Computation of semi -averages.

Year	outputs	Semi totals	Semi-averages	Semi average points
1999	20			
2000	16	60	20	$\overline{X_1}$
2001	24			
2002	30			
2003	28	90	30	$\overline{X_2}$
2004	32			

With the above semi averages of 20 for 2000 and 30 for 2003 the trend line is fitted to the given series as under:

Graphic representation of the trend line by the method of semi averages.

Out put



6. Computation of trend values by location on the trend line

Year	1999	2000	2001	2002	2003	2004	2005
Output in '000 units	15.67	20	23.33	26.67	30	33.33	36.67

Alternatively.
Computation of the trend values by the average change the average is given by

$\overline{X_c} = \frac{\overline{X_2} - \overline{X_1}}{N}$

$\frac{30-20}{3} = 3.33$

Let the average of origin be $\overline{X_1}$ i.e., 20
Thus the trend values will be computed as under:

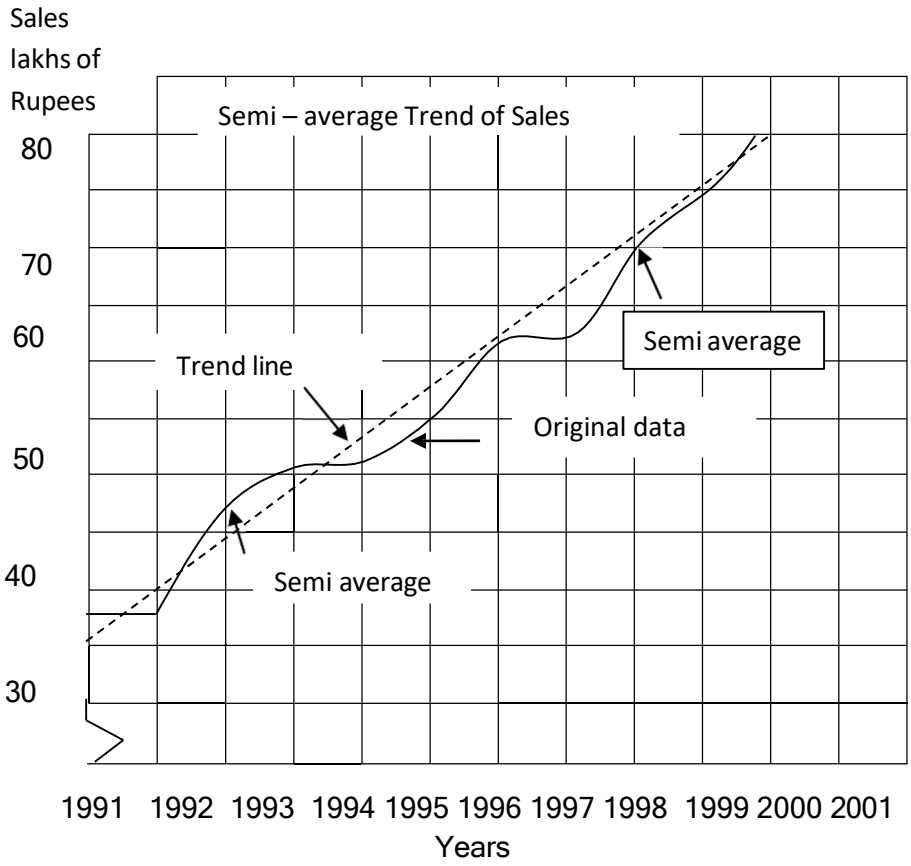
Year	Time dvs. from the time of origin = 2000 (x)	Trend values $T = \overline{X_1} + \left(\frac{\overline{X_2} - \overline{X_1}}{N}\right).x$
1999	-1	20 + 3.33 (-1) = 16.67
2000	0	20 + 3.33 (0) = 20.00
2001	1	20 +3.33 (1) = 23.33
2002	2	20 + 3.33 (2) = 26.67
2003	3	20 + 3.33 (3) = 30.33
2004	4	20 + 3.33 (4) = 33.33
2005	5	20 + 3.33 (5) = 36.67

From the given, it must be noticed that the required trend value for 2002 is 26.67. Further, it must be seen that the trend values computed as above are the same as those located from the trend line shown above.

7. Compute trend by the semi average method of the following data:

Year	Sales (lakhs of Rs.)	Semi-total	Semi- average
1991	38		
1992	40		
1993	46	224	44.8
1994	49		
1995	51		
1996	55		
1997	61		
1998	63	345	69.0
1999	69		
2000	72		
2001	80		

These two semi-averages are plotted in the middle of the respective time spans. Thus 44.8 is plotted against 1993; and 69.0 against 1999. These two points are then connected by a straight line as shown in following figure.



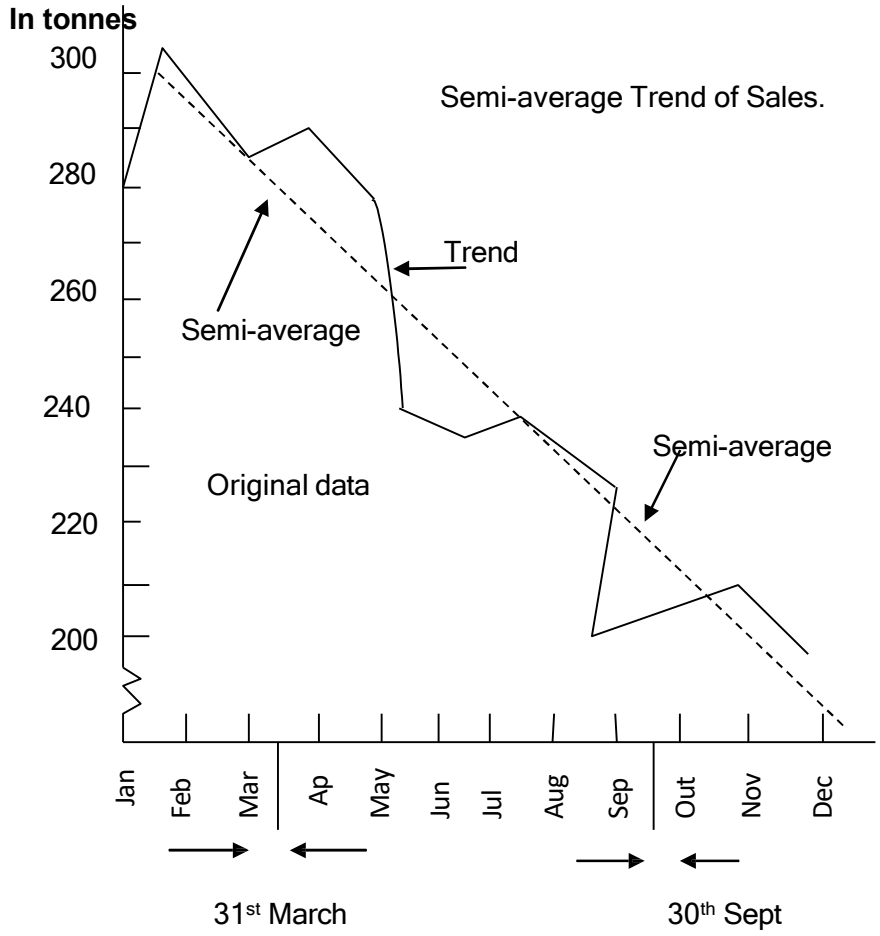
8. The sale of a commodity in ton's varied from January 2000 to December 2000 in the following manner:

280	300	280	280	270	240
230	230	220	200	210	200

Fit a trend by the method of semi-average.

Solution:

	Sales	Semi - average	Sales	Semi - average	
January	280		July	230	
February	300		August	230	
March	280		September	220	
		275			215
April	280		October	200	
May	270		November	210	
June	240		December	200	



8.9 Summary

The method of moving average is another way of obtaining trend. Beginning with a certain number of time periods, average is calculated and then successive averages are calculated by dropping the first of the values and including the next one. For time series that have a significant seasonal effect, the moving averages technique is generally preferred. When moving averages are used for identifying trend components, the period of the average must coincide with the cycle of the data being analyzed. This is done in order to remove possible cyclical fluctuations. Even-period moving averages must be centred in order that their values coincide with actual time points.

8.10 Self-Assessment Questions

1. Explain the method of semi-averages method.
2. Explain about the moving average method.
3. State the merits and limitations of semi-averages method.
4. State the merits and demerits of moving averages method.
5. Apply three yearly moving averages to obtain the trend-free series for the years 2 to 6.

Year	1	2	3	4	5	6	7
Exports (Rs. Lakhs)	126	130	137	141	145	155	159

(Ans: The trend values for the years 2 to 6 are: 131,136,141,147,153 respectively)

6. Calculate four yearly moving averages from the following data relating to production of tea in a certain tea estate:

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Production (kg.)	464	515	518	467	502	540	557	571	586	612

(Ans: The 4 yearly moving averages for 2003 to 2008 are: 495.70, 503.60, 511.60, 529.50, 553 and 572.50 respectively)

7. The following table gives the annual sales (in Rs. '000) of a commodity:

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Sales	710	705	680	687	757	629	644	783	781	805	872

Calculate Five yearly moving averages.
(Ans: The five yearly moving averages from 2002 to 2008 are: 707.80, 691.6, 679.4, 700, 718.8, 728.4 and 777 respectively).

8. Find the trend line to the following series by the method of semi averages.

Day	Sun	Mon	Tues	Wed	Thur	Fri	Sat
Sales (in Rs. '000)	125	130	135	140	105	110	120

8.11 Reference Books:

1. S. C. Gupta, Fundamentals of Statistics, Himalaya Publishing House.
2. Digambar Patri., D. N. Patri, Quantitative Techniques, Kalyani publications.
3. P. N. Arora and S. Arora, Statistics for Management: S. Chand & Comp. Ltd.
4. G. V. Shenoy, Uma K. Srivastava, S. C. Sharma: Business Statistics
5. B. M. Agarwal: Business statistics
6. Gupta S. P.: Statistical Methods

Lesson Writer
Dr. J. Pratapa Reddy

9. Time Series – Methods of Curve fitting

Objectives

After completion of this chapter, you should be able to:

- Understand the basic idea of principles of least squares;
- Know how to find the trend values;
- Understand about fitting of different curves.

Structure

9.1 Introduction:

The principle of least squares provides us mathematical or an analytical device to find an objective fit to the trend of the given time series. Most of the time series data relating to business and commerce conform to definite laws of growth or decay and accordingly in such a situation analytical trend fitting will be more reliable for forecasting and predictions. This technique can be used to fit linear as well as non-linear trends. Curve fitting method of least squares is the most commonly employed and a very satisfactory method to describe a trend is to use a mathematical equation. The type of equation chosen depends upon the nature of the variable under study.

9.2 Straight line method of least squares:

The equation of the straight line is of the form Y = a + bx

where Y = values of the variable

X = time

a and b are constants.

By the method of principles of least squares, the constant values a and b can be determined by assuming the error sum of squares is minimum. The error is E = (y – a - bx)

Squaring an both sides and taking summation over i from 1 to n,

$S = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (Y_i - a - bX_i)^2$ is error sum of squares.

Differentiating 'S' partially, with respect to a and b and equating to zero, then we get Normal equations as

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

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

If the no. of time periods are add then the deviations are taken from the middle value is

$$x = X - A$$

If the no. of time periods are even then we define 'x' as

$$x = \frac{X - (\text{Average of middle two years})}{h/2}$$

where h = width of the years.

Hence, from the normal equations the values of a and b are obtained as

$$a = \bar{Y} = \frac{\Sigma Y}{n} \text{ and } b = \frac{\Sigma XY}{\Sigma X^2}$$

9.3 Second degree parabola of Least squares

The equation of line second degree parabolic curve is $Y = a + bX + cX^2$

where a, b and c are the constants.

The constant values can be estimate by the method of principles of least squares technique. According to the principles of least squares technique the error sum of squares is minimum. The error is $E = Y - a - bX - cX^2$

Squaring on both sides and taking summation over i from 1 to n,

$$S = \sum_{i=1}^{n} \Sigma_i^2 = \sum_{i=1}^{n} \left(Y_i - a - bX_i - CX_i^2 \right)$$

Differentiate 'S' partially with respect to a, b and c, and equating them to zero, we get the Normal equations as

$$\Sigma Y = na + b\Sigma X + \Sigma X^2$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3$$

$$\Sigma X^2 Y = a\Sigma x^2 + b\Sigma X^3 + c\Sigma X^4$$

Solving the above equations, we get the constant values a, b and c; with these values $Y = a + bx + cx^2$ is the best fitted second degree parabolic curve for the given time series data.

9.4 Exponential Curve of Least Squares:

The exponential curve is given by the equation $Y = ab^X$.

The above equation is not in the linear form. So, convert it in to linear form by applying 'log' of both on sides.

$\log Y = \log (ab^X)$

$= \log a + \log b^X$

$\log Y = \log a + x \log b$

$V = A + B U$, is the linear form.

Where $V = \log Y$

$A = \log a \Rightarrow a = A.\ L\ (A)$

$U = X$

$B = \log b \Rightarrow b = A.L\ (B)$

The constant A and B are obtained by the technique of principles of least squares. The Normal equations are $\Sigma V = nA + B\Sigma U$

$$\Sigma UV = A\Sigma U + B\Sigma U^2$$

Solving the above two equations, we can find A and B values and then calculate a and b with these values of a and b, the equation $Y = ab^X$ is the best fitted exponential curve for the given time series data.

9.5 Merits and de-merits of trend fitting by principle of least squares:-

Merits:- The method of least squares is the most popular and widely used method of fitting mathematical functions to a given set of observations. It has the following advantages:

i) Because of its analytical or mathematical character, this method completely eliminates the element of subjective judgment or personal bias on the part of the investigator.

ii) Unlike the method of moving averages, this method enables us to compute the trend values for all the given time periods in the series.

iii) The trend equation can be used to estimate or predict the values of the variable for any period t in future or even in the intermediate periods of the given series and the forecasted values are also quite reliable.

iv) The curve fitting by the principle of least squares is the only technique which enables us to obtain the rate of growth per annum, for yearly data, if linear trend is fitted. If we fit the linear trend Y = a + bx, where x is obtained from t by change of origin such that $\Sigma x = 0$, then for the yearly data, the annual rate of growth is b or 2b according as the number of years is odd or even respectively.

Demerits:-

i) The most serious limitation of the method is the determination of the type of the trend curve to be fitted, viz., whether we should fit a linear or a parabolic trend or some other more complicated trend curve. Assumptions about the type of trend to be fitted might introduce some bias.

ii) The addition of even a single new observation necessitates all the calculations to be done afresh which is not so in the case of moving average method.

iii) This method requires more calculations and is quite tedious and time consuming as compared with other methods. It is rather difficult for a non-mathematical person to understand and use.

iv) Future predictions or forecasts based on this method are based only on the long term variations, i.e., trend and completely ignore the cyclical, seasonal and irregular fluctuations.

9.6 SOLVED PROBLEMS

1. Fit a straight line for the given data

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

Sol:- Let the straight line be y = a + bx

From the principle of least squares, normal equations are

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \qquad \text{- ------ (1)}$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum x^2 \qquad \text{----------(2)}$$

x_i	y_i	x_i^2	$x_i y_i$
1	1	1	1
3	2	9	6
4	4	16	16
6	4	36	24
8	5	64	40
9	7	81	63
11	8	121	88
14	9	196	126
$\sum_{i=1}^{n} x_i = 56$	$\sum_{i=1}^{n} y_i = 40$	$\sum_{i=1}^{n} x_i^2 = 524$	$\sum_{i=1}^{n} x_i y_i = 364$

From the normal equations (1) & (2) we get

40 = 8a + 56b

364 = 56a + 524b

By solving these equations, we get

a = 0.5452 and b = 0.6364

∴ The best straight line for the given data is

y = 0.5452 + (0.6364) x

(2) Fit a straight line of the form y = a + bx to the following data

x	2	4	6	8	10	12
y	10	14	19	25	31	36

Sol:- Let the straight line be Y = a + bx

Normal equations are $\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$ -------- (1)

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \quad \text{----------(2)}$$

x_i	y_i	x_i^2	$x_i y_i$
2	10	4	20
4	14	16	56
6	19	36	114
8	25	64	200
10	31	100	310
12	36	144	432
$\sum_{i=1}^{n} x_i = 42$	$\sum_{i=1}^{n} y_i = 135$	$\sum_{i=1}^{n} x_i^2 = 364$	$\sum_{i=1}^{n} x_i y_i = 1132$

From the normal equations (1) and (2) we get

135 = 6a + 42b

1132 = 42a + 364b

By solving these equations, we get

a = 3.8

b = 2.67

∴ The best straight line for the given data is

Y = 3.8 + (2.67) x

3) Fit a straight line of the form y = a + bx to the following data Estimate value of y when x = 7

x	1	2	3	4	5	6

y	18	51	90	120	140	150

Sol:- Let the straight line be

$$y = a + bx$$

Normal equations are

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \text{ ----------(1)}$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \text{ ----------(2)}$$

x_i	y_i	x_i^2	$x_i y_i$
1	18	1	18
2	51	4	102
3	90	9	270
4	120	16	480
5	140	25	700
6	150	36	900
$\sum_{i=1}^{n} x_i = 21$	$\sum_{i=1}^{n} y_i = 569$	$\sum_{i=1}^{n} x_i^2 = 91$	$\sum_{i=1}^{n} x_i y_i = 2470$

From the normal equations (1) and (2) we get

$$569 = 6a + 21b$$

$$2470 = 21a + 91b$$

By solving these equations, we get

$$a = -0.8667, \quad b = 27.34$$

∴ The best straight line for the given data is

$$y = -0.8667 + (27.34)\, x$$

If x = 7, then estimated value of y is

$$y = -0.8667 + (27.34)7$$

$$= 190.5133$$

4) Fit a second degree parabola for the following data

x	0	1	2	3	6

y	1	1.8	1.3	2.5	6.3

Sol:- Let the second degree parabola be

$$y = a + bx + cx^2$$

Normal equations are

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 \qquad \text{--------(1)}$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 \qquad \text{---------- (2)}$$

$$\sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 \qquad \text{---------- (3)}$$

x_i	y_i	x_i^2	x_i^3	x_i^4	$x_i y_i$	$x_i^2 y_i$
0	1	0	0	0	0	0
1	1.8	1	1	1	1.8	1.8
2	1.3	4	8	16	2.6	5.2
3	2.5	9	27	81	7.5	22.5
4	6.3	16	64	256	25.2	100.8
$\sum_{i=1}^{n} x_i$ =10	$\sum_{i=1}^{n} y_i$ =12.9	$\sum_{i=1}^{n} x_i^2$ =30	$\sum_{i=1}^{n} x_i^3$ =100	$\sum_{i=1}^{n} x_i^4$ =354	$\sum_{i=1}^{n} x_i y_i$ =37.1	$\sum_{i=1}^{n} x_i^2 y_i$ =130.3

From the normal equations

$12.9 = 5a + 10b + 30c$

$37.1 = 10a + 30b + 100c$

$130.3 = 30a + 100b + 354c$

By solving these equations, we get

$a = 1.42$

$b = -1.07$

$C = 0.55$

∴ The best fit of parabola for the given data is

$$Y = 1.42 - 1.07x + 0.55\,x^2$$

5) Speed and resistance of a train is given below

Fit a parabola and estimate resistance of a train if speed is 140 Km/hr

Speed in (Km/hr)	20	40	60	80	100	120
Resistance	5.5	9.1	14.9	22.8	33.3	46.0

Sol:- Let parabola be

$$Y = a + bx + cx^2$$

Normal equations are

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 \qquad \text{-------- (1)}$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 \qquad \text{---------- (2)}$$

$$\sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 \qquad \text{---------- (3)}$$

x_i	y_i	x_i^2	x_i^3	x_i^4	$x_i y_i$	$x_i^2 y_i$
20	5.5	400	8000	160000	10	2200
40	9.1	1600	64000	2560000	364	14560
60	14.9	3600	216000	12960000	894	53640
80	22.8	6400	512000	40960000	1824	145920
100	33.3	10000	10,00,000	100000000	3330	333000
120	46.0	14400	17,28,000	207360000	5520	662400
$\sum_{i=1}^{n} x_i$ =420	$\sum_{i=1}^{n} y_i$ =131.6	$\sum_{i=1}^{n} x_i^2$ =36400	$\sum_{i=1}^{n} x_i^3$ =35,28,000	$\sum_{i=1}^{n} x_i^4$ =36,4000000	$\sum_{i=1}^{n} x_i y_i$ =12042	$\sum_{i=1}^{n} x_i^2 y_i$ =1211720

From the normal equations

$131.6 = 6a + 420b + 364000c$

$12042 = 420a + 36400b + 3528000c$

$12,720 = 36,400a + 352800b + 364000000c$

By solving these equations, we get

$a = 4.35 \quad b = 0.0024 \quad c = 0.0028$

∴The best fit of parabola for the given data is

$$Y = 4.35 + (0.0024)\, x + (0.0028)x^2$$

if x = 140 Km/hr

$Y = 4.35 + (0.0024)\, 140 + (0.0028)\, (140)^2$

$= 59.566$

6. Fit a linear trend to the following data by the least squares method.

Year :	1990	1992	1994	1996	1998
Production:	18	21	23	27	16

Also estimate the production for the year 1999.

Sol:- Here n = 5 i.e., odd.

Hence, we shift the origin to the middle of the time period viz., the year 1994.

Let x = t- 1994 ---------- (i)

Let the trend line of Y (Production) on x be:

Y = a + bx (origin 1994)---------- (ii)

COMPUTION OF STRAIGHI LINE TREND

Year (t)	Production ('000 units) (Y)	X = t – 1994	x^2	Xy	Trend values ('000 units) (Ye) = 21 + 0.1x
1990	18	-4	16	-72	21-0.4 = 20.6
1992	21	-2	4	-42	21-0.2 = 20.8
1994	23	0	0	0	21.0
1996	27	2	4	54	21+ 0.2 = 21.2
1998	16	4	16	64	21 + 0.4 = 21.4

	$\Sigma Y = 105$	$\Sigma x = 0$	$\Sigma x^2 = 40$	$\Sigma xy = 4$	

The normal equations for estimating a and b in (ii) are:

$\Sigma Y = na + b\Sigma x$ and $\Sigma XY = a\Sigma x + b\Sigma x^2$

105 = 5a + b × 0 4 = a × 0 + b × 40

a = $\frac{105}{5} = 21$ b = $\frac{4}{40} = \frac{1}{10}$ = 0.1

Substituting in (ii) the straight line trend equation is given by:

Y = 21 + 0.1x, (origin: 1994) ----------(iii)

[x units = 1 year and y = production (in '000 units).]

putting x = -4, -2,0,2 and 4 in (iii), we obtain the trend values (y_e) for the years 1990, 1992, ------ 1998 respectively, as given in the last column of the above table.

Estimated production for 1999.

Taking t = 1999 in (i), we get x = 1999 – 1994 = 5.

Substituting x = 5 in (iii), the estimated production for 1999 is given by.

(y_e) 1999 = 21 + 0.1 × 5 = 21 + 0.5 = 21.5 thousand units.

(7) Below are given the figures of production (in thousand tons) of a sugar factory:

Year:	1989	1990	1991	1992	1993	1994	1995
Production:	77	88	94	85	91	98	90

(i) Fit a straight line by the method of 'least squares' and show the trend values.

(ii) what is the monthly increase in production?

(iii) Eliminate the trend.

Sol:-

COMPUTION OF STRAIGHT LINE TREND

Year	Production (in '000 tons) (Y)	X = t-1992	xy	x^2	Trend values (in '000 tons) Ye = 89 + 2x
1989	77	-3	-231	9	83
1990	88	-2	-176	4	85
1991	94	-1	-94	1	87
1992	85	0	0	0	89
1993	91	1	91	1	91
1994	98	2	196	4	93
1995	90	3	270	9	95
Total	$\Sigma Y = 623$	$\Sigma x = 0$	$\Sigma xy = 56$	$\Sigma x^2 = 28$	$\Sigma y_e = 623$

i) Let the straight line trend of Y on X be given by: y = a + bx ---------- (i)

where the origin is July 1992 and x units = 1 year.

The normal equations for estimating a and b in (i) are:

$\Sigma Y = na + b\Sigma x$ and $\Sigma xy = a\Sigma x + b\Sigma x^2$

$\Rightarrow$ a = $\frac{\Sigma y}{n} = \frac{623}{7}$ = 89 and b = $\frac{\Sigma xy}{\Sigma x^2} = \frac{56}{23}$ = 2 [$\because \sum x = 10$].

Hence, the straight line trend is given by the equation:

Y = 89 + 2x (origin: 1992) ---------- (ii)

[x units = 1 year and y = Annual product of sugar (in '000 tons)]

putting x = -3, -2, -1,0,1,2,3 in (ii), we get the trend values for the years 1989 to 1995 respectively and are shown in the last column of the table. It may be checked that $\sum y = \sum y_e$, as required by the principle of least squares.

(x) from (i) it is obvious that the trend values increase by a constant amount 'b' units every year. Thus, the yearly increase in production is 'b' units. i.e., 2 × 1000 = 2000 tons.

Hence, the monthly increase in production = $\frac{2000}{12}$ = 166. 67 tons.

(y) Assuming multiplicative model, the trend values are eliminated on dividing the given values (z) by the trend values (y_e). However, if we assume the additive model, the trend eliminated values are given by ($y - y_e$). The resulting values contain short-term variations and irregular variations. Since the data are annual, the seasonal variations are absent.

ELIMINATION OF TREND

Year	Trend eliminated values based on	
	Additive model ($y - y_e$)	Multiplicative model (y / y_e)
1989	77 -83 = -6	77 ÷ 83 =0.93
1990	88 – 85 = 3	88 ÷ 85 = 1.04
1991	94 – 87 = 7	94 ÷ 87 = 1.08
1992	85 -89 = -4	85 ÷ 89 = 0.96
1993	91 -91 = 0	91 ÷ 91 = 1.00
1994	98 -93 = 5	98 ÷ 93 = 1.05
1995	90 – 95 = -5	90 ÷ 95 = 0.95

8) The sales of a company in million of rupees for the years 1994 – 2001 are given below.

Year	1994	1995	1996	1997	1998	1999	2000	2001
Sales	550	560	555	585	540	525	545	585

i) Find the linear trend equation.

ii) Estimate the sales for the year for the year 1993.

iii) Find the slope of the straight line trend.

Sol:- i) In this case, since n, the number of pairs is even, viz., 8, we shift the origin to the time which is the arithmetic mean of the two middle times, viz., 1997 and 1998 and we take:

$$x = \frac{t - \left(\frac{1997 + 1998}{2}\right)}{\frac{1}{2}.(\text{Interval})} = 2(t - 1997.5) = 2t - 3995 \text{ ----------(i)}$$

Thus taking

t = 1997, we get x = 3994 – 3995 = -1; t = 1996, we get x = 3992 – 3995 = -3. and so on. Let the linear trend equation between y and x be given by:

y = a + bx, x = 2 (t – 1997.5) ---------- (ii)

where x units = $\frac{1}{2}$ year and

y = Annual sales in million of Rs.

COMPUTATION FOR LINEAR TREND

Year (t)	Sales (y)	x = 2 (t – 1997.5)	xy	x^2	Trend values (in million Rs.) Y_e = 555.63 + 0. 21x
1994	550	-7	-3850	49	555.63 -7 × 0.21 = 554.16
1995	560	-5	-2800	25	555.63 –5 × 0.21 = 554.58
1996	555	-3	-1665	9	555.63-3 × 0.21 = 555.00
1997	585	-1	-585	1	555.63 -1 × 0.21 = 555.42
1998	540	1	540	1	555.63 +1 × 0.21 = 555.84
1999	525	3	1575	9	555.63 +3 × 0.21 = 556.26
2000	545	5	2725	25	555.63 + 5 × 0.21 = 556.68
2001	585	7	4095	49	555.63 +7 × 0.21 = 557.10
Total	∑y = 4445	∑x = 0	∑xy = 35	$\sum x^2 = 168$	

The normal equations for estimating a and b in (ii) are:

$\Sigma Y = na + b\Sigma x$ | $\Sigma xy = a\Sigma x + b\Sigma x^2$

$\Rightarrow$ 4445 = 8a + 0 | $\Rightarrow$ 35 = a × 0 + 168b

$\Rightarrow$ a = $\frac{4445}{8}$ = 555.63 | $\Rightarrow$ b = $\frac{35}{168}$ = 0.21

Substituting in (ii), The straight line trend is given by the equation:

y = 555.63 + 0.21x ---------- (iii)

Putting x = -7, -5, -3, -1, 1, 3, 5 and 7 in (iii), we get the trend values of sales for the years 1994 to 2001 respectively.

The trend values are shown in the last column of the above table.

ii) The estimated sales for 1993 are obtained on putting t = 1993 in:

x = 2 (t – 1997.5) = 2 (1993 – 1997.5)

= -9.

Substituting in (iii), the estimated sales for 1993 are:

$(y_e)_{1993} = 555.63 + 0.21 \times (-9) = 555.63 - 1.89$

$= 553.74$ million Rs.

iii) The slope of straight line trend (iii) is given by b = 0.21.

9.7 Summary

The linear trend is obtained by fitting a straight line to the given data. It is fitted on the principle of least squares. The trend line enables to isolate the trend component in the given data and its extension into future enables making forecasts. The slope co-efficient of the line indicates the average rate of change over time of the variable. The annual trend equations can be changed on a monthly or quarterly basis, and reverse is also possible. The parabolic trend involves fitting a second degree parabola to the given data. The exponential trend is appropriate where the variable in consideration grows or declines exponentially. The exponential trend involves fitting a straight line to the log values of the variable.

9.8 Self – Assessment Questions

1. Explain the various methods of least squares used in finding the trend values from a time series.

2. Explain in brief the merits and demerits of trend fitting by principle of least squares.

3. Fit a straight line trend to the following data by the method of least squares.

Year	2008	2009	2010	2011	2012	2013	2014
sales	80	90	92	83	94	99	92

(Ans: The fitted straight line trend is: Y = 90 + 2 X)

4. Fit a second degree parabola to the following data.

Year	2011	2012	2013	2014	2015
Profit (in '000Rs)	50	80	60	100	70

(Ans: The fitted second degree parabola is $Y = 80.57 + 6X + 4.29 X^2$)

5. The sales of a company in lakh of rupees for the years 2005 to 2009 are given below:

Year (x)	2005	2006	2007	2008	2009
Sales (y)	65	92	132	190	275

Fit the curve of the type $y = ab^x$ and also estimate the sales for the year 2010.

(Ans: Fitted curve is $Y = 132.7\ (1.435)^x$. The estimated sales for the year 2010 are 392.13 (lakh. Rs)

6. Fit a straight line trend by the method of least squares to the following data:

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Number	110	125	115	135	150	165	155	175	180	200

Also calculated the trend values.

(Ans: The fitted straight line trend is: Y = 151 + 4. 758 X and the trend values are: 108.18, 117.7, 127.21, 136.23, 146.24, 155.76, 165.27, 174.79, 184. 31 and 193.82 respectively).

7. Below are given the annual production of a commodity:

Year	2005	2006	2007	2008	2009	2010	2011
Production (in tons)	70	80	90	95	102	110	115

Fit a straight line trend by the method of least squares

(Ans: The fitted straight line trend is: Y = 94.57 + 7.39 X)

8. Fit a second degree parabolic curve to the data given below and estimate the value for 2016 and comment on it.

Year	2010	2011	2012	2013	2014
Sales (in '000 Rs)	10	12	13	10	8

(Ans: The fitted second degree parabola is: $Y = 12.314 - 0.6\ X - 0.857\ X^2$

The trend values are: 10.086, 12.057, 12.314, 10.857 and 7.686 and the sales in 2016 are -3.798. Since the sales cannot be negative, the given second degree parabolic curve is not good fit to the given data)

9. The sales of a company in lakh of rupees for the years 2005 to 2009 are given below:

Year	2005	2006	2007	2008	2009
sales	65	92	132	190	275

Fit an exponential trend curve $Y = a\ b^x$ to the data.

(Ans: The fitted exponential trend curve is: $Y = (132.7)\ (1.435)^x$.)

9.9 Reference Books

1. S.C Gupta, Fundamentals of Statistics, Himalaya Publishing House.

2. Digambar Patri., D.N. Patri, Quantitative Techniques, Kalyani publications.

3. P.N. Arora and S. Arora, Statistics for Management: A. Chand & Comp. Ltd.

4. G.V. Shenoy, Uma K. Srivastava, S.C.Sharma.: Business Statistics

5. B.M.Agarwal,: Business statistics

6. Gupta S.P.: Statistical Methods

Lesson Writer

Dr. J. Pratapa Reddy

10. PROBABILITY BASICS

Objectives

After completion of this chapter, you should be able to

* Know the relationship between deterministic and probabilistic models.
* Understand the basic concepts of probability
* Explain the basic concepts of set theory

Structure:

10. PROBABILITY BASICS

10.1 Introduction

We know that these are only few things that happen when we know it should happen. When a unique thing happens, and we know the outcome, it is called deterministic, such as a definite chemical reaction or physical law. Sometimes the results cannot be predicted, these are called probabilistic. Most of the managerial decisions are uncertain. In these cases, we are forced to take a chance under certain „risk‟ level. This risk or uncertainty is called probability. Probable origin of the word “Probability”, is from the games of gambling. In simple language, it refers to the chance of happening of an event.

In modern mathematics, the concept of sets is a very useful fundamental development. It has very wide ranging applications in general life, business, management and economics. Many complex problems, the various business decisions are based and resolved by logical interpretation of set theory.

10.2 Deterministic and Non – deterministic relationship:

If an experiment is performed repeatedly under essentially homogeneous and identical conditions, the result or outcome of the experiment is unique or certain then it is said to be deterministic. In a deterministic situation, the conditions under which an experiment is performed, uniquely determine the outcome of the experiment.

Ex (i) In case of perfect gas, we have Boyle‟s law which states

Pressure x Volume = Constant

i.e., PV = Constant => $V \propto 1/P$, provided the temperature is constant.

ii) If dilute sulfuric acid is added to zinc, we get hydrogen.

iii) The distance (s) covered by a particle after time (t) is given by

$s = ut + \frac{1}{2} at^2$

All the above cases the result can be known and certain.

If an experiment is performed repeatedly under essentially homogeneous and identical conditions, the result or outcome of the experiment is not known in advance, but the result may be one of the several possibilities then it is said to be probabilistic. In a probabilistic situation, the result is uncertain and cannot be predicted certainly.

Ex (i) If an electric bulb has lasted 5 months, nothing can be said about its future life.

(ii) A producer cannot ascertain the future demand of his product with certainty.

(iii) In toss of a uniform coin, we are not sure if we shall get head or tail.

In all the above cases the result is not known exactly. Even in our day-to-day life we hear phrases like “It may rain today”; “Probably I will get a first class in the semester examination” etc. In all these cases there is involved an element of chance or uncertainty.

10.3 Permutations and Combinations

Permutation is nothing but “Arrangement”. It is denoted by “P”. If we have „n‟ different objects then „r‟ different objects can be arranged into ${}^nP_r$ ways. It is given as

$${}^nP_r = \frac{n!}{(n-r)!}$$

Here $n = n(n-1)(n-2)\ldots\ldots 5 \times 4 \times 3 \times 2 \times 1$

= product of “n” objects or terms

Combination is nothing but “Selection”. It is denoted by “c”. If we have “n” objects then “r” objects can be selected into ${}^nC_r$ ways. It is given as

$${}^nC_r = \frac{n!}{r!(n-r)!}$$

Ex.:

1. ${}^5P_2 = \frac{5!}{(5-2)!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1} = 5 \times 4 = 20$

2. ${}^5C_2 = \frac{5!}{2!(5-2)!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1 \times 3 \times 2 \times 1} = 10$

3. 0! = 1

4. 1! = 1

5. The number permutations of "n" different objects taken "r" at a time with repetitions is ${}^nP_r = n^r$.

10.4 Basic terminology in Probability

RANDOM EXPRIMENT: "If an experiment repeatedly conducted under essentially homogeneous conditions, the result is not known in advance or not unique but it may be one of the several possibilities" is called Random Experiment.

TRIAL: Performing of a random experiment is called a "Trial".

Ex: Tossing a coin, throwing a Dice etc.

EVENT: The outcome or combination of outcomes in a trial is called "Event".

Ex: i) Tossing a coin is a Trial and getting a head or tail is an event.

ii) Throwing a dice is a trial and getting any one of the faces 1, 2, 3, 4, 5 or 6 is an event.

Exhaustive Events: The total number of possible outcomes of random experiment is called "Exhaustive Events or Cases".

Ex: i) In a toss of single coin, head and tails are exhaustive cases. If two coins are tossed the exhaustive cases are $4 = 2^2$ (HH, HT, TH, TT). If three coins are tossed then the exhaustive cases are $8 = 2^3$. In general, if "n" coins are tossed then exhaustive cases are 2^n.

ii) In a throw of single dice, the exhaustive cases are 6 (1, 2, 3, 4, 5, 6). If two dice are thrown then the exhaustive cases are 6^2. In general, if „n" dice are thrown then the exhaustive cases are 6^3.

Mutually Exclusive events: Events are said to be mutually exclusive or disjoint, if the happening of any one of them excludes the happening of all others in the same experiment.

Ex: In toss of a coin, the events „head" and „tail" are mutually exclusive. In throw of dice, all the six faces are mutually exclusive.

Note: If A and B are mutually exclusive events then $A \cap B = \phi$

Equally likely cases: Events are said to be Equally likely or equally probable, if none of them is expected to occur in preference to other.

Ex.: Tossing a coin, all the outcomes (head and tail) are equally likely. Throwing a dice all the outcomes (1, 2, 3, 4, 5, 6) are equally likely.

Independent events: Events are said to be independent, if the happening of one of them is not effect the happening of any one of the other.

Sample Space: The totality of all possibilities of a random experiment is called sample space. It is denoted by „S".

10.5 Basic concepts of Set theory

A set is a well defined collection or aggregate of objects having given properties and specified according to a well defined rule. Sets are generally denoted by the capital letters of alphabet, like A, B, C etc.

If „1" is an element of the set then it is denoted by $1 \in A$ and „1" is not an element in the set then it is denoted by $1 \notin A$.

A set having no element at all is called a Null set or empty set. It is denoted by "ϕ".

A set "A" is said to be a proper subset of B, if every element of A is also an element of B and there is atleast one element of B which is not an element of A and we write $A \subset B$. Every set is a subset of itself.

Two sets A and B are said to be equal, if every element of A is an element of B and if every element of B is an element of A, then A and B are said to be equal.

A set describing all the objects that are possible points of interest in a problem is called universal set.

10.6 Algebra of Sets

(i) Union of the sets: If A and B are two sets then the union of the two sets is denoted by $A \cup B$ and is defined as "A set of elements which belong to either A or B or both". If x is an element then

$A \cup B = \{x : x \in A \text{ or } x \in B\}$

For example, if A = {1, 2, 3} and B = {3, 4, 5} then $A \cup B$ = {1, 2, 3, 4, 5}

(ii) Intersection of the sets: If A and B are two sets then the intersection of the two sets is denoted by $A \cap B$ and is defined as "A set of elements belong to both A and B". If x is an element then

$A \cap B = \{x : x \in A \text{ and } x \in B\}$

For example, if A = {1, 2, 3} and B = {3, 4, 5} then $A \cap B$ = {3}

(iii) Disjoint (or) Mutually exclusive events:

Two sets A and B are said to be mutually exclusive event, if they do not have any common point. If A and B are said to be disjoint if their intersection is a null set i.e., $A \cap B = \phi$.

(iv) Complement of a set: If „A" is an element then complement of a set is denoted by $\bar{A}$ or A' or A^c and is the set of elements which do not belong to the set A. The sets A and $\bar{A}$ are disjoint sets.

(v) Difference of sets: If A and B are two sets then the difference of two sets is denoted by "A – B" and is the set of elements which belong to A but not to B. If x is an element then

$A - B = \{x : x \in A \text{ and } x \notin B\}$

10.7 Laws o set theory:

The different laws of set theory are:

Commutative laws:

$A \cup B = B \cup A$

$A \cap B = B \cap A$

Associative laws:

$A \cup (B \cup C) = (A \cup B) \cup C$

$A \cap (B \cap C) = (A \cap B) \cap C$

Distributive laws:

$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

Difference laws:

$A - B = A \cap \overline{B}$

$A - B = A - (A \cap B) = (A \cup B) - B$

Complementary laws:

$A \cup \overline{A} = S$, $A \cap \overline{A} = \phi$

$A \cup S = S$, $A \cap S = A$

$A \cup \phi = A$, $A \cap \phi = \phi$

Here S is the universal set or sample space.

De Morgan"s laws:

$$(A \cup B)' = A' \cap B'$$
$$(A \cap B)' = A' \cup B'$$

Idempotency laws:

$$A \cup A = A \text{ and } A \cap A = A$$

10.8 Solved Problems

1. If an office, there are 500 persons who smoke and 400 persons drink. There are 200 persons who smoke and drink. Find how many persons do not drink but only smoke.

Sol: Let A be the set of persons who smoke and B is the set of persons who drink.

Given that $n(A) = 500$
$n(B) = 400$
$n(A \cap B) = 200$

$A \cap B$ will be the set of persons who smoke and drink.

$A \cap \overline{B}$ will be the set of persons who smoke but do not drink.

$n(A \cap \overline{B}) = n(A) - n(A \cap B)$
$= 500 - 200$
$= 300$

Hence, only 300 persons only smoke, but do not drink.

2. For sets A = {1, 2, 3, 4, 5, 6}, B = {2, 4, 6, 8}, C = {1, 3, 5, 6}. Determine (i) $A \cap B$ (ii) $A \cup B$ (iii) $A \cap B \cap C$ iv) $A \cup B \cup C$

Sol: Given that A = {1, 2, 3, 4, 5, 6}
B = {2, 4, 6, 8}
C = {1, 3, 5, 6}

(i) $A \cap B = \{1, 2, 3, 4, 5, 6\} \cap \{2, 4, 6, 8\}$
$= \{2, 4, 6\}$

(ii) $A \cup B = \{1, 2, 3, 4, 5, 6\} \cup \{2, 4, 6, 8\}$
$= \{1, 2, 3, 4, 5, 6, 8\}$

(iii) $A \cap B \cap C = \{1, 2, 3, 4, 5, 6\} \cap \{2, 4, 6, 8\} \cap \{1, 3, 5, 6\}$
$= \{6\}$

(iv) $A \cup B \cup C = \{1, 2, 3, 4, 5, 6\} \cup \{2, 4, 6, 8\} \cup \{1, 3, 5, 6\} = \{1, 2, 3, 4, 5, 6, 8\}$

3. A bag contains 4 white, 5 red and 6 green balls. Three balls are drawn at random. Find

i) Exhaustive number of cases

ii) Favorable number of cases of getting one ball of each color.

Sol.: (i) Total number of balls in the bag
$= 4 + 5 + 6 = 15$ balls

Out of 15 balls 3 balls can be selected into $^{15}C_3$ ways.

$\therefore$ Total number of cases (or) exhaustive number of cases = $^{15}C_3$

(ii) Total number of white balls = 4

Total number of red balls = 5

Total number of green balls = 6

Out of 4 white balls one white ball can be selected into $^4C_1$ ways.

Out of 5 red balls one red ball can be selected into $^5C_1$ ways.

Out of 6 green balls one green ball can be selected into $^6C_1$ ways.

Favorable number of cases to select one ball of each color = $^4C_1 \times {}^5C_1 \times {}^6C_1$ ways.

4. A bag contains 20 tickets marks with numbers 1 to 20. One ticket is drawn at random. Compute

(i) Exhaustive number of cases.

(ii) Favorable number of cases to get number on the ticket is a multiple of 2 as well as 5.

Sol: (i) Total number of tickets in the bag are 20.

Out of 20 tickets one ticket can be selected into $^{20}C_1$ ways.

Exhaustive number of cases = $^{20}C_1$

(ii) Cases favorable to get number of the ticket as a multiple of 2 are 2, 4, 6, 8, 10, 12, 14, 16, 18, 20.

$\therefore$ Favorable number of cases = 10 ways.

Cases favorable to get number of the ticket as a multiple of 5 are 5, 10, 15 and 20.

$\therefore$ Favorable number of cases = 4 ways.

5. A party of 3 ladies and 4 gentlemen is to be formed from 8 ladies and 7 gentlemen. In how many different ways can the party be formed if Mrs. A and Mr. B refuse to join the same party?

Sol: Out of 8 ladies, 3 ladies can be selected into $^8C_3$ ways. Out of 7 gentlemen, 4 gentlemen can be selected into $^4C_1$ ways.

The number of ways of choosing the committee = $^8C_3 \times {}^7C_4$ ways.

If both Mrs. A and Mr. B are members, there remain to be selected 2 ladies from 7 ladies and 3 gentlemen from 6 gentlemen.

This can be done into $^7C_2 \times {}^6C_3$ ways.

$\therefore$ The number of ways of forming the party in which Mrs. A and Mr. B refuse to join $= {}^8C_3 \times {}^7C_4 - {}^7C_2 \times {}^6C_3$

$$= \frac{8!}{3!5!} \times \frac{7!}{4!3!} - \frac{7!}{2!4!} \times \frac{6!}{3!3!}$$

$= 1960 - 420$

$= 1540$

6. The question paper of the micro economics contains ten questions divided into two groups of five questions each. In how many ways can an examinee answer six questions taking atleast two questions from each group?

Sol: An examinee answers six questions in the following ways.

(i) 2 questions from I group + 4 questions from II group.

(ii) 3 questions from I group + 3 questions from II group.

(iii) 4 questions from I group + 2 questions from II group.

(i) Two questions can be chosen from I group into $^5C_2$ ways and 4 questions can be chose from II group into $^5C_4$ ways. So, total number of ways of selecting 6 questions from both the groups are $^5C_2 \times {}^5C_4$ ways.

(ii) 3 questions from I group and 3 questions from second group can be selected into $^5C_3 \times {}^5C_3$ ways.

(iii) 2 questions from I group and 4 questions from II group can be selected into $^5C_2 \times {}^5C_4$ ways.

$\therefore$ Total number of ways $= {}^5C_2 \times {}^5C_4 + {}^5C_3 \times {}^5C_3 + {}^5C_2 \times {}^5C_4$

$= 10 \times 5 + 10 \times 10 + 5 \times 10$

$= 200$ ways

7. There are 5 boys and 3 girls. In how many ways can they stand in a row so that no two girls are together?

Sol: Total number of boys = 5

Total number of girls = 3

Since no two girls are to be together, each girl must be placed between two boys. Place the 5 boys that $\star B_1 \star B_2 \star B_3 \star B_4 \star B_5 \star$. In order that no two of the girls be together, they must be placed in the places marked "$\star$". There are six such places and

so the 3 girls can be placed in ${}^6P_3$ ways. Further, the 5 boys can be arranged among themselves in 5! ways.

$\therefore$ Total number of arrangements $= {}^6P_3 \times 5!$

$$= \frac{6!}{3!} \times 5!$$

$$= 14{,}400$$

10.10 Self assessment questions

1) In a class of 25 students, 12 students have taken economics; 8 have taken economics but not politics. Find the number of students who have taken economics and politics. [Ans: 4]
2) Define permutations and combinations with an example.
3) State any five laws of set theory.
4) Distinguish between deterministic and probabilistic relations with an example.
5) If A = {1, 2, 3, 4}, B = {2, 4, 6, 8} and U = {1, 2, 3, ….8} be the universal set then find (i) $A \cup B$ (ii) $A \cap B$ (iii) A' (iv) $(A \cup B)'$ (v) $(A \cap B)'$
   [Ans: (i) $A \cup B = \{1, 2, 3, 4, 6, 8\}$ (ii) $A \cap B = \{2, 4\}$
   (iii) $A' = \{5, 6, 7, 8, 9\}$ (iv) $(A \cup B)' = \{5, 7, 9\}$
   (v) $(A \cap B)' = \{1, 3, 5, 6, 7, 9\}$
6) Find the number of permutations of the word (i) ACCOUNTANT (ii) ENGINEERING.
   [Ans: (i) 2,26,800 (ii) $\frac{1!}{3!3!2!2!}$]

7) For an examination, a candidate has to select seven subjects from three different groups A, B and C. The three groups A, B, C contain 4, 5, 6 subjects respectively. In how many different ways can a candidate make his selection if he has to select at least 2 subjects from each group?
   [Ans: 2700]
8. Two unbiased dice are thrown. Find favorable number of cases of
   (i) Both dice shows the same number
   (ii) First dice shows six.
   (iii) Total of the numbers on the dice is greater than 8.
   [Ans: (i) 6 (ii) 6 (iii) 10]

10.09 Summary

Most of the physical and chemical sciences are of a deterministic nature. However, there exists a number of phenomenal where we cannot make predictions with certainty or complete reliability and are known as probabilistic phenomenon. Such phenomena are frequently observed in Economics, business or even in our day-to-day life.

In almost whole of the business mathematics the set theory is applied in one form or the other. The probability theory based on the basic principles of set theory, permutations and combinations. Permutations and combinations help us, how to arrange and select the items from as group of objects.

10.11 Reference Books

1. S. C. Gupta: Fundamentals of Statistics
2. K. Chandra Sekhar: Business Statistics
3. K. V. Sarma: Statistics made simple, Prentice Hall of India
4. D. C. Sancheti, V. K. Kapoor, Business Mathematics

Lesson Writer
Prof. K. Chandan

11. Probability – Theorems

Objectives:

After completion of this chapter, you should be able to

- Learn the theorems of Probability
- Identify various definitions of Probability
- Understand the concept of condition Probability

Structure:

11. Probability – Theorems

11.1 Introduction

The various approaches to probability have evolved, mainly to cater to the three different types of situations under which probability measures are normally sought. The objective of this chapter is to introduce you to the theory of probability. Accordingly, the various definitions of probability, Addition theorem, multiplication theorems of probability are presented, followed by problems on there theorems. Finally, conditional Probability and problems based on conditional Probability are presented.

11.2 Various definitions of Probability

The various definitions of Probability are

(1) Mathematical or classical or priori definition of probability:

If a random experiment in "n" exhaustive, naturally exclusive and equally likely cases out of which "m" are favorable to the happening of an event A, then the probability of occurrence of A, usually denoted by P(A) and is given by:

P(A) = Favorable number of cases to A / Total number of Cases = m/n

The non – favorable number of cases are "n-m". the probability of non – happening of the event "A" is denoted by $P(\overline{A})$ and is defined

$P(\overline{A})$ = Favorable number of cases to $\overline{A}$ / Total number of Cases = n - m /n

= 1 – m/n = 1- P(A)

$P(A) + P(\overline{A}) = 1$

Limitations:- the limitations of classical probability and fails in the following cases:

(i) the exhaustive number of outcomes of the random experiment is finite.

(ii) If the various outcomes of the random experiment are not equally likely

(iii) If the actual value of "n" is not known.

(2) Statistical or Empirical definition of Probability

If an experiment is performed repeatedly under essentially homogeneous and similar conditions, then the limiting value of the ratio of the number of times the event occurs to the number of trials, as the number of trials becomes large, is called the probability of happening of the event, it being assumed that the limit is finite and unique. If "A" is an event then

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

Here m = Favorable no. of cases

n = Total no. of cases

Limitations: the limitations are

(i) the experimental conditions may not remain essentially homogeneous and identical in a large number of repetitions of the experiment.

(ii) the relative frequency m/n, may not attain a unique value.

(3) Modern (or) Axiomatic definition of Probability

This definition of Probability was introduced by Russian mathematic A.N Kolomogrov in 1930"s. The various Axioms of probability are

(i) Non – Negativity: - If "A" is an event in the sample space "S" then the Probability of the event "A" is always positive is $P(A) \geq 0$

(ii) **Totality**: the total probability is always unity is $P(A) + P(\overline{A}) = 1$

(iii) Additivity: If A and B are two mutually exclusive event then

$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B)$ Since $P(A \cap B) = 0$

11.3 Theorems on Probability

(1) Addition theorem of Probability

If A and B are two events in the sample space „S" then

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

If A, B and C are the events in the sample space „S" then

$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$

If A and B are two mutually exclusive events in the sample space then

$P(A \cup B) = P(A) + P(B)$ $(\therefore P(A \cap B) = 0)$

If A, B and C are mutually exclusive events in the sample space „S" then

$P(A \cup B \cup C) = P(A) + P(B) + P(C)$

(2) $P(\overline{A}) = 1 - P(A)$

(3) $P(\overline{A} \cap B) = P(B) - P(A \cap B)$

(4) $P(A \cap \overline{B}) = P(A) - P(A \cap B)$

(5) If $A \subset B$, then $P(A) \leq P(B)$

(6) $P(\overline{A} \cap \overline{B}) = 1 - P(A \cup B)$

(7) Multiplication theorem of Probability or Compound theorem of Probability:

If A and B are two events then

$P(A \cap B) = P(A) \cdot P(B/A)$ (or)

$P(A \cap B) = P(B) \cdot P(A/B)$

If A,B and C are three events then $P(A \cap B \cap C) = P(A). P(B/A). P(C/A \cap B)$

If A and B are two independent event then $P(A \cap B) = P(A). P(B). P(C)$

If A, B and C are independent events then $P(A \cap B \cap C) = P(A). P(B). P(C)$

(8) If A and B are independent then

(i) A and $\overline{B}$ are also independent

(ii) $\overline{A}$ and B are also independent

(iii) $\overline{A}$ and $\overline{B}$ are also independent

11.4 Conditional Probability

If A and B are two events then

$P(A/B) = P(A \cap B)/P(B)$, $P(B) \neq 0$

And $P(B/A) = P(A \cap B)/P(A)$, $P(A) \neq 0$

Here P (A/B) is the conditional Probability of the happening of the event A, where the event "B" has already happened similarly, P(B/A) is the conditional Probability of the event B, where the event "A" has already happened. The events A and B are independent then

$P(A/B) = P(A \cap B)/P(B) = P(A). P(B) / P(B) = P(A)$

And $P(B/A) = P(A \cap B)/P(A) = P(A). P(B) / P(A) = P(B)$

11.5 Solved Problems

1) What is the probability that a leap year, selected at random, will have 53 Fridays?

Solution:- there are 366 days in a leap year and it has 52 complete weeks 2 days over. There two extra days may be

(i) Sunday and Monday (ii) Monday and Tuesday

(iii) Tuesday and Wednesday (iv) Wednesday and Thursday

(V) Thursday and Friday (vi) Friday and Saturday

(vii) Saturday and Sunday

$\therefore$ No. of exhaustive cases = 7

No. of Favorable cases = 2 [there are two cases which have Friday]

Required Probability = 2/7

(2) What is the probability of getting a total of more than 10 in a single throw with two dice?

Solution: Where two dice are thrown then the sample space consists of 36 sample points
$\therefore$ Total no. of cases are = $6^2 = 36$
The favorable cases of getting more than 10 are (5,6), (6,5), (6,6) is : 3
Required Probability = 3/36 = 1/12

(3) A man and his wife appear for an interview for two posts. The probability of the husband selection is 1/7 and that of the wife‟s selection is 1/5 what is the probability that one of them will be selected?

Solution:- The Probability of husband selection is 1/7
The probability that husband not selected = 1 – 1/7 = 6/7
The probability that wife‟s selection is 1/5
The probability that wife is not selected = 1 -1/5 = 4/5
The probability that only husband is selected = 1/7 × 4/5 = 4/35
The Probability that only wife is selected = 1/5 × 6/7 = 6/35
The probability that only one of them is selected = 4/35 + 6/35 = 10/35 = 2/7.

(4) Start a committee of 4 persons is to be appointed from 3 officers of the production department, 4 officers of the purchase department 2officer of the sales department and 1 charted accountant. Find the probability of forming the committee as:
(i) There must be one from each department
(ii) It should have at least one from the purchase department
(iii) The charted accountant must be in the committee

Solution:- Total no. of persons = 4 + 3+2+1 = 10
Total no. of persons should be in the committee = 4
Out of 10 persons, 4 persons can be selected in to ${}^{10}C_4$ ways
$\therefore$ Total no. of cases = ${}^{10}C_4$ = 10 !/4! 6! = 210
(i) The no. of favorable cases for the committee to consists of one member from each department is;

$${}^{3}C_{1} \times {}^{4}C_{1} \times {}^{2}C_{1} \times {}^{1}C_{1} = 3 \times 4 \times 2 \times 1 = 24$$

$\therefore$ Required Probability = 24/210 = 4/35 = 0.1143

(ii) The Probability that the committee of 4 has at least one member from the purchase department = P[1 from purchase department and 3others] +P[2 from purchases department and 2 other] +P[3 from purchases department and 1 other] + P[4from purchase department only]

$$= \frac{{}^{4}C_{1} \times {}^{6}C_{3}}{{}^{10}C_4} + \frac{{}^{4}C_{2} \times {}^{6}C_{2}}{{}^{10}C_4} + \frac{{}^{4}C_{3} \times {}^{6}C_{1}}{{}^{10}C_4} + \frac{{}^{4}C_{4}}{{}^{10}C_4}$$

= 80 + 90 + 24 +1/210 = 195/210 = 0.9286

(iii) The Probability that charted accountant must be in the committee = P[Charted accountant and 3 others]

$$\frac{{}^{4}C_{1} \times {}^{9}C_{3}}{{}^{10}C_4} = 4/10 = 0.4$$

(5) A problem in Economics is given to three students A,B and C, whose chance of solving it are 1/3, ¼ and 1/5 respectively. Find the probability that the problem will be solved if they all try independently.

Solution:- Let A_1,A_2, and A_3 denote the events that the problem is solved by A, B and C respectively. We are given that

$P(A_1) = 1/3$ $P(A_2) = ¼$ and $P(A_3) = 1/5$

$P(\bar{A}_1) = 1- P(A_1) = 1-1/3 = 2/3$

$P(\overline{A}_2) = 1- P(A_2) = 1-1/4 = ¾$

$P(\overline{A}_3) = 1- P(A_3) = 1 -1/5 = 4/5$

The problem will be solved, if at least one of the three is able to solve it. The required probability that the problem will be solved is given by

$P(A_1 \cup A_2 \cup A_3) = 1 - P(\overline{A}_1 \cap \overline{A}_2 \cap \overline{A}_3)$

$= 1- P(\overline{A}_1).P(\overline{A}_2).P(\overline{A}_3)$

[$\therefore$ if A_1, A_2 and A_3 are independent then $\overline{A}_1$, $\overline{A}_2$ and $\overline{A}_3$ are also independent]

$= 1- 2/3 \times 3/4 \times 4/5$

$= 1- 2/5$

$= 3/5$

(6) The odds that A speaks the truth are 3:2 and the odds that B speaks the truth are 5:3. in what percentage of cases are they likely to contradict each other or an identical point?

Solution: Let the events A_1 and A_2 are defined as

A_1 = A speaks the truth, then

$\overline{A}_1$ = A tell a lie

A_2 = B speaks the truth, then

$\overline{A}_2$ = B tell a lie

We are given that

$P(A_1) = 3/3+2 = 3/5$ $P(\overline{A}_1) = 1- P(A_1) = 1-3/5 = 2/5$

$P(A_2) = 5/5+3 =5/8,$ $P(\overline{A}_2) = 1- 5/8 = 3/8$

The probability that A and B contradict each other on an identical point can happen in the following disjoint cases.

(i) A Speaks the truth and B tells lie

i.e., $A_1 \cap \overline{A}_2$ happens

(ii) A tells lie and B speaks the truth

i.e., $\overline{A}_1 \cap A_2$ happens

Hence by the addition theorem of probability, the required probably = P(i) + P(ii)

$= P(A_1 \cap \overline{A}_2) + P(\overline{A}_1 \cap A_2)$

$= P(A_1). P(\overline{A}_2) + P(\overline{A}_1).P(A_2)$

$= 3/5. 3/8 + 2/5 .5/8$

$= 9/40 + 10/40 = 19/40 = 0.475 = 47.5\%$

So A and B contradict each other on an identical point in 47.5% of the cases.

(7) A bag contains 8 Red and 5 White balls. Two successive drawings of 3 balls are made such that (i) balls are replaced before the second draw.(ii) Balls are not replaced before the second draw. Find the probability that the first draw will give 3 white balls and the second draw will give 3 red balls.

Solution: - Let A_1, A_2 are the events defined as

A_1: Drawing 3 white balls in the first draw

A_2: Drawing 3 red balls in the second draw

(i) Draws with replacement:- If the balls drawn in the first draw are replaced back in the bag before the second draw, then the events A_1 and A_2 are independent.

Then the required Probability is $P(A_1 \cap A_2) = P(A_1). P(A_2)$

1st Draw: Total no. of balls = 8 + 5 =13 out of 13 balls 3 balls can be drawn into $^{13}C_3$ ways

Total no. of white balls = 5

Out of 5 white balls 3 white balls can be drawn into ${}^5C_3$ Ways

$\therefore$ $P(A_1) = {}^5C_3 / {}^{13}C_3$

II nd draw: when the balls drawn in the first draw are replaced before second draw, the bag again consists of 13 balls. The probability of drawing 3 red balls in the second draw is

$P(A_2) = {}^8C_3 / {}^{13}C_3$

$P(A_1 \cap A_2) = {}^5C_3 / {}^{13}C_3 \times {}^8C_3 / {}^{13}C_3$

(ii) Draws without replacement: if the balls drawn are not replaced before the second draw, then the event A_1 and A_2 are not independent. In this case

$P(A_1 \cap A_2) = P(A_1).\ P(A_2 / A_1)$

$P(A_1) = {}^5C_3 / {}^{13}C_3$

If the first drawn balls are not replaced back before the second draw, then the total no. of balls left in the bag are 13- 3 =10. out of 10 balls 3 balls can be drawn in to ${}^{10}C_3$ ways

$P(A_2 / A_1) = {}^8C_3 / {}^{10}C_3$

$P(A_1 \cap A_2) = P(A_1).\ P(A_2 / A_1) = {}^5C_3 / {}^{13}C_3 \times {}^8C_3 / {}^{10}C_3$

(8) The odds against a student X solving a business economics problems are 8 to 6 and odds in favor of the student Y solving the problem are 14 to 16.

(a) What is the probability that the problem will be solved if they both try independently of each other?

(b) What is the Probability that none of them is able to solve the problem?

Solution:- Let A and B are two events defined as

A: The student X solves the Problem

B: The student Y solves the Problem

$\therefore$ P(A) = 6/8+6 = 6/14

P (B) = 14/ 14 +16 = 14/30

(a) The Probability that the problem will be solved = P [At least one of them solves the problem] $P(A \cup B)$

$= P(A) + P(B) - P(A \cap B)$

$= P(A) + P(B) - P(A \cap B)$

$= P(A) + P(B) - P(A).\ P(B)$

$= 6/14 + 14/30 - 6/14 \times 14/30 = 73/105$

(b) The Probability that none of them will solve the problem is

$P(\overline{A} \cap \overline{B}) = P(\overline{A}).\ P(\overline{B})$

$= 8/14.\ 16/30$

$= 32/105$

11.6 Summary

Probability in common paralance means the chance of occurrence of an event. Conditional Probability can be applied if an event has already happened. In different types of situations, we can apply the classical, statistical and aromatic approaches. Based on the nature of the problems, we will apply addition theorems of Probability as well as multiplication theorems of probability certain results in probability theory which are helpful in choosing the valid decisions or alternatives among the various alternating.

11.7 Self Assessment Questions

(1) Explain various definitions of Probability

(2) State Addition and multiplication theorems of probability.

(3) Define Conditional Probability
(4) There are four hotels in a city. If 3 men check into hotels in a day, what is the probability they each are into a different hotel? [Answer: - 24/64 = 0.375]
(5) What is the probability that a leap year selected at random will contain 53 Thursday or 53 Fridays? [Answer: 3/7]
(6) A problem in statistics is given to two students A and B. The odds in favour of A solving the problems are 6 to 9 and against B solving the problem are 12 to 10. if both A and B attempt, find the Probability of the Problem being solved. [Answer: - 37/55]
(7) A card is drawn from well shuffled pack playing cards. Find the Probability that it is either a diamond or an ace. [Answer: 4/13]
(8) A bag contains 5 white and 3 black balls; another bag contains 4 white and 5 black balls. From any one of these bags a single draw of two balls is made. Find the probability that one of them would be white and the other black ball? [Answer: 0.5456]
(9) The probability that a contractor will get a plumbing contract is 2/3, and the probability that he will not get an electric contact is 5/9. If the probability of getting at least one contract is 4/5. What is the probability that he will get both the contracts?
[Answer: - 14/45]

Reference Books

1. S.C Gupta: Fundamentals of Statistics
2. K. Chandra Sekhar: Business Statistics
3. K.V.Sarma: Statistics made simple, Prentice Hall of India

Lesson Writer
Prof. K. Chandan

12. INVERSE PROBABILITY

Objectives

After completion of this chapter, you should be able to

- Known the concept of Inverse Probability
- Understand the Baye's theorem
- Explain how to apply Baye's theorem in numerical problems

Structure

12. INVERSE PROBABILITY

12.1 Introduction:

The most important application of probability is in the computation of unknown probabilities on the basis of the information given by the experiment or past records. There the conditional probability that an event has occurred through one of the various mutually disjoint events is called posterior Probabilities or inverse probability. The Probability are calculated based on Baye's Rule are also known as inverse probabilities. Thus, the main purpose of use Baye's theorem is to compute posterior Probabilities. These probabilities are based on the relation effect to causes but not cause to effect.

12.2 Bayer's theorem

Baye's theorem is based on the concept that the probabilities should be revised when some additional or new information is available. Baye's theorem offers as powerful statistical method of evaluating new information and revising our prior estimates of probability modern decision theorem is often is called Bayesian Design theory.

Probabilities before revision, by Baye's rule are called Prior or Priory probabilities. The posterior Probabilities is the revision of probability with new information. Posterior probabilities are also called revised probabilities.

12.3 Statement of Bayer's theorem

If an event B can only occur in the conjunction with one of the "n" mutually exclusive and exhaustive events $A_1, A_2, \ldots A_n$ and if "B" actually happens, them the probability that it was Preceded by the particular event A_i(i =1,2,….n) is given by

$$P(Ai/B) = \frac{P(Ai).P(B/Ai)}{\sum_{i=1}^{n} P(Ai).P(B/Ai)}$$

Here

$P(A_1)$, $P(A_2)$, …..$P(A_n)$ are called Prior probabilities

$P(B/A_1)$, $P(B/A_2)$,….$P(B/A_n)$ are called Conditional probabilities

$P(A_1/B)$, $P(A_2/B)$. …. $P(A_n/B)$ are called posterior (or) inverse probabilities

12.4 Solved Problems

(1) In a bolt factory, Machines A, B and C manufacture respectively 25%, 35% and 40% of the total. Of their output 5%, 4% and 2% are known to be defective bolts. A bolt is drawn at random from the product and is found to be defective. What are the probabilities it was manufactured by machines A, B or C?

Solution:- Let A_1, A_2 and A_3 are three events defined as the bolts are manufactured by the machines A, B and C respectively.

$P(A_1) = 25\% = 0.25$

$P(A_2) = 35\% = 0.35$

$P(A_3) = 40\% = 0.40$

Let "B" be the event defined as the bolt is defective. From the problem,

$P(B/A_1) = 5\% = 0.05$

$P(B/A_2) = 4\% = 0.04$

$P(B/A_3) = 2\% = 0.02$

According to Beye's theorem the probability that the defective bolt manufactured by machine A is

$P(A_1/B) = P(A_1).\ P(B/A_1) / P(A_1).\ P(B/A_1) + P(A_2).\ P(B/A_2) + P(A_3).\ P(B/A_3)$

$= 0.25 \times 0.05 / (0.25 \times 0.05) + (0.35 \times 0.04) + (0.40 \times 0.02)$

= 0.0125 / 0.0125 +0.0140 + 0.0080

= 0.0125 / 0.0345 = 0.36

The probability that the defective bolt manufactured by machine "B" is

$P(A_2/B) = P(A_2).P(B/A_2) / P(A_1).P(B/A_1) + P(A_2).P(B/A_2) + P(A_3).P(B/A_3)$

= 0.35 × 0.04 /(0.25 ×0.05) + (0.35 × 0.04) + (0.40 ×0.02)

= 0.0140/ 0.0125 +0.0140 + 0.0080

= 0.0140/0.0345 = 0.41

The probability that the defective bolt manufactured by machine "C" is

$P(A_3/B) = P(A_3).P(B/A_3)/ P(A_1).P(B/A_1) + P(A_2).P(B/A_2) + P(A_3).P(B/A_3)$

= 0.40 × 0.02 /(0.25 ×0.05) + (0.35 × 0.04) + (0.40 ×0.02)

= 0.0080/0.0125 +0.0140 + 0.0080

= 0.0080/0.0345 = 0.23

(2) In a class of 75 students, 15 were considered to be very intelligent, 45 as medium and the rest below average. The probability that a very intelligent student fails in a viva – voce examination is 0.005; the medium student failing has a probability 0.05; and the corresponding probability for a below average student is 0.15. if a student is known to have passed the viva – voce examination what is the probability that he is below average?

Solution: The events A_1, A_2 and A_3 are defined as

A_1:- The student is very intelligent

A_2:- The Student is medium

A_3:- The student is below average

Also let the event "B" as

B: The student passes in the viva – voce examination

From the given problem

$P(A_1) = 15/75 = 0.2$

$P(A_2) = 45/75 = 0.6$

$P(A_3) = 15/75 = 0.2$

Also $P(B/A_1) = 1- 0.005 = 0.995$

$P(B/A_2) = 1 - 0.5 = 0.95$

$P(B/A_3) = 1- 0.15 = 0.85$

According to Baye's theorem if a student is known to have passed the viva – voce examination that he is below average student is given by

$P(A_3/B) = P(A_3).P(B/A_3)/ P(A_1).P(B/A_1) + P(A_2).P(B/A_2) + P(A_3).P(B/A_3)$

= 0.2 ×0.85 / (0.2 ×0.995) + (0.6 × 0.95) + (0.2 × 0.85)

= 0.170 / 0.199 + 0.570 +0.170

= 0.170 / 0.939

= 0.181

(3) The contents of Boxes I, II and III are as follows:

Box I: 1 white, 2 Black and 3 Red balls

Box II: 2 White, 1 Black and 1 Red balls

Box III: 4 white, 5 Black and 3 Red balls

One Box is chosen at random and two balls drawn. They happen to be white and red. What is the probability that they came from urns I, II and III?

Solution:- Let A_1, A_2 and A_3 denote the events of choosing First, Second and Third boxes respectively. The probability of selecting any box is 1/3

$\therefore P(A_1) = 1/3, P(A_2) = 1/3, P(A_3) = 1/3$

Let g be the event define as that the two balls drawn from the selected box are White and Red.

$P(B/A_1) = 1 \times 3 / {}^6C_2 = 1/5$

$P(B/A_2) = 2 \times 1/ {}^4C_2 = 1/3$

$P(B/A_3) = 4 \times 3 / {}^{12}C_2 = 2/11$

According to the Baye's rule, the probability that the drawn two red and white balls are came from Box I is

$P(A_1/B) = P(A_1).\ P(B/A_1) / P(A_1).\ P(B/A_1) + P(A_2).\ P(B/A_2) + P(A_3).\ P(B/A_3)$

$= 1/3 \times 1/5 /(1/3 \times 1/5) + (1/3 \times 1/3) + (1/3 \times 2/11)$

$= 1/15 / 118/495 = 33/118$

Similarly

$P(A_2/B) = P(A_2).\ P(B/A_2) / P(A_1).\ P(B/A_1) + P(A_2).\ P(B/A_2) + P(A_3).\ P(B/A_3)$

$= 1/3 \times 1/3 /(1/3 \times 1/3) + (1/3 \times 1/3) + (1/3 \times 2/11)$

$= 1/9 / 118/495 = 55/118$

And $P(A_3/B) = 1- P(A_1/B) - P(A_2/B)$

$=1- 33/118 - 55/118$

$= 30/118$

12.5 Summary

The basic idea developed by Bayer's is to compute the posterior probabilities. In conditional probability we studied only those problems on probability in which our knowledge of factors affecting the event was sufficient to enable as to determine the chances of happening of the event. Posterior probability are always conditional probabilities, the conditional event being the sample information. Every time some new information is available, we do revise our odds mentally. This revision of probability with added information is formalized in probability theory in terms of a theorem known as Bayer's theorem. Baye's theorem is also known as inverse Probability theorem. the idea of revising the probability is used by all of us in daily life even though we may not be knowing anything about probability.

12.6 Self assessment questions

(1) Define inverse Probability. State the Bayer's Theorem

(2) A factory has two machines A and B. past records shows that machine A produces 30%. Of the total output and machine B the remaining 70%. Machine A produces 5% defective articles and machine B produces 1% defective items. An item is drawn at random and found to be defective. What is the probability that it was produced by (i)Machine A (ii) Machine B

[Answer:- the probability that the defective article came from machine A is 0.682 and machine B is 0.318]

(3) Two sets of candidates are competing for the position on the board of directors of a company. The probability that the first and second sets will win are 0.6 and 0.4 respectively. if the first set wins, the probability of introducing a new product is 0.8 and corresponding probability if the second set wins is 0.3. What is the probability that the product will be introduced by the second? [Answer: 0.2]

(4) There are 4 boys and 2 girls in room A and 5 boys and 3 girls in room B. A girl from one of the two rooms laughed loudly what is the probability that the girl who laughed was from room B. [Answer 9/17]

(5) In a Railway reservation office, two clerks are engaged in checking reservation forms. On an average, the first clerk checks 55% of the forms, while the second does the remaining. The first clerk has an error rate of 0.03 and second has an error rate of 0.02. A reservation form is selected at random from the total number of forms checked during at day, and is found to have an error. Find the probability that it was checked by first clerk.
[Answer: 11/17]

12.7 References

(1) P.N Arora; S. Arora: Statistics for management: S. Chand and company limited
(2) S.C Gupta: Fundamentals of Statistics: Himalaya Publishing House
(3) K. Chandrasekhar: Business Statistics

Lesson Writer
Prof. K Chandan

13. JOINT AND MARGINAL PROBABILITES

Objectives

After completion of the chapter you should be able to

- Understand random variables and types
- Explain the probability most and density functions
- Know about the Joint and Marginal probability

Structure

13.1 Introduction

Frequency distribution with single variable is called unavailable frequency distribution. Probability distribution with single variable is called probability distribution. Probability distributions are key distributions to compute various statistics like mean, ,median, mode etc. the joint and marginal probability distribution are same as bivariate frequency distribution as and marginal frequency distribution except we consider joint probabilities in the place of joint frequencies. In joint and marginal probabilities, the probabilities are associated with random variables. The random variables may be discrete or continuous.

13.2 Random variable and types

Random variable, we mean a real number X associated with the outcomes of at random experiment. Random variable can take any one of the various possible values each with as definite probability. For example, in toss of a coin if X is denotes number of tails, then X is a random variable which can take any one of the two values: 0 (no tail) of 1 (tail), each which equal probability ½. Similarly, in throw of a die, if Y denotes the number on the die, then Y is a random variables which can take any one of the values 1,2,3,4,5 or 6 each with equal probability 1/6.

Random variable may also defined as a real function valued function on the sample space taking values on the real line $R(-\infty, \infty)$." It is also defined as "Random variable is a real valued function which takes real values, which are obtained by the outcomes of a random experiment." Generally random variables are denoted by the capital letter of English alphabet as X, Y,Z . . . etc and values which are taken by the random variable are denoted by corresponding small letters of English alphabet.

Random variables are classified as

(i) Discrete Random variable

(ii) Continuous Random Variable

If the random variable X assumes only a finite or countable infinite set of values then it is known as discrete Random variable Ex: Marks obtained by a student in an examination, no. of students in a college, no. of accidents, no. of defective items in a box etc are all discrete random variables.

If the random variable X assumes infinite and uncountable set of values (or) if a random variable X assumes all possible values in a certain limits or range then it is known as continuous random variable. Examples of continuous random variables are age, height or weight of students in a class etc.

13.3 Probability mass function and probability density function

If X be a discrete random variable, which takes the values are $x_1, x_2, \ldots x_i, \ldots x_n$ and corresponding probabilities are $P(x_1), P(x_2), \ldots P(x_i) \ldots\ldots\ldots P(x_n)$ respectively then the probability $p(x_i)$ or P_i or $P(x = x_i)$ associated to "xi) is called probability mass function if it satisfies the following conditions.

(i) All probability are non – negative is. positive. i.e.,. $P(x_i) \geq 0, \forall x_i$

(ii) Total probability is one is $P(x_1) + P(x_2) + \ldots\ldots + P(x_n) = 1$

$$\sum_{i=1}^{n} P(x_i) = 1$$

Let X be the continuous random variable taking values on the interval [a, b]. The probability function f(x) is said to be probability density function of the continuous random variables „X" if it statistics the following properties:

(i) The probability density function is always positive i.e.,: $f(x) \geq 0$, $\forall x \in [a, b]$

(ii) Total area under the probability curve is unity. Is $\int_a^b f(x)\ d_x = 1$

For a continuous random variable, the probability at a point is always zero. Is if „C" is point then P(x = c) = 0

13.4 Joint Probability distributions

Let X and Y are two discrete random variables. Let us suppose that the random variable X assumes „m" values $x_1, x_2, \ldots x_i, \ldots x_m$ and the random variable Y assumes „n" values $y_1, y_2, \ldots y_j, \ldots y_n$ respectively. by going through the pairs of values (x_i, y_i), i = 1,2,. . .m j=1,2.....n we can obtain the joint probabilities $P(x_i, y_j)$ or $P(x = x_i, y = y_j)$ or f_{ij} as shown in the following table, then the function $p(x_i, y_i)$ or $P(x = x_i, y = y_i)$ is called joint probability mass function if it satisfies the following properties.

(i) The joint probabilities are always positive is $P(x_i, y_j) \geq 0$, i= 1,2 . . .m, j = 1, 2 . . n

(ii) Total Probability is always are

$$\sum_{i=1}^{m} \sum_{j=1}^{n} P(x_i y_j) = 1$$

Y → / X ↓	Y_1	Y_2	- - - - -- - - Y_j - - - - - - - -- -	Y_n	Total
X_1	$P(x_1,y_1)$	$P(x_1,y_2)$	- - --- - $P(x_1,y_j)$ ------	$P(x_i,y_n)$	$P(x_1)$
X_2	$P(x_2,y_1)$	$P(x_2,y_2)$	- - --- - $P(x_2,y_j)$ ------	$P(x_2,y_n)$	$P(x_2)$
„	„	„	„	„	„
„	„	„	„	„	„
X_i	$P(x_i,y_1)$	$P(x_i,y_2)$	- - --- - $P(x_i,y_j)$ ------	$P(x_i,y_n)$	$P(x_i)$
„	„	„	„	„	„
„	„	„	„	„	„
„	„	„	„	„	„
X_m	$P(x_n,y_1)$	$P(x_n,y_2)$	- - --- - $P(x_n,y_j)$ ------	$P(x_n\ y_m)$	$P(x_m)$
Total	$P(Y_1)$	$P(Y_2)$	- - - - -- - - -$P(Y_j)$ ----------	$P(Y_n)$	1

If X and Y are the continuous random variables then f(x, y) is said to be joint probability density function, if it satisfies the following properties.

(i) The joint probability density function is always positive. i.e., $f(x, y) \geq 0$

(ii) Total area under the probability curve is unity is

$\int\int f(x, y)\ d_x\ d_y = 1$

13.5 Marginal probability distributions

If X and Y are the discrete random variables, which takes the values are $x_1, x_2, - - - - x_i, - - x_m$ and $y_1, y_2, - - - y_j - - - y_n$ respectively and $P(x_i, y_j)$ or $P(x = x_i, y = y_j)$ is the joint probability mass function then $P_x(x_i)$ or $P(x_i)$ is called marginal probability mass function of x which is defined as:

$P(x_i) = P(x_i, y_j) + P(x_i, y_2) + ... + P(x_i, y_m)$ the marginal probability distribution of „x" is

$X = x_i$	x_1 x_2 - - - - x_i - - x_m	total
$P(x_i)$	$P(x_1)$ $P(x_2)$ - - - -- $P(x_i)$--------$P(x_m)$	1

The marginal probability mass function of Y is denoted by $P_y(y)$ or $P(y_j)$ and is defined as

$P(y_j) = P(x_1, y_j) + P(x_2, y_j) + \ldots + P(x_m, y_j)$ the marginal probability distribution of „Y" is

$Y = y_j$	y_1 y_2 - - - - - - - - - - y_j----------y_m	Total
$P(y_j)$	$P(y_1)$ $P(y_2)$ - - - - - - - - P(yj) ----------$P(y_n)$	1

13.6 Solved problems

(1) Let (x, y) be the pair of discrete random variables each taking three values 1, 2 and 3 which the joint probability distribution as

X \ Y	1	2	3
1	5/27	4/27	2/17
2	1/27	3/27	3/27
3	3/27	4/27	2/27

obtain the marginal probability distributions x and y.

Solution:- The given Joint probability distribution is

X \ Y	1	2	3	Total
1	5/27	4/27	2/17	11/27
2	1/27	3/27	3/27	7/27
3	3/27	4/27	2/27	9/27
Total	9/27	11/27	7/27	1

The marginal distribution of X is

x	P(x)
1	11/27
2	7/27
3	9/27
Total	1

The marginal distribution of Y is

Y	1	2	3	Total
P(y)	9/27	11/27	7/27	1

(2) Let X and Y be two random variables each taking three values -1, 0 and 1, and having joint probabilities distribution as

Y \ X	-1	0	1
-1	0.0	0.1	0.1
0	0.2	0.2	0.2
1	0.0	0.1	0.1

Obtain marginal probability distribution of X and Y

Solution:- The joint probability distribution of X and Y is

X ► / Y ▼	-1	0	1	Total
-1	0.0	0.1	0.1	0.2
0	0.2	0.2	0.2	0.6
1	0.0	0.1	0.1	0.2
Total	0.2	0.4	0.4	1

the marginal probability distribution of X is

X=x	-1	0	1	Total
P(x)	0.2	0.4	0.4	1

The marginal probability distribution of Y is

Y =y	-1	0	1	Total
P(y)	0.2	0.6	0.2	1

(3) The joint Probability distribution of X and Y is

Y / X	1	2	3
1	a	2a	3a
2	2a	2a	2a
3	4a	2a	3a

Obtain the marginal Probability distribution of X and Y

Solution:- The joint Probabilities are expressed in terms of the constant “a”. the constant can be estimated using total probability properly.

We know that total probability is one is a +2a +3a +2a+2a+2a+4a+2a+3a =1

$\Rightarrow$ 21 a = 1

$\therefore$ a = 1/21

$\therefore$ the Probability distribution of X and Yis

Y / X	1	2	3	Total
1	1/21	2/21	3/21	6/21
2	2/21	2/21	2/21	6/21
3	4/21	2/21	3/21	9/21
Total	7/21	6/21	8/21	1

The marginal probability distribution of X is

X=x	1	2	3	Total
P(x)	6/21	6/21	9/21	1

The marginal probability distribution of Y is

Y =y	1	2	3	Total
P(y)	7/21	6/21	8/21	1

13.7 Summary

Probability distributions with two variables is called joint probabilities and probability distribution with unique random variable is called marginal probability. The marginal probability can be easily computed when the joint probability are known.

13.8 Self assessment questions

1. Define Random variable with an example
2. Explain about Joint Probability distributions and marginal Probability distributions.
3. Define probability mass function and Probability density function.
4. The discrete random variables x and y each taking three values 4,5 and 6, and having joint Probability distribution as follows.

Y / X	4	5	6
4	0.1	0.0	0.1
5	0.2	0.1	0.2
6	0.2	0.1	0.0

Obtain marginal probability distribution of x and y.

5. The random variables x and y, each taking the values -1, 0 and 1, and having the following joint Probability distribution.

Y / X	-1	0	1
-1	c	2c	3c
0	2c	3c	4c
1	2c	2c	3c

(i) find „C"

(ii) Compute the marginal probability distribution of X and Y

13.9 Reference Books

1. S.C. Gupta: Fundamentals of Statistics
2. K. Chandra Sekhar: Business Statistics
3. K. V. Sarma: Statistics made simple, pent ice Hall of India

Lesson Writer
Prof. K. Chandan

14. Binomial distribution

Objectives

After completion of this chapter, you should be able to:

- Understand about the Binomial distribution;
- Explain the characteristics of Binomial distribution;
- Know the conditions to apply the Binomial distribution;
- Apply the binomial distribution to a variety of problems.

Structure

14. 1 Introduction

Binomial distribution was discovered by James Bernoulli in the year 1700 and was first published posthumously in 1713, eight years after his death. Let a random experiment be performed repeatedly, each repetition being called a trial and let the happening of an event in a trial be called a success and its non-happening of an event in a trial be called a failure. 'Bi' at the beginning of a word generally denotes the fact that the meaning involves 'two' and binomial is no exception. A random variable follows a binomial distribution when each trial has exactly two possible outcomes. For example, when sarah, a practiced archer, shoots an arrow at a target the either hits or misses each time. If X is ' the number of hits sarah scores in five shots', then the probabilities associated with 0, 1, 2, 3, 4 5 hits can be expected to follow a particular pattern, known as Binomial distribution.

14.2 probability function of Binomial distribution:

A discrete random variable X is said to follows Binomial distribution, if it assumes non-negative values and it has the probability mass function,

$P(x) = n_{cx}. P^x. q^{n-x}$, x = 0, 1, 2,….., n

Here n = no. of independent trials

P = Probability of success in each trial

q = Probability of failure in each trial

X = Number of success

And n, p are called the parameters of Binomial distribution.

14.3 Conditions for applications of Binomial distribution.

1. The variable should be discrete, i.e, the values of x could be 1, 2, 3, 4, 5 etc. and never 1.5, 2.1 & 3.4, etc.
2. A dichotomy exists. In other words, the happening of events must have two alternatives. It must be either a success of a failure.
3. The number of trails 'n' should be finite and small.
4. The trails or events must be independent, the happening of one event must not effect the happening of other events. In other words, statistical independence must exist.
5. The trails or events must be repeated under identical conditions.

14.4 Characteristics of Binomial Distribution:

1. It is a discrete distribution which gives the theoretical probabilities.
2. It depends on a parameters p or q, the probability of success or failure and n (the number of trails). The parameter n is always a positive integer.
3. The distribution will be symmetrical if p = q.

It is skew-symmetric or a symmetric if $p \neq q$ although with n tending to large it is approximately so.

4. The statistics of the binomial distribution are mean = np; variance = npq; and standard deviation = $\sqrt{npq}$
5. The mode of the Binomial distribution is equal to that value of x which has the largest frequency.
6. It can be represented graphically, taking the x – axis to represent the number of success and y-axis to represent the probabilities or frequencies. Its graph will be vertical lines, with spaces in between them. Drawing a smooth curve by free hand is inadmissible as the variable x is a discrete one.
7. The shape and location of Binomial distribution charges as p changes for a given n or n changes for a given P.
8. The Binomial coefficients are given by the Pascalls Triagle.

14.5 Constants of Binomial distribution:

1. The mean of the Binomial distribution is 'np'
2. The variance of the Binomial distribution is 'npq'

3. The expected or theoretical frequencies are obtained as E = N × P (x) here P (x) is the probability function of Binomial distribution.
4. The recurrence relation of probabilities of Binomial distribution is P (x + 1) =
5. The standard deviation of Binomial distribution is
6. The moment generating function of Binomial distribution is $m_x(t) = (q + pe^t)^n$
7. The characteristic function of Binomial distribution is $Q_x(t) = (q + pe^{it})^n$.

14.6 Solved problems:

1. The overall percentage of failures in a certain examination is 30. What is the probability that out of a group of 6 candidates at least 4 passed the examination?

Solution: Probability of a student failing in examination p = 0.30

Probability of a student passing the examination q = 1-p = 0.70

Probability that out of 6 candidates at least 4 passed the examination.

$= 6_{c4}(0.7)^4(0.3)^{6-4} + 6_{c5}(0.7)^5(0.3)^{6-5} + 6_{c6}(0.7)^6(0.3)^{6-6}$

$= (0.7)^4\left[\frac{6!}{4!\times 2!}(0.3)^2 + \frac{6!}{5!\times 11}(0.7)(0.3) + \frac{6!}{6!\times 0}(0.7)^2\right]$

$= (0.7)^4\left[\frac{6\times 5}{2\times 1}\times 0.09 + 6\times 0.21 + 0.49\right]$

= 0.2401 (1.35 + 1.26 + 0.49)

= 0.74431

2. If the probability that a man aged 60 will live to be 70 is 0.65, what is the probability that out of 10 men now 60, at least 7 will upto 70?

Solution: Probability of man living upto 70 = 0.65

$= \frac{65}{100} = \frac{13}{20} =$ p (say)

$\therefore$ probability of dying = $1-\frac{13}{20} = \frac{7}{20}$ = q (say)

Total number of men = 10 and at least 7 will live up to 70. We have the following four cases;

Case I: The probability of 8 living and 2 dying

$10_{c_8}P^8 q^2 = \frac{10\times 9}{2\times 1}\left(\frac{13}{20}\right)^8\left(\frac{7}{20}\right)^2 = p_1$

Case II: The probability of 7 living and 3 dying = $10_{c_7}P^7 q^3$

$= \frac{(10)!}{7!3!}\left[\frac{13}{20}\right]^7\left[\frac{7}{20}\right]^3 = \frac{10\times 9\times 8}{3\times 2\times 1}\times\left[\frac{13}{20}\right]^7\times\left[\frac{7}{20}\right]^3 = P_2$

Case III: The probability of 9 living and 1 dying:

$10_{c_9}\left(\frac{13}{20}\right)^9\left(\frac{7}{20}\right) = 10\left(\frac{13}{20}\right)^9\left(\frac{7}{20}\right) = P_3.$

Case IV: The probability of all the 10 living = $10_{c10}P^{10} = \left(\frac{13}{20}\right)^{10} p_4$

Required probability = $P_1 + P_2 + P_3 + P_4$

$= \frac{10\times 9}{2\times 1}\left(\frac{13}{20}\right)^8\left(\frac{7}{20}\right)^2 + \frac{10\times 9\times 8}{3\times 2\times 1}\times\left[\frac{13}{20}\right]^7\times\left(\frac{7}{20}\right)^3 + 10\left(\frac{13}{20}\right)^9\left(\frac{7}{20}\right) + \left[\frac{13}{20}\right]^{10}$

$= \left(\frac{13}{20}\right)^7 \left(\frac{120\times 343}{8000} + \frac{45\times 13\times 49}{8000} + \frac{10\times 69\times 7}{8000} + \frac{2197}{8000}\right)$

$= (0.65)^7 (5.145 + 3.583 + 1.479 + 0.275)$

$= (0.65)^7 (10.482) = 0.04902 \times 10.482 = 0.5139.$

Hence the required probability = 0.5139

3. The incidence of occupational disease in an industry is such that the workmen have a 25% chance of sufficient from it. What is the probability that out of six workmen 4 of more will contact the disease?

Solution: Let P denotes chance of suffering and q chance not suffering.

P = 25% = ¼ and q = ¾.

The Binomial distribution is: $P(x = r) = n_{c_r}\, p^r\, q^{n-r}$

The probability of 4 of more (i 6 4, 5 and 6)

Success = P (4) + P (5) + P (6)

$15q^2 p^4 + 6qp^5 + P^6 = 15\left|\frac{1}{4}\right|^4 \left[\frac{3}{4}\right]^2 + 6\left|\frac{1}{4}\right|^5 \left|\frac{3}{4}\right| + \left|\frac{1}{4}\right|^6$

$= \left(\frac{15\times 9}{4096} + \frac{6\times 3}{4096} + \frac{1}{4096} = \frac{135+18+1}{4096} = \frac{154}{4096}\right) = 0.0376$

4. What is the probability of guessing correctly at least six of ten answers in TRUE/FALSE objective test?

Solution: Probability 'P' of guessing an answer correctly is p = ½

$\Rightarrow$

q = 1-p = 1/2

Probability of guessing correctly x answers in 10 questions $p(x) = n_{cx}\, p^x\, q^{n-x} = 10_{cx}(1/2)^{10-x}$, x = 0, 1, 2, …..10

Required probability = p (0) + p (7) + p (8) + P (9) + p (10)

$= \left(\frac{1}{2}\right)^{10} \left[10_{c6} + 10_{c7} + 10_{c8} + 10_{c9} + 10_{c10}\right]$

$= \frac{1}{1024}\left(\frac{10\times 9\times 8\times 7}{46} + \frac{10\times 9\times 8}{36} + \frac{10\times 9}{2!} + 10 + 1\right)$

$= \frac{1}{1024}(210 + 120 + 45 + 10 + 1)$

$\frac{386}{1024} = \frac{193}{512}$

5. Out of 320 families with 5 children each, what percentage would be expected to have

(i) 2 boys and 3 girls. (ii) at least one boy?

Assume equal probability for boys and girls.

Solution: Here n = 5, N = 320.

Probability of a boy = p =1/2;

Probability of a girl = q = ½

Binomial Distribution is: p (r) = ${}^{n}c_{r}\ p^{r}\ q^{n-r}$

(i) Probability of 2 Boys and 3 girls:

P (r = 2) = ${}^{5}c_{2}\left(\frac{1}{2}\right)^{3}\left(\frac{1}{2}\right)^{2} = \left(\frac{10}{32}\right)$

Expected number of families with 2 boys and 3 girls i N. P (r) = $320 \times \left(\frac{10}{32}\right) = 100.$

Percentage of families excepted to have 2 boys and 3 girls = $\left(\frac{10}{32}\right) \times 100 = 31.25\%$

(ii) Probability of at least one boy = p ($r \geq 1$) = 1- probability of no boy

P (r = 0) ${}^{5}c_{0}\left(\frac{1}{2}\right)^{5}$ $(1/2)^0 = 1/32,$ $P(r \geq 1) = 1 - P(r=0) = 1 - (1/32) = 31/32$

Percentage of families expected to have at least one boy = $\left(\frac{31}{32}\right) \times 100 = 96.875\%$

expected number of families with at least one boy:

$= 320 \times \left(\frac{31}{32}\right) = 310.$

(6) Ten unbiased coins are tossed simultaneously. Find the probability of obtaining.

(i) Exactly 6 heads (ii) At least 8 heads.

(iii) No head (iv) At least one head.

(v) not more than three heads

(vi) At least 4 heads.

Solution:If p denotes the probability of head, the p = q = ½. Here n = 10. If the random variable x denotes the number of heads, then by the binomial probability law.

1. The overall percentage of failures in a certain examination is 30. What is the probability that out of a group of 6 candidates at least 4 passed the examination?

solution

p (r) = ncr pr q n-r =

Thus, the probability that at least three dice show 5 or 6 is p (3) + P (4) + P (5) + P (6). Hence, in 729 throws of 6 dice each, the required frequency of getting at least 3 successes is N × p (3) + P (4) + P (5) + P (6)

$= 729 \times 1/729\ (6C_3 \times 2^3 + 6C_4 \times 2^2 + 6C_5 \times 2 + 6C_6 \times 1)$

$= (8 \times 20 + 15 \times 4 + 6 \times 2 + 1) = 160 + 60 + 12 + 1 = 233.$

14.7 Summary: Binomial distribution applies to a situations in which each outcome is either a 'success' or a failure. In n independent trails each have probability P of being a success and x

denotes the number of success. The probabilities of success as well as failure are fixed in each trial. The mean of the binomial distribution is always larger than the variance. If the number of the binomial distribution is always larger than the variance. If the number of trials are very large and probability of success is very small in each trial then Binomial distribution follows the poisson distribution. Binomial distribution is also a limiting case of Normal distribution. The parameters involved in the Binomial distribution are n and P.

14.8 Self assessment questions:

(1) An account is to audit 24 accounts of a firm. Sixteen of these are highly valued customers. If the accountant selects u of the accounts at random, what is the probability that he chooses at least one highly valued account?

Ans: 80/81.

(2) The average percentage of failures in a certain examination is 40. what is the probability that out of a group of 6 candidates, at least 4 passed in the examination.

Ans: 0.54432

(3) The incidence of occupational diseases, in an industry in such that the work men have 20% chance of suffering from it. What is the probability that out of six workmen, 4 of more will contract the disease?

Ans: 53/3125 = 0.0169.

It is known from the past experience that 80% of the students in a school do their home work. Find the probability that during a random check of 10 students. (i) all have done this home work.

(i) at least one has not done the home work.

(ii) at least one has not done the home work.

Ans: (i) 0.1074 (ii) 0.6778 (iii) 0.8926.

(5) An accountant is to audit 24 accounts of a firm. Sixteen of these are of highly valued customers. If the accountant selects 4 of the accounts of random, what is the probability that he chooses at least one highly valued account?

Ans: 1-(1/3)4.

(6) Assuming that it is true that 2 in 10 industrial accidents are due to fatigue, fit the probability that

(i) exactly 2 of 8 industrial accidents will due to fatigue,

(ii) at least 2 of 8 industrial accidents will be due to fatigue.

Ans: (i) Required probability = p (2) = 8c2 (0.2)2 = 0.2936

(ii) Required probability = p ($x \geq 2$) = 1-(0.8)8 + 8c1 (0.2) (0.8)7

= 0.4967

(7) Explain the characteristics of Binomial distribution.

(8) Define Binomial distribution. State the conditions to apply binomial distribution.

14.9 Reference Books

1. S. C Gupta: Fundamentals of Statistics.
2. K. Chandra Sekhar: Business Statistics
3. K. V. Sarma: Statistics made simple, Prentice Hall of India

Lesson writer
Prof. K. Chandan

15: POISSON DISTRIBUTION

Objectives

After completion of this chapter, you should be able to:-

- Understand about the Poisson Distribution:
- Explain the characteristics of Poisson Distribution
- Know the constants of Poisson Distribution:
- Explain the practical situations to apply Poisson distribution:
- Apply the Poisson distribution to a variety of problems.

Structure

15: POISSON DISTRIBUTION

15.1 Introduction

Some times we come across a rare event which occurs once in many trials. In such a situation we can apply Poisson distribution instead of Binomial distribution. Poisson distribution was discovered by the French Mathematician and physicist Simeon Denis Poisson who published it in 1837. It is a discrete distribution and is very widely used. Poisson distribution is a limiting form of the Binomial distribution in which n, the number of trial, becomes very large and p, the probability of success of the event is very small such that np is a finite quantity.

Moreover, in Binomial distribution there is a sample of definite size and one can count the number of times a certain event is observed but in Poisson distribution one cannot count the number of times an event occurs or does not occur such as the accident in a company, or the number of deaths in a city in one year by a rare disease. In such a situation we use Poisson distribution. These cases we never apply the Binomial distribution.

15.2 Probability function of Poisson distribution

A discrete random variable X is said to follows Poisson distribution, if it assumes non – negative Values and it has the Probability function,

$P(x) = \frac{e^{-m}.m^x}{x!}$, x =0,1,2,3 -- - - -

Here λ is called Parameter of the Poisson distribution.

x! = x(x-1) (x -2) ----------5×4×3×2×1

Product of „x" terms.

The probability of o,1,2,3, ----- x successes are given by

e^{-m}, $\frac{e^{-m}.m^1}{1!}$, $\frac{e^{-m}.m^2}{2!}$, - - - $\frac{e^{-m}.m^x}{x!}$ respectively also e =2.7183

15.3 Importance and practical situations to apply Poisson distribution

The condition under which Poisson distribution is obtained as a limiting case of the Binomial distribution and also the conditions for the general model underlying Poisson distribution suggest that Poisson distribution can be used to explain the behavior of a discrete random variables where the probability of occurrence of the event is very small and the total number of possible cases is sufficiently large. Such Poisson distribution has found application in a variety of fields such as Insurance, Physics, Biology, Business, Economics and industry. Some practical Situations where Poisson distribution can be used are.

(i) The number of telephone calls arriving at a telephone switch board in unit time (say, per unit)

(ii) The number of customers arriving at the super market: say, per hour.

(iii) The number of defects per unit of manufactured product (This is done for the construction of control chart for number of defects (1) Industrial (Quality Control)

(iv) To count the number of radio -active disintegrations of radio – active element per unit of time (physics)

(v) To count the number of packing manufactured by a good concern.

(vii) The number of suicides repported in a Particular day of the number of causalities. Due to rare disease such as heart attack or cancer or snake bile in a year.

(viii) The number of accidents taking place per day on a busy road.

(ix) The number of typographical errors per page in a typed material of the number of printing mistakes per page in a book.

15.4 Characters tics of Poisson distribution

(1) Poisson distribution is a discrete distribution.
(2) It depends mainly on the value of the mean m.
(3) This distribution is positively skewed to the left. With the increase in the value of the mean my the distribution shifts to the right and the skewness diminishes.
(4) Its arithmetic mean in relative distribution is P and in absolute distribution is np.
(5) If n is large and P is small, this distribution gives a close approximation to Binomial distribution since the arithmetic mean of Poisson is same as that of Binomial, so the Poisson distribution can be used instead of Binomial distribution if n or p is not known.
(6) Poisson distribution has only one Parameter, viz: m, the arithmetic mean. Thus the entire distribution can be determined once the arithmetic mean is known.

15.5 Constants of Poisson distribution

(i) The mean of the Poisson distribution is "m"
(ii) The variance of the Poisson distribution is "m"
(iii) The standard deviation of Poisson distribution is " $\sqrt{m}$.
(iv) The moment generating function of Poisson distribution is M x(t) = $e^{\lambda (e^t - 1)}$
(v) The characteristic function of Poisson distribution is ϕx(t) = $e^{\lambda(e^{it} - 1)}$
(vi) The recurrence relation of the Poisson distribution is P(x+1) = [λ/x+1] P(x).

15.6 Binomial approximation to Poisson distribution

Poisson distribution can be derived from Binomial distribution under the following conditions.
(i) P, the probability of the occurrence of the events is very small.
(ii) n is very large, where n is number of trails, i.e $n \to \infty$
(iii) n p is finite quantity say np = m then m is called the parameter of the Poisson distribution. In Poisson distribution, the probability of r success is given by

P(r) or P(x =x) = $\frac{e^{-m}.m^x}{x!}$, x = 0,1,2. - -

15.7 Solved problems

(1) if 5% of the electric bulbs manufactured by a company are defective, use Poisson distribution to find the probability that in a sample of 100 bulbs.
(i) none is defective.
(ii) 5 bulbs will be defective. (Given e^{-5} = 0.007)
Solution:- Here we are given: n = 100
P = Probability of defective bulb b = 5% =0.05.

Since P is small n is large, we may approximate the given distribution by Poisson distribution. Hence, the parameter m of the Poisson distribution is:

M = n p = 100 × 0.05 = 5

Let the random variable X denote the number of defective bulbs in a sample of 100. Then

$$P(x = r) = \frac{e^{-m}.m^r}{r!} = \frac{e^{-5}.5^r}{r!}, \quad r = 0, 1, 2, - - - -$$

(i) The Probability that none of bulbs is defective is given by

$P(X = 0) = e^{-5} = 0.007.$

(ii) The Probability of 5 defective bulbs is given by

$P(x = 5) = \frac{e^{-5} \times 5^5}{5!} \times 0.007 \times 625/24 = 4.375/24 = 0.1823$

(2) It is known from past experience that in a certain plant there are on the average u industrial accidents per months. Find the probability that in a given year there will be less than is accident. Assume Poisson distribution ($e^{-4} = 0.0183$)

Solution:- In the usual notations we are given m =4, If the random variable X denotes the number of accidents in the plant per month then by poison distribution probability law,

$P(x = r) = \frac{e^{-m}.m^r}{r!} = \frac{e^{-5}.4^r}{r!}$,

The required probability that there will be less than 4 accidents is given by

$P(x < 4) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3)$

$= e^{-4}\left[1 + 4 + \frac{4^2}{2!} + \frac{4^3}{3!}\right] = e^{-4}$ [1+4+8+10.67]

$= e^{-4} \times 23.67 = 0.0183 \times 23.67$

$= 0.4332.$

(3) It is given that 3% of the electric bulbs manufactured by a company are defective. Using Poisson distribution, find the probability that a sample of 100 bulbs will contain no defective bulb. Given that $e^{-3} = 0.05$

Solution: Let P be the Probability of defective bulb. Then

P = 3/100, n =100 Also n p = 100 ×3/100 = 3

We know that h : $P(r) = \frac{e^{-\lambda}.\lambda^r}{r!}$

Now probability that a sample will contain no defective bulb is

$P(0) = \frac{e^{-3}(3)^0}{0!} = e^{-3} = 0.05.$

4) It is known from the past experience that in a certain plant there are on the average 4 industrial accident per month. Find the probability that in a given year, there will be less than 4 accidents. Assume poison distribution [Given $e^{-4} = 0.0183$]

Solution:- let the random variable x denote the number of accidents in the plant per month

$P(x = r) = \frac{e^{-m}.m^r}{r!}$, r = 0, 1, 2 - - -

Here m =4.

$P(x = r) = \frac{e^{-4}.4^r}{r!}$, r = 0, 1, 2 - - -

Again $P(x < 4) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3)$

$= e^{-4}\left[1 + 4 + \frac{4^2}{2!} + \frac{4^3}{3!}\right] = e^{-4}$ [1 +4+8+10.67]

$= e^{-4} \times 23.67 = 0.0183 \times 23.67$

$= 0.4332.$

(5) An office Switch board receives telephone calls at the rat of 3 calls per minute on average. What is the probability of receiving no calls in one minute internal? [e^{-4} = 0.04979]

Solution:- Let the random variable x denote the number of telephone calls per minute, then x follows Poisson distribution with parameter m =3 and probability density function.

P(x =r) = $\frac{m^r e^{-m}}{r!} = \frac{3^r e^{-3}}{r!}$, r=0, 1, 2 - - - -

$\therefore$ m =3

Probability of no calls in one minute = P(0), where

P(0) = $\frac{3^0 e^{-3}}{0!}$ = 0.04979.

(6) Assuming that the typing mistakes per page committed by a follows a Poisson distribution, find the expected frequencies for the following distribution of typing mistakes:

No. of mistakes per page :- 0 1 2 3 4 5

No. of Pages:- 40 30 20 15 10 5 [$e^{-1.5}$ = 0.22313]

Solution:- Here N = 120

Mean m = $\frac{40\times 0 + 30\times 1 + 20\times 2 + 3\times 15 + 4\times 10 + 5\times 5}{120} = \frac{180}{120}$=1.5

Frequencies are : P(0), P(1), P(2)--------P(5), where

P(0) = $e^{-1.5}$ = 0.22313

P(1) = $e^{-1.5}$ ×1.5 = 0.334695

P(2) = $e^{-1.5}$ × $\frac{(1.5)^2}{2!}$ =0.25

P(3) = $e^{-1.5}$ × $\frac{(1.5)^3}{3!}$ =0.13

P(4) = $e^{-1.5}$ × $\frac{(1.5)^4}{4!}$ =0.05

P(5) = $e^{-1.5}$ × $\frac{(1.5)^5}{5!}$ =0.01

The expected frequencies are given by

N× $\frac{e^{-1.5}(1.5)}{r!}$; r = 0, 1,2,3,4,5 - - - -

No. of mistakes	No. of Pages	Expected frequency $Ne^{-m}m^x/x!$
0	40	120 × 0.22313 = 27
1	30	120 ×0.334695 =40
2	20	120 ×0.25 = 30
3	15	120 ×0.13 = 16
4	10	120 × 0.05 = 6
5	5	120 ×0.01 = 1
	N = 120	120

15.8 Summary

The Poisson distribution refers to the counts of items that occur at random points in time or space. It is the one of the most important discrete distribution. The Poisson distribution which is presented in this chapter have wide use in engineering, scientific and management applications. In this distribution, mean and variance are equal. It has only one parameter. The Poisson distribution usually is used to analyze phenomena that produce rare occurrences. The only information required to generate a Poisson distribution pertains to occurrences over some interval. The assumptions are that each occurrence is independent of other occurrences and that the value of the parameter remains constant throughout the experiment.

15.9 Self Assessment Questions

(1) Between 2 and 4 p. m the average number of phone calls per minute coming into the switchboard of a company is 2.5. Find the probability that desire of one particular minute there will be (i) no phone call at all (ii) Exactly 3 calls
(Given $e^{-2} = 0.13534$ and $e^{-0.5} = 0.60650$)
Answer:-

$$P(0) = \frac{e^{-2.5}(2.5)^0}{0!} = e^{-2.5} \times e^{-0.5} = 0.13534 \times 0.60650 = 0.0821$$

$$P(3) = \frac{e^{-2.5}(2.5)^3}{3!} = 0.214$$

(2) One fifth percent of the blades produced by a blade manufacturing factory turn out to be defective. The blades an supplied in packets of 10. Use Poisson distribution to calculate the approximate number of packets containing no defective, one defective and two defective blades repetitively in a consignment of 1,00,000 packets. Gives $e^{-0.2} = 0.9802$.

$M = 10 \times 1/500 = 0.02$

$$P(x) = \frac{e^{-0.02}(0.02)^x}{x!},$$

N P(1) = 1960, N P(2) = 20.

(3) The distribution of number of road accidents per day in a city is Poisson with mean 4. find the number of days out of 100 days when there will be (i) no accident (ii) at least 2 accident and (iii) at most 3 accidents
Answer:- (i) No accident = NP(0) = 1.83 ≅ 2 days.
(ii) At least 2 accident = 100 – NP(0) + NP(1)] = 100 –(2+7) = 91 days
(iii) al least 3 accidents = NP(0) +NP(1) +NP(2) + NP(3) = 2 +7+15 +20 =44 days.
4) The Probability that an individual suffers from a bad reaction from an infection of a gives serum is 0.001. Determine the probability that out 2000 individuals.(i) exactly 3 and (ii) more then 2 individual who suffer from bad section.
Answer:- (i) $4e^{-2}/3$ (ii) $1\text{-}6e^{-2}$
(5) The following mistakes per page were observed in a book;

No. of mistakes per page	0	1	2	3	4
No. of times the mistakes accrued	211	90	19	5	0

Fit a Poisson distribution to the data.

Answer: $\frac{\sum fx}{\sum x} = \frac{211\times 0 + 1\times 90 + 2\times 19 + 3\times 5 + 4\times 0}{143} = 0.44$

Expected frequencies are N P(r) =N × $\frac{e^{-m}.m^r}{r!}$, r = 0, 1, 2,3,4 - - -

A)209.30; 92.09; 20.26; 2.97, 0.34
(6) Define Poisson distribution state the constants of Poisson distribution.
(7) Explain about the characteristics of Poisson distribution.
(8) Write the applications of Poisson distribution.

15.10 References

1. S. C. Gupta: Fundamentals of Statistics
2. K. Chandra Sekhar : Business Statistics.
3. K. V. Sarma: Statistics made simple, Prentice Hall of India.

Lesson Writer
Dr. J. PRATAPA REDDY

LESSON 16: NORMAL DISTRIBUTION

Objectives

After completion of this chapter, you should be able to:

- Understand about the Normal distribution;
- Explain the properties of Normal distribution;
- Know the uses of Normal distribution;
- Apply the Normal distribution to a variety of problems.

Structure.

16.1 Introduction:

Normal distribution is a continuous probability distribution in which the relative frequencies of a continuous variable are distributed according to the normal probability law. It is a symmetrical distribution. Normal distribution was first discovered by British mathematician De-Movire in 1733. Normal distribution is also known as Gaussian distribution. Now a day's normal probability model is one of the most important probability model in statistical analysis.

The normal distribution of a variable, when represented graphically, takes the shape of a symmetrical curve, known as the normal curve. This curve is asymptotic to base line on its either side. It is also known as Normal curve of error. The shape of the curve is Bell- shape curve.

16.2 Probability functions of normal distribution:

A random variable "X" is said to follows Normal distribution with mean "μ" and variable "σ^2" if it has the probability function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} . e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$

Here

μ and σ^2 are called the parameters of normal distribution

σ Is called standard deviation

e= 2.7183.

$\sqrt{2\pi}$ = 2.5060.

"X" is a continuous random variable and it assumes the values between -∞ and +∞.

16.3 Standard normal distribution

A continuous random variable Z is said to follows normal distribution with mean "0" and variance "1" if it has the probability functions

$$F(z) = \frac{1}{\sqrt{2\pi}} . e^{-\frac{z^2}{2}} \;,\; -\infty < Z < \infty \text{ here } Z = \frac{X - \mu}{\sigma}$$ Is called standard normal variable follows normal distribution with mean "0" and variance "1`" respectively. σ Is called the standard deviation. The standard normal distribution is called as Unit normal distribution (or) Z- distribution

The area under any normal curve is found from the table of a standard normal probability distribution showing the area between mean and any value of the normally distributed random variable. The standard normal curve helps as to find the areas within two assigned limits under the curve.

16.4 Properties of normal distribution:

1. The equation of the curve is and it is bell shaped. The top of bell is directly above mean is.

2. The curve is symmetrical about the line $x=\mu$ and x ranges from $-\infty$ to $+\infty$

3. x- axis is asymptote to the curve.

4. The points of inflection of the curve areal at $x= \mu+\sigma$, $x= \mu-\sigma$ and the curve changes from concave to convex at $x= \mu+\sigma$ to $x= \mu- \sigma$.

5. The total area under the normal curve is equal to unity and the percentage distribution of area under

The normal curve is given below and is shown also figure.

(i) About 68% of the area falls between $\mu-\sigma$ and $\mu+\sigma$

(ii) About 95.5% of the area falls between $\mu-2\sigma$ and $\mu+2\sigma$.

(iii) About 99.7% of the area falls between $\mu-3\sigma$ and $\mu+3\sigma$

6. The maximum ordinate lies at the mean, i.e. at $x=\mu$

7. The curve of normal distribution has a single peak, i.e., it is unimodal.

8. The two tails of the curve extend indefinitely and never touch the horizontal line.

9. No portion of the curve lies below the x-axis as f(x), being the probability function can naver be negative.

16.5 Constants of Normal Distribution

(i) The mean of the Normal Distribution is "μ".
(ii) The variance of the Normal Distribution is "σ^2" and the standard deviation is "σ".
(iii) The median of the Normal Distribution is "μ".
(iv) The mode of the Normal Distribution is "μ".
(v) The quartile deviation of the Normal Distribution is 0.675σ (Q.D)
(vi) The mean deviation of the Normal Distribution is 0.8σ (MD)
(vii) In a Normal Distribution, Q.D: MD: S.D :: 10:12:15

16.6 Uses of Normal Distribution:

1) The Normal Distribution can be used to approximate the Binomial and Poisson distribution.

2) It has extensive use in sampling theory.
3) It has a wide use in testing statistical hypothesis and tests of significance in which it is always assumed that the population from which the samples have been drawn should have Normal Distribution.
4) It has significant applications in a statistical quality control as the control chart in statistical quality control is closely related to Normal Distribution.
5) It can be used for smoothing and graduating a distribution which is not normal, simply by a contracting a normal case.
6) It serves as a guiding instrument in the analysis and interpretation of statistical data.

16.7 Solved Problems:

1) The scores made by a candidate in ascertain tests are normally distributed with mean 500 and standard deviation 100. What percentage of candidate receives the scores between 400 and 600?

Sol: Let x be the normal variate showing scores of candidates. Its mean μ = 500, σ = 100.

Now $z = \frac{x-\mu}{\sigma} = \frac{x-500}{100}$ is standard normal variate N(0, 1)

When x = 400, then $z = \frac{400-500}{100} = -1$

When x = 600, then $z = \frac{600-500}{100} = 1$

When x lies between 400 and 600, then z lies between -1 and 1.

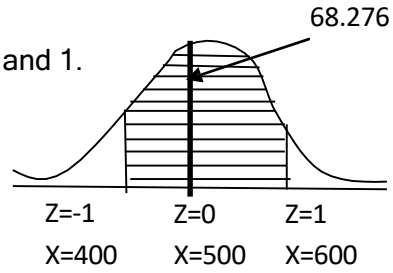
Area under the curve from (z=-1 to z=1)

= Area under the curve from

(z = -1 to z = 0) + Area under the curve from (z=0 to z=1)

=2 x (area under the curve from z=0 to z=1)

=2 x 0.34134 = 0.68268 ≅ 0.6827



∴ Percentage of candidates securing marks between 400 and 600.

= 0.6827 x 100 = 68.27%

2) 1,000 light bulbs with a mean life of 120 days are installed in a new factory. Their length of life is normally distributed with a standard deviation of 20 days.

a) How many bulbs may expire in less than 90 days?

b) If it is decided to replace all bulbs together, what interval should be allowed between replacement if not more than 10% should expire before replacement?

Sol: Let x be the normal variate of life of light bulbs. Then its mean μ = 120 and standard deviation σ = 20.

$z = \frac{x-\mu}{\sigma} = \frac{x-120}{20}$ is an standard normal variate

a) When x = 90, then $z = \frac{90-120}{20} = -1.5$

∴Number of bulbs expected to expire less than 90 days out of 1,000 bulbs = 1000 x 0.0668 ≅ 67 days.

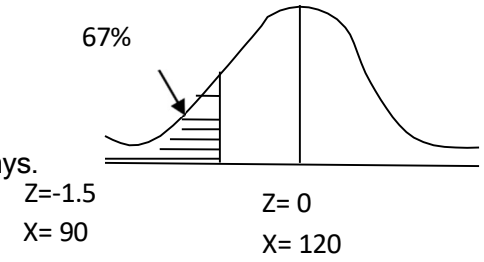
b) Since not more than 10% or 0.1 expire before replacement so the value of standard normal variate z to an area 0.5 – 0.1 = 0.4 is 1.28.

Thus the value of z is less than -1.28

$\frac{x - 120}{20}$ = -1.28

=> x = 120 – 25.6 = 94.4 of 94 days

Then the bulbs may be replaced after 94 days.



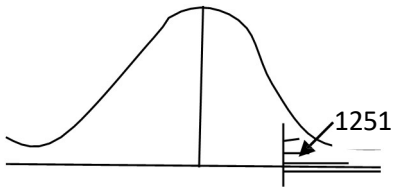
3) Assume the mean height of children to be 68.22 cm with a variance of 10.8 cm. How many children in a school of 1000 would you expect to be over 72 cm tall?

Sol: Let the distribution of height be normal. Let x be a normal variate with mean 68.22 and standard deviation $\sqrt{10.8}$

$z = \frac{x - \mu}{\sigma} = \frac{x - 68.22}{\sqrt{10.8}}$

when x = 72 cm, then

$z = \frac{72 - 68.22}{\sqrt{10.8}} = \frac{3.78}{3.286} = 1.15$



Now P(x>72) = P(z>1.15) = Area to the right of the coordinate at z = 1.15 = (Total area on the right of z = 0) – (Area from z = 0 to z = 1.15) = (0.5000 – 0.3749) = 0.1251

Expected number of children to be above 72 cm out of 1000 = 0.1251 x 1000 = 125.1 or children.

4. The marks obtained in a certain examination follow normal distribution with mean 45 and standard deviation 10. If 1000 students appeared at the examination, calculate the number of students scoring (i) less than 40 marks and (ii) more than 60 marks.

Sol: Given $\bar{x}$ = 45, σ = 10, then $z = \frac{x - \bar{x}}{\sigma} = \frac{x - 45}{10}$

(i) For x = 40 => $z = \frac{40 - 45}{10}$ = -0.5

P(x<40) = P(z<-0.5)

Or P(z>0.5) = 0.5 - P(0<z<0.5) = 0.5 - 0.1915 = 0.3085

Number of students = 1000 x 0.3085 ≅ 309

(ii) For x = 60 => $z = \frac{x - \bar{x}}{\sigma} = \frac{60 - 45}{10}$ = 1.5

P(X>60) = p(Z>1.5) = 0.5 – p(0<2<1.5)

= 0.5 – 0.4332

= 0.0668

Number of students = 1000 x 0.0668 ≅ 67

5. The life time of a certain kind of batteries has a mean life of 400 hours and standard deviation as 45 hours. Assuming the distribution of life time to be normal, find

(i) The percentage of batteries with a life time of atleast 470 hours.

(ii) The proportion of batteries with life time between 385 and 415 hours.

(iii) The minimum life of best 5% batteries.

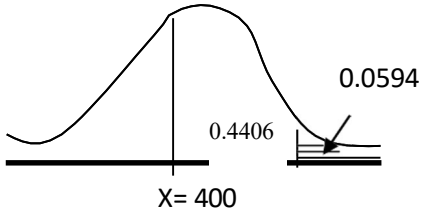
Sol: Here mean: $\bar{x}$ = 400 hours; standard deviation σ = 45 hours.

(i) Standard Normal variate z for x = 470 is

$$z = \frac{x - \bar{x}}{\sigma} = \frac{470 - 400}{45} = \frac{70}{45} = 1.56$$

∴ Area between z = 0 and z = 1.56 = 0.4406

Probability that life time of batteries is at least 470 hours = 0.5 – 0.4406 = 0.0594 = 5.94% or 6%



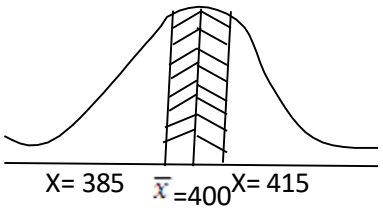
The value of z corresponding to x = 385 is

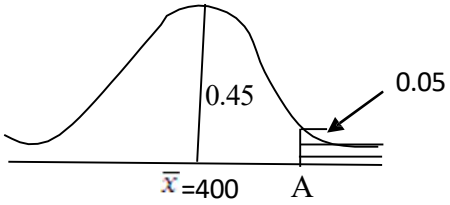
$$z_1 = \frac{x - \bar{x}}{\sigma} = \frac{385 - 400}{45} = \frac{-15}{45} = -1/3 = -0.33$$

Also when x = 415, then $z_2 = \frac{415 - 400}{45} = \frac{15}{45} = 1/3 = 0.33$

Area between z = 0 and z = 0.33 is 0.1293.

Area between x = 385 and x = 415 hours = (0.1293 + 0.1293) = 0.2586 or 26%





(iii) Life of best 5% of batteries starts at A.

The value of x at A gives minimum life of best 5% batteries. Value of z corresponding to area 0.45 to right of $\bar{x}$ = 400 = 1.65

⇨ $\bar{x}$ = 1.65 x 45 + 400

= 74.25 + 400

= 474.25 i.e., 47 hours

6. The average test marks in a particular class is 79. The standard deviation is 5. If the marks are normally distributed, how many students in a class of 200 did not receive marks between 75 and 82?

Sol: Here x = 79 and σ = 5

The value of z when x = 7 is

$z = \frac{x - \bar{x}}{\sigma}$ => $z_1 = \frac{75 - 79}{5} = \frac{-4}{5}$ = -0.8

Area between z = 0 and z = -0.8 is – 0.2881 -> (1)

The value of z when x = 82 is

$\frac{x - \bar{x}}{\sigma}$ => $z = \frac{82 - 79}{5} = \frac{3}{5}$ = 0.6

Area between z = 0 and z = 0.6 is = 0.2257

Total area between x = 75 and x = 82

= 0.2881 + 0.2257 = 0.5138

Number of students getting marks between 75 and 82

= 200 x 0.5138 = 102.76 or 103 students.

Number of students did not getting marks between 75 and 82

= 200 – 103 = 97 students.

7. A multiple choice quiz has 200 questions. Each with 4 possible answers, of which only 1 is the correct answer. What is the probability that sheer guess work yields from 25 to 30 correct answers for 80 problems about which the student has no knowledge?

Sol: Since each question has 4 possible answers and for n = 80 problems about which the student has no knowledge and uses sheer guess work, the probability of correct answer is given by

P = ¼, q = 1 – p = ¾

If the random variable x denotes the number of correct answers then using normal approximation to binomial distribution $X \sim N(\mu, \sigma^2)$ where

μ = mean = np = 80 x ¼ = 20

σ^2= variance = npq = 80 x ¼ x ¾ = 15 => σ = $\sqrt{15}$ = 3.87

The required probability that sheer guess work yields from 25 to 30 correct answer is:

$P(25 \le x \le 30) = P(\frac{25 - 20}{3.87} \le z \le \frac{30 - 20}{3.87})$

= P(1.29≤ z ≤ 2.58) = P(0≤ z ≤1.29)

= 0.4951 - 0.4015 = 0.0936

16.8 Summary:

Normal distribution is a continuous distribution. In fact, in continuous distributions, the probability at any discrete point is always zero. It is the most widely used distribution, comparing with all the distributions. Many phenomena are normally distributed, including characteristics of most machine-parts, many measurements of the biological and natural environment, and many human characteristics such as height, weight, IQ and achievement scores. The normal curve is continuous, symmetrical, unimodel, and asymptotic to the x-axis; actually, it is a family of curves. The Normal distribution can be used to work certain types of Biological and Poisson problems.

16.9 Self Assessment Questions:

1. The weekly wages of 1000 workers are normally distributed with a mean of Rs.70 and standard deviation of Rs.5. Estimate the lowest weekly wages of 100 highest paid workers.

[Ans: 76.4]

2. The mean life time of 60 watt light bulb produced by bright light bulbs company is 200 hours. It is keeping the standard deviation 20 hours. Assuming that the life times of bulbs are normally distributed, what are probabilities that single 60 watt light bulb extracted from the production lot will (i) burn out between 180 hours and 210 hours. (ii) burn out at a time greater than 250 hours.

[Ans: (i) 0.5328 (ii) 0.0062]

3. The burning time of an experimental socket is a random variable which has normal distribution with $\mu = 4.36$ seconds and $\sigma = 0.04$ seconds. What are the probabilities that his kind of socket will burn for

(i) Less than 4.5 seconds?

(ii) More than 4.40 seconds?

(iii) Between 4.30 to 4.42 seconds?

[Ans: (i) 0.003 (ii) 0.1587 (iii) 0.8664]

4. In a manufacturing organization the distribution of wages was perfectly normal and the number of workers employed in the organization was 5000. The mean wages of the workers were calculated as Rs.800 pm and the standard deviation was worked out to be Rs.200. On the basis of the information estimate.

(i) The number of workers getting salary between Rs.700 and Rs.900.

(ii) Percentage of workers getting salary above Rs.1000.

(iii) Percentage of workers getting salary below Rs.600.

[Ans: (i) 1915 (ii) 15.87% (iii) 16%

5. The income of a group of 10,000 persons was found to be normally distributed with mean equal to Rs.750 and standard deviation equal to Rs.50. What was the lowest income among the richest 250?

[Ans: Rs.848]

6. Explain the properties of Normal distribution?
7. State the uses of Normal distribution.
8. Define Normal distribution.

16.10 Reference Books

1. S.C. Gupta: Fundamentals of Statistics.
2. K. Chandra Sekhar: Business Statistics.
3. K. V. Sarma: Statistics made simple, Prentice Hall of India.

Lesson Writer

Prof. M. Koteswara Rao

17. Testing of Hypothesis

Objectives

After completion of this chapter, you should be able to:

- Understanding the assumptions of statistical hypothesis testing;
- Construct Null and Alternative Hypothesis;
- Know how to formulate a statistical hypothesis;
- Understand about basic concepts in testing of Hypothesis.

Structure

17. 1 Introduction

One of the aims of statistics is to make inferences about the unknown parameters of a population, based on the information contained in a sample that is selected from this population. The goal of making such valid inferences may be achieved by testing hypothesis about the plausible values of these unknown parameters. Testing of hypothesis is a phenomenon that we deal with in everyday life. The type of problem that arise during hypothesis testing concerns whether a sample could reasonably have come from a population having a specified distribution. Managers have to make decisions with minimum risk in an environment characterized by uncertainty. Acceptance or rejection of decision depends on acceptance or rejection of a hypothesis. A suitable hypothesis formulation and testing it would be help the managers to take the right decision. Testing of Hypothesis is the right option, which is useful in making the sound decisions.

17.2 Basic terminology and definitions.

POPULATION:- Population means the number of inhabitants in a well-defined area. Population (universe) in statistics means the whole of the information, which comes under the purview of statistical investigation. It is a totality of persons, objects, items or anything conceivable pertaining to pertain characteristics. For example, the population of students in a university, the no. of workers in the jute mill, etc. The population size is denoted by „N"

The population may be finite or infinite. If the no. of objects are finite or countable then it is known as finite population otherwise it is known as infinite population.

PARAMETER:- Any statistical measure computed from population data is known as parameter. For example population mean(μ), population standard deviation (σ), population proportion (P), etc are the parameters.

SAMPLE:- A finite part of a population or a subset of a set of sampling units, selected by some process in known as sample. The no. of objects in a sample is known as sample size and it is denoted by „n". For example, we select 50 items to test the quality of the product, from a lot of products manufactured by a factory in a month then 50 selected products constitute a sample and n=50 is the sample size.

STATISTIC:- Any statistical measure computed from sample data is known as statistic. For example sample mean ($\overline{X}$), sample standard deviation (s), sample proportion (p), etc are the statistics.

SAMPLING DISTRIBUTION OF A STATISTIC:- If we draw a sample of size „n" from a given finite population of size „N" then the total number of possible samples is N_{c_n} = k (say) ways.

For each of these samples we can compute some statistic like t, $\overline{X}$ and s^2 etc; as given below:

Sample No.	Statistics		
	t	$\overline{X}$	s^2
1	t_1	$\overline{X_1}$	s_1^2
2	t_2	$\overline{X_2}$	s_2^2
3	t_3	$\overline{X_3}$	s_3^2
.	.	.	.
.	.	.	.
.	.	.	.
k	t_k	$\overline{X_k}$	s_k^2

The set values of statistic so obtained, one for each sample constitute is called sampling distribution of the statistic.

STANDARD ERROR (S.E):- The standard deviation of sampling distribution of a statistic is known as standard error and is abbreviated as S.E. In cases of large sample tests the statistics of various statistics are shown in the following table.

S.No	Statistic	Standard error
1	Single Sample mean $\left(\overline{X}\right)$	$\sqrt{\frac{\sigma^2}{n}}$ (or) $\sqrt{\frac{s^2}{n}}$
2	Single standard deviation (s)	$\sqrt{\frac{\sigma^2}{2n}}$ (or) $\sqrt{\frac{s^2}{2n}}$
3	Sample proportion (p)	$\sqrt{\frac{PQ}{n}}$ (or) $\sqrt{\frac{pq}{n}}$

4	Difference of two sample means $\left(\overline{X}-\overline{Y}\right)$	$\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$ (or) $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$
5	Difference of two sample standard deviations ($s_1 - s_2$)	$\sqrt{\frac{\sigma_1^2}{2n_1}+\frac{\sigma_2^2}{2n_2}}$ (or) $\sqrt{\frac{s_1^2}{2n_1}+\frac{s_2^2}{2n_2}}$
6	Difference of two proportions ($P_1 - P_2$)	$\sqrt{PQ\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ Here $P = \frac{n_1p_1+n_2p_2}{n_1+n_2}$ or $P = \frac{X_1+X_2}{n_1+n_2}$ and $Q = 1 - P$

17.3 Hypothesis and statistical hypothesis:-

Hypothesis:- A hypothesis is a small statement about certain theoretical concepts based on one"s own experience or knowledge gathered directly or indirectly. It is definite statement about the population parameter. A hypothesis is used as a basis for investigation and reasoning.

According to Pro. Morris Hamburg. "A hypothesis in statistics is simply a quantitative statement about a population."

Statistical Hypothesis:- "A statistical hypothesis is some statement about a population or equivalently about the probability distribution characterizing a population which we want to verify on the basis of information available from a sample".

If the statistical hypothesis specifies the population completely then it is known as simple Statistical hypothesis, otherwise it is known as Composite statistical hypothesis

For example, if X_1, X_2.....X_n is a random sample of size „n" drawn from a normal population with mean μ and variance σ^2, then the hypothesis

H: $\mu = \mu_0$, $\sigma^2 = \sigma_0^2$ is a simple hypothesis.

The hypothesis (i) H: $\mu > \mu_0$, $\sigma^2 = \sigma_0^2$ (ii) H: $\mu > \mu_0$, $\sigma^2 = \sigma_0^2$

(iii) H: $\mu = \mu_0$, $\sigma^2 = \sigma_0^2$ etc are the Composite Hypothesis.

17.4 Characteristics of Hypothesis:-

The characteristics of hypothesis are:

(i) Hypothesis should be clear and precise.

(ii) Hypothesis should be capable of being tested.

(iii) If the hypothesis is relational hypothesis, its states the relationship between the variables.

(iv) Hypothesis should be consistent with most known facts.

(v) Hypothesis should be amenable to testing within a reasonable time.

(vi) Hypothesis must explain the facts that gave rise to the need for explanation.

17.5 Null and Alternative hypothesis:-

Null Hypothesis: - A hypothesis of no difference is called Null Hypothesis. For applying the test of significance we first set up a hypothesis, which is a definite statement about the population parameters, such a hypothesis which is usually a hypothesis of no difference is called a null hypothesis. According to prof. R.A. Fisher, "Null Hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true". It is denoted by "H_0".

In case of single statistic, null hypothesis is defined as "There is no significant difference between the sample statistic and population parameter". OR "The population parameter has a specific value". For example, H_0 : The average age of the engineering student is 18 years. i.e.; $\mu = 18$ (μ_0 say)

In case of two statistics the null hypothesis is defined as "There is no significance difference between the two sample statistics" or "The population parameters are equal". For example, H_0 : There is no significance difference between the two sample means or The population means are equal. i.e.; $\mu_1 = \mu_2$

Alternative Hypothesis:- Any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis, usually it is denoted by H_1. Given the example of null hypothesis H_0 : $\mu_1 = \mu_2$, the alternative hypothesis could be

(a) $H_1 : \mu_1 \neq \mu_2$. (b) $H_1 : \mu_1 > \mu_2$ (c) $H_1 : \mu_1 < \mu_2$

The alternative hypothesis in (a) is known as a two tailed alternative and the alternatives of (b) and (c) are known as right – tailed and left – tailed alternatives respectively.

17.6 One tailed and Two tailed tests:-

In any test of statistical hypothesis, the alternative hypothesis is one tailed (right tailed or left tailed) is called one tailed test.

For example, a testing of the two means, $H_0 : \mu_1 = \mu_2$

against the alternative hypothesis: $H_1 : \mu_1 > \mu_2$ or $H_1 : \mu_1 < \mu_2$, is a single tailed or one – tailed test.

In any test of statistical hypothesis, the alternative hypothesis is two tailed then it is called two tailed test. For example, a testing of the two means,

$H_0 : \mu_1 = \mu_2$

against the alternative hypothesis:

$H_1 : H_1 : \mu_1 \neq \mu_2$, is a two tailed test.

17.7 PROCEDURE FOR TESTING OF HYPOTHESIS:-

The following are the steps involved in testing hypothesis.

Step I: Set up the null hypothesis H_0

The following Steps must be taken into consideration while setting up a null hypothesis:

(a) If we want to test the significance of the difference between a *statistic* and the parameter or between two independent sample statistics then we set up the null hypothesis that the difference is not significant.

(b) If we want to test any statement about the population, we set up the null hypothesis that it is true.

Sep II: Set up the Alternative Hypothesis H_1.

Step III: Computer the test statistic "Z" under the H_0.

Step IV: Choose the appropriate level of significance (α) . Generally, we take α value as 5% or 1%.

Srep V : We compare the computed value of Z in step IV with the significant value Zα at given level of significance, „α".

If calculated value of Z or $|Z|$ is less than the tabulated value of Z at α% level of significance we accept H_0 otherwise we reject H_0.

17.8 Limitations of Testing of Hypothesis:-

Some of the limitations of Testing of Hypothesis are:

(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.

(ii) Statistical inferences based on the significance tests cannot be said to be entirely right evidence relating the truth of the hypothesis.

(iii) Significance tests results are based on probability and as such cannot be expressed with full of certainty.

(iv) Tests do not explain the reasons as to why does the differences exists, say between the means of two samples or standard deviations of two sample or proportions of two samples etc. They simply indicates 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 main difference.

17.9 Summary

In this chapter, we have learned various type of hypothesis. Hypothesis testing begins with the drawing of a sample and calculating its characteristics. A hypothesis is a statement about the population distribution that may or may not be true. We use hypothesis tests to make an inference about some population parameter of interest. The first and foremost steps in testing of hypothesis are to define null and alternative hypothesis. Testing of hypothesis is a two action problem – accept the hypothesis or reject the hypothesis.

17.10 Self-Assessment Questions

1. Define (a) Parameter (b) Statistic (c) Standard Error
2. Explain about the Procedure of Testing of Hypothesis
3. State the chief characteristics of Testing of Hypothesis
4. Define Null hypothesis and alternative hypothesis with suitable examples.
5. State the limitations of Testing of Hypothesis.
6. Define one tailed and two tailed tests.

17.11 Reference Books

1. S.C. Gupta, Fundamentals of Statistics, Himalaya Publishing House.
2. Digambar Patri., D.N. Patri, Quntitative Techniques, Kalyani publications.
3. P.N,Arora and S. Arora, Statistics for Management: S.Chand & Comp.Ltd.
4. G.v. Shenoy, Uma K.Srivastava, S.C.Sharma. : Business Statistics
5. B.M.Agarwal,: Business statistics
6. Gupta S.P. : Statistical Methods

Lesson Writer

Dr. J. Pratapa Reddy

18. Hypothesis testing - Z test

Objectives

After completion of this chapter, you should be able to:

- Understanding the assumptions of Large sample test
- Apply the Z –test for means, Standard deviations and proportions;

Structure

18.1 Introduction

The main objective of the sampling theory is the study of the tests of Hypothesis or tests of significance. Most of the situations, we are to make decisions about the population on the basis of the sample information. For example, on the basis of sample data, (i) a statistician has to decide whether a given die is unbiased or not; whether a coin is unbiased or not (ii) a quality control engineer is to determine whether a process is properly working, (iii) a drug chemist is to decide whether a new drug is really effective in curing a disease etc. Such decisions are generally called statistical decisions. The theory of testing of hypothesis employs various statistical techniques to arrive at such decisions on the basis of the sample study.

18.2 Large samples test and assumptions: -

Large samples test are a parametric tests. If the sample size (n) is greater than 30 then those samples may be regarded as large samples. The basic assumptions of large samples tests are:

(i) The observations are independent

(ii) The observations should be drawn from normal population

(iii) Sampling values are sufficiently close to the population value and can be used for the calculation of standard error of a statistic.

18.3 Critical values of Z – test:-

We have to apply Two – tailed test or right – tailed test is based on alternative hypothesis. For large samples, the critical values of 'Z' at various levels are obtained from normal distribution area tables are shown directly in the following table.

Test	Level of significance (α)			
	1%	2%	5%	10%
Two – tailed test	$\left\|Z_{\alpha/2}\right\|$ = 2.58	$\left\|Z_{\alpha/2}\right\|$ = 2.33	$\left\|Z_{\alpha/2}\right\|$ = 1.96	$\left\|Z_{\alpha/2}\right\|$ = 1.645
Right – tailed test	Z_α = 2.33	Z_α = 2.05	Z_α = 1.645	Z_α = 1.28
Left - tailed test	- Z_α = - 2.33	- Z_α = - 2.05	- Z_α = - 1.645	- Z_α = - 1.28

18.4 Z - test for Single Mean

Suppose we want to test whether there is any significant difference between sample mean $(\bar{x})$ and population (μ) (or) the population mean has a specific value (say μ_0).

Working procedure:

Step 1: Null Hypothesis: There is no significant difference between sample mean $(\bar{x})$ and population (μ) (or) the population mean has a specific value (say μ_0).

i.e., $H_0 : \mu = \mu_0$

Step 2: Alternative Hypothesis: $H_1: \mu \neq \mu_0$ (sometimes $\mu \geq \mu_0$ or $\mu \leq \mu_0$)

Step 3: If σ is known then the required test statistic for testing the above null hypothesis is

$$Z = \frac{\bar{x} - \mu_0}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

If σ is not known (For large samples $\sigma^2 \cong s^2$) then the required test statistic for testing the above null hypothesis is

$$Z = \frac{\bar{x} - \mu_0}{\sqrt{s^2/n}} \sim N(0,1)$$

Where $\bar{x}$ = sample mean and n = sample size.

Step IV: Chosen the appropriate level of significance (α). Generally, we take α value as 5% or 1%

Step V: We compare the computed value of Z in step IV with the significant value of $Z\alpha$ at given level of significance, 'α'.

If calculate value of Z or $|Z|$ is less than the tabulated value of Z at α% level of significance we accept H_0 otherwise we reject H_0.

Note: The probable limits for the population mean are

$$\bar{x} \pm Z_{\alpha/2} \sqrt{s^2/n}$$

18.5 Z – test for Two Means:-

Suppose we want to test whether there is any significant difference between two sample means $(\bar{x}_1 - \bar{x}_2)$ (or) the population means are equal.

Working procedure:

Step 1: Null Hypothesis: There is no significant difference between the two sample means or the population means are equal i.e., $H_0: \mu_1 = \mu_2$

Step 2: Alternative Hypothesis $H_1: \mu_1 \neq \mu_2$ (sometimes $\mu_1 \geq \mu_2$ or $\mu_1 \leq \mu_2$)

Step 3: If σ_1^2 and σ_2^2 are known then the required test statistic for testing the above null hypothesis is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

If σ_1^2 and σ_2^2 are not known (For large samples $\sigma_1^2 \cong s_1^2$ and $\sigma_2^2 \cong s_2^2$) then the required test statistic for testing the above null hypothesis is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1)$$

where $\bar{x}_1$ = mean of the first sample and n_1 = size of the first sample.

$\bar{x}_2$ = mean of the second sample and n_2 = size of the second sample.

Step IV: Choose the appropriate level of significance (α). Generally, we take α value as 5% or 1%

Step V: We compare the computed value of Z in step IV with the significant value $Z\alpha$ at given level of significance, 'α'.

If calculate value of Z or $|Z|$ is less than the tabulated value Z at α% level of significance we accept H_0 otherwise we reject H_0.

Note: the probable limits for the difference between two means are

$$\bar{x}_1 - \bar{x}_2 \pm Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

18.6 Z – test for Single standard deviation: -

Suppose we want to test whether there is any significant difference between sample standard deviation (s) and population standard deviation (σ) (or) the population standard deviation has a specific value (Say σ_0).

Working procedure:

Step 1: Null Hypothesis: There is no significant difference between sample standard deviation (s) and population standard deviation (σ) (or) the population mean has a specific value (say σ_0). i.e., H_0: $\sigma = \sigma_0$

Step 2: Alternative Hypothesis: H_1: $\sigma \neq \sigma_0$ (sometimes $\mu \geq \mu_0$ or $\sigma \leq \sigma_0$)

Step 3: If σ is known then the required test statistic for testing the above null hypothesis is

$$Z = \frac{s - \sigma}{\sqrt{\sigma^2/2n}} \sim N(0,1)$$

If σ is not known (For large samples $\sigma^2 \cong s^2$) then the required test statistic for testing the above null hypothesis is

$$Z = \frac{s - \sigma_0}{\sqrt{s^2/2n}} \sim N(0,1)$$

where $\bar{x}$ = sample mean and n = sample size.

Step IV: Chosen the appropriate level of significance (α). Generally, we take α value as 5% or 1%

Step V: We compare the computed value of Z in step IV with the significant value $Z\alpha$ at given level of significance, 'α'.

If calculate value of Z or $|Z|$ is less than the tabulated value of Z at α% level of significance we accept H_0 otherwise we reject H_0.

18.7 Z –test for Two Standard deviations:-

Suppose we want to test whether there is any significant difference between two sample standard deviations (σ_1 - σ_2) (or) the population standard deviations are equal.

Working procedure:

Step 1: Null Hypothesis: There is no significant difference between the two sample standard deviation or the population standard deviation are equal

i.e., H_0: $\sigma_1 = \sigma_2$

Step 2: Alternative Hypothesis: H_1: $\sigma_1 \neq \sigma_2$ (sometimes $\sigma_1 \geq \sigma_2$ or $\sigma_1 \leq \sigma_2$)

Step 3: The required test statistic for testing the above null hypothesis is

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}} \sim N(0,1)$$

If σ_1^2 and σ_2^2 are not known (For large samples $\sigma_1^2 \cong s_1^2$ and $\sigma_2^2 \cong s_2^2$) then the required test statistic for testing the above null hypothesis is

$$Z = \frac{s_1 - s_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}} \sim N(0,1)$$

Where $\bar{x}_1$ = mean of the first sample and n_1 = size of the first sample.

$\bar{x}_2$ = mean of the second sample and n_2 = size of the second sample.

Step IV: Choose the appropriate level of significance (α). Generally, we take α value as 5% or 1%

Step V: We compare the computed value of Z in step IV with the significant value $Z\alpha$ at given level of significance, 'α'.

If calculate value of Z or $|Z|$ is less than the tabulated value of Z at α% level of significance we accept H_0 otherwise we reject H_0.

18.8 Z – test for Single Proportion

Suppose we want to test whether there is any significant difference between sample proportion (p) and population proportion, (P) (or) the population proportion has a specific value (Say P_0)

Working procedure:

Step 1: Null Hypothesis: There is no significant difference between sample proportion and Population proportion (or) the population proportion has a specific value say P_0

i.e., $H_0: P = P_0$

Step 2: Alternative Hypothesis: $H_1: P \neq P_0$ (sometimes $P \geq P_0$ or $P \leq P_0$)

Step 3: The required test statistic for testing the above null hypothesis is

$$Z = \frac{p - P}{\sqrt{PQ/n}} \sim N(0, 1)$$

Where p =no. of persons possessing the certain attribute (favour to the proposal) = X/n

Step IV: Choose the appropriate level of significance (α). Generally, we take α value as 5% or 1%

Step V: We compare the computed value of Z in step IV with the significant value $Z\alpha$ at given level of significance, 'α'.

If calculated value of Z or $|Z|$ is less than the tabulated value Z at α% level of significance we accept H_0 otherwise we reject H_0.

Note: the probable limits for the observed proportion of successes are

$$P \pm Z_{a/2}\sqrt{PQ/n}$$

18.9 Z –test for two Proportions

Suppose we want to test whether there is any significant difference between two sample proportions or (Or) the population proportions are equal.

Working procedure:

Step 1: Null Hypothesis: There is no significant difference between two sample proportions or the population proportions are equal.

i.e., $H_0: P_1 = P_2$

Step 2: Alternative Hypothesis: $H_1: P_1 \neq P_2$ (some times $P_1 \geq P_2$ or $P_1 \leq P_2$)

Step 3: The required test statistic for testing the above null hypothesis is

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

Where $\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ or $\hat{P} = \frac{X_1 + X_2}{n_1 + n_2}$

$\hat{Q} = 1 - \hat{P}$

X_1 = No. of persons possessing the attribute (quality) 1

X_2 = No. of persons possessing the attribute (quality) 2

$p_1 = \frac{X_1}{n_1}$ and $p_2 = \frac{X_2}{n_2}$

Step IV: Choose the appropriate level of significance (α). Generally, we take α value as 5% or 1%

Step V: We compare the computed value of Z in step IV with the significant value $Z\alpha$ at given level of significance, 'α'.

If calculated value of Z or $|Z|$ is less than the tabulated value Z at α% level of significance we accept H_0 otherwise we reject H_0.

Note: The probable limits for the difference between two proportions are

$$p_1 - p_2 \pm Z_{\alpha/2}\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

18.10 Solved Problems

(1) A random Sample of 1600 students has a mean score 99. Test whether the sample has been drawn from a population with mean score 100 and S.D. 15

Sol:- Given that n=1600, $\bar{x}$ = 99 μ =100, σ =15

σ is given (Known)

Null Hypothesis:- H_0:μ =100

i.e., The sample has been drawn from a population

Alternative Hypothesis

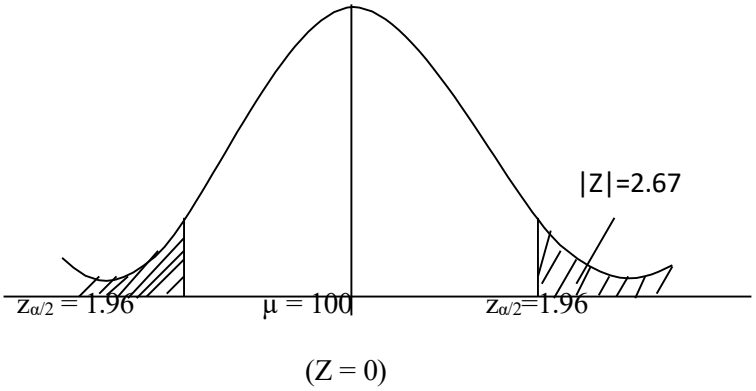
H_1:μ ≠ 100 (Two tailed test)

i.e., the sample has not drawn from the population

Test Statistic under H_0: $Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$

$$= \frac{99-100}{15/\sqrt{1600}}$$

$$= -2.67$$



Inference $|Z|$ = 2.67

Zα = ± 1.96 at 5% level of significance and for two tailed test from standard normal tables

$|Z| > Z\alpha \Rightarrow H_0$ is rejected

i.e., the Sample has not drawn from the populations

(2) A Sample of 100 students is taken from a large population. The mean height of the students in the sample is 160 cms. Can it be reasonably regarded that, in the population, the mean height is 165 Cm and the S.D is 10 Cm

Solution:- Given that n=100, $\bar{x}$ =160, μ =165 σ =10

Null Hypothesis: H_0:μ =165

i.e., the sample has drawn from the population with mean height 165 cms.

Alternative Hypothesis:

H_1:μ ≠ 165 (Two tailed test)

i.e., the sample has not considered from the population with mean height 165 Cm

Test Statistic under H_0: $Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$

$$= \frac{160 - 165}{10 / \sqrt{100}}$$

= - 5

Inference $|Z|$ = 5

Since $|Z| > 3$, always we reject H_0

i.e., the given Sample has not drawn from the population with the mean height 165 Cm.

(3) A random Sample of 100 items, drawn from a universe with mean value 64 and S.D. 3 has a mean value 63.5. Is the difference in the means significant? What will be your inference if the sample had 200 items.

Solution: Given that n = 100, $\bar{x}$ =64, S = 3, μ =63.5

σ is not known

Null Hypothesis: H_0: μ = 63.5

i.e., the difference in the means not significant

Alternative Hypothesis:

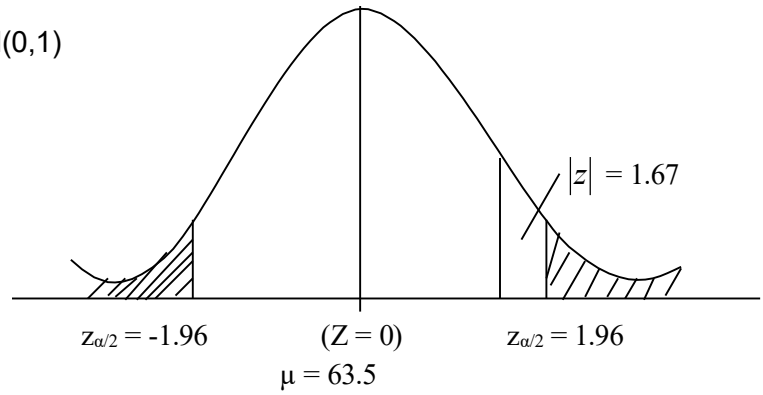
$H_1: \mu \neq 63.5$ (Two tailed test)

i.e., the difference in the means is significant

Test statistic under H_0: $Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim N(0,1)$

$= \frac{64 - 63.5}{3/\sqrt{100}}$

$= 1.67$



Inference: $|Z| = 1.67$

$Z\alpha = 1.96$ at 5% level of significance and for two tailed test from standard normal tables

$\therefore |Z| < Z\alpha \Rightarrow H_0$ is accepted

i.e., the difference between means is not significant if n = 200 then test statistic under H_0

$$Z = \frac{64 - 63.5}{3/\sqrt{100}} = 2.36$$

Inference: $|Z| = 2.36$

$Z\alpha = 1.96$

$\therefore |Z| > Z\alpha \Rightarrow H_0$ is rejected

In this case (if n =200) the difference between means is significant

(4) In a random sample of size 500, the mean is found to be 30, in another random Sample of size 400, the mean is 35. Could the samples have been drawn from the same population with SD. 8

Solution:- Given that $n_1 = 500$, $n_2 = 400$, $\bar{x} = 30$, $\bar{y} = 35$ and $\sigma_1 = \sigma_2 = \sigma = 8$

Null Hypothesis: $H_0: \mu_1 = \mu_2$

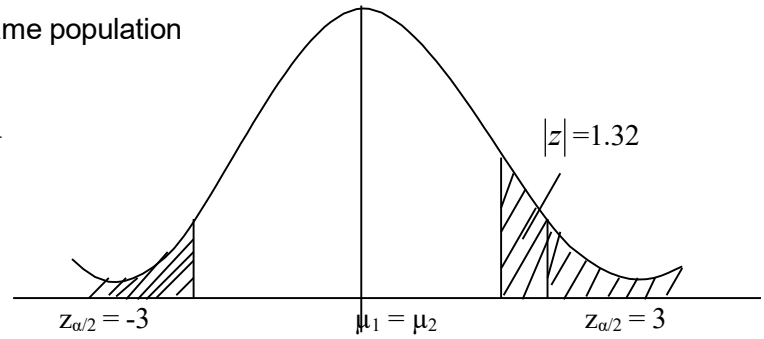
i.e., two samples have been drawn from the same population

Alternative Hypothesis: $H_1 : \mu_1 \neq \mu_2$ (two tailed test)

i.e., two samples have not drawn from the same population

Test Statistic under H_0;

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



$$= \frac{30 - 35}{8\sqrt{\frac{1}{500} + \frac{1}{400}}}$$

$$= -9.32$$

Inference: $|Z| = 9.32$

If $|Z| > 3$, we always reject H_0

i.e., two samples were not drawn from the some population.

(5) The means of two large samples of 1000 and 2000 members are 67.5″ and 68.0” respectively. Can the samples be regarded as drawn from the same population with standard deviation 2.5”.

Solution:- Given that $n_1 = 1000$, $n_2 = 2000$, $\bar{x} = 67.5$, $\bar{y} = 68$

And $\sigma_1 = \sigma_2 = \sigma = 2.5$

Null Hypothesis: $H_0: \mu_1 = \mu_2$

i.e., two samples have drawn from the same population

Alternative Hypothesis: $H_1: \mu_1 \neq \mu_2$ (Two tailed test)

i.e., two samples have not drawn from the same population

Test statistic under H_0: $Z = \dfrac{\bar{x} - \bar{y}}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

$= \dfrac{67.5 - 68}{(2.5)\sqrt{\dfrac{1}{1000} + \dfrac{1}{2000}}} = -5.16$

Inference: $|Z| = 5.16$

If $|Z| > 3$, H_0 is always rejected i.e., two sample have not come from the same population.

(6) Intelligence tests were given to two groups of boys and girls of the same age group chosen from the same college and the following results were obtained

	Size	Mean	S.D
Boys	100	73	10
Girls	60	75	8

Examine whether the difference between the mean is significant or not

Solution: Given that $n_1 = 100$ $n_2 = 60$

$\bar{x} = 73$ $\bar{y} = 75$

$S_1 = 10$ $S_2 = 8$

σ_1 and σ_2 are not known

$\hat{\sigma}_1 = S_1 = 10$, $\hat{\sigma}_2 = s_2 = 8$

Null Hypothesis:

$H_0: \mu_1 = \mu_2$

i.e., there is no significant difference between the Sample means

Alternative Hypothesis:

$H_1: \mu_1 \neq \mu_2$ (Two tailed test)

i.e., there is a significant difference between the sample mean

Test statistic under H_0:

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1)$$

$$= \frac{73 - 75}{\sqrt{\frac{(10)^2}{100} + \frac{(8)^2}{60}}} = -1.39$$

Conclusion: $|Z| = 1.39$

$Z\alpha = 1.96$ at 5% level of significance and for two tailed test from standard normal tables

$\therefore |Z| < Z\alpha \Rightarrow H_0$ accepted

i.e., there is no significant between the sample means

(7) A random sample of size 50 has S.D.11.8 drawn from a normal population. Can we assume that the sample has been drawn from the population with S.D. 10

Solution:- Given that n =50, s =11.8, $\sigma = 10$

Null Hypothesis: $H_0: \sigma = 10$

i.e., the sample has drawn from a population

Alternative Hypothesis: $H_1: \sigma \neq 10$ (Two – tailed test)

i.e., The sample has not drawn from the population

Test Statistic under H_0: $Z = \frac{s - \sigma_0}{\sigma_0 / \sqrt{2n}} \sim N(0,1)$

$$= \frac{11.8 - 10}{10 / \sqrt{2 \times 50}} = 1.8$$

Inference: $|Z|$ = 1.8

$Z\alpha$ = 1.96 at 5% of significance and for two tailed test from standard normal tables

$\therefore |Z| < Z\alpha \Rightarrow$ we accept H0

i.e., the sample has drawn from the population.

(8) The standard deviations of two samples of sizes 1000 and 500 are 2.6 and 2.7 respectively. Assuming the sample are independent find whether the two samples have come from the populations with same S.D

Solution:- Given that $n_1 = 1000$ $n_2 = 500$

$S_1 = 2.6$ $s_2 = 2.7$

σ_1 and σ_2 are not given, then they estimated by $\hat{\sigma}_1 = s1 = 2.6$ $\hat{\sigma}_2 = s_2 = 2.7$

Null Hypothesis: $H_0: \sigma_1 = \sigma_2$

i.e., two samples have come from the same population

Alternative Hypothesis:

$H_1: \sigma_1 \neq \sigma_2$

i.e., two samples have not drawn from the same population

Test statistic under H_0: $Z = \dfrac{s_1 - s_2}{\sqrt{\dfrac{s_1^2}{2n_1} + \dfrac{s_2^2}{2n_2}}} \sim N(0,1)$

$$= \frac{2.6 - 2.7}{\sqrt{\dfrac{(2.6)^2}{2\times 1000} + \dfrac{(2.7)^2}{2\times 500}}} = -0.3397$$

Inference: $|Z|$ = 0.3397

$Z\alpha$= 1.96 at 5% level of significance and for two tailed test from standard normal tables

$\therefore |Z| < Z\alpha \Rightarrow H_0$ may be accepted

i.e., two samples were drawn from the same population.

(9) The data of two random samples was given below. Test the significance difference between the two samples standard deviations at 1% level of significance.

	Sample – I	Sample - II
Size	50	80
S.D	70	60

Solution: Given that $n_1 = 50$ $n_2 = 80$

$s_1 = 70$ $s_2 = 60$

σ_1, σ_2 are not given, they estimated by $\hat{\sigma}_1 = s_1 = 70, \hat{\sigma}_2 = s_2 = 60$

Null Hypothesis: $H_0: \sigma_1 = \sigma_2$

i.e., there is no significant difference between the two sample standard deviations

Alternative Hypothesis: $H_1: \sigma_1 \neq \sigma_2$ (Two – tailed test)

i.e., there is a significant difference between the two sample standard deviations

Test Statistic under H_0: $Z = \dfrac{s_1 - s_2}{\sqrt{\dfrac{s_1^2}{2n_1} + \dfrac{s_2^2}{2n_2}}} \sim N(0,1)$

$$= \frac{70 - 60}{\sqrt{\dfrac{(70)^2}{2 \times 50} + \dfrac{(60)^2}{2 \times 80}}}$$

$= 1.183$

Inference: $|Z| = 1.183$

$Z\alpha = 1.96$ at 5% level of significance and for two – tailed test from standard normal tables.

$\therefore |Z| < Z\alpha \Rightarrow H_0$ may be accepted.

i.e., there is no significant difference between the two sample standard deviations.

(10) Two random samples drawn from two countries gave the following data relating to the heights of the males.

	Country – I	Country – II
Sample size	1000	1200
Mean height (in inches)	67.42	67.25
S.D (In inches)	2.58	2.50

(i) Is the difference between the means significant?

(ii) Is the difference between the standard deviations significance?

Solution: Given that $n_1 = 1000$ $n_2 = 1200$

$\bar{x} = 67.42$ $\bar{y} = 67.25$

$S_1 = 2.58$ $s_2 = 2.50$

σ_1 and σ_2 are not known, estimated by

$\hat{\sigma}_1 = s_1 = 2.58$ $\hat{\sigma}_2 = s_2 = 2.50$

1) Test for means:

Null Hypothesis: $H_0: \mu_1 = \mu_2$

i.e., there is no significant difference between the sample means

Alternative Hypothesis: $H_1: \mu_1 \neq \mu_2$ (two tailed test)

i.e., these is no significant difference between the sample means

Test statistic under H_0:

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0,1)$$

$$= \frac{67.42 - 67.25}{\sqrt{\frac{(2.58)^2}{1000} + \frac{(2.5)^2}{1200}}} = 1.56$$

Inference: $|Z|$ = 1.56

$Z\alpha$ = 1.96 at 5% level of significance, two tailed test from standard normal tables

$\therefore |Z| < Z\alpha \Rightarrow H_0$ is accepted

(ii) Test for standard deviations:

Null Hypothesis: $H_0: \sigma_1 = \sigma_2$

i.e., There is no significant difference between the sample standard deviations

Alternative Hypothesis: $H_1: \sigma_1 \neq \sigma_2$ (Two – tailed test)

i.e., there is a significant difference between the sample standard deviations

Test statistic under H_0:

$$Z = \frac{S_1 - S_2}{\sqrt{\frac{s_1^2}{2n_1} + \frac{s_2^2}{2n_2}}} \sim N(0,1)$$

$$= \frac{2.58 - 2.50}{\sqrt{\frac{(2.58)^2}{2\times1000} + \frac{(2.50)^2}{2\times1200}}} = 1.03$$

Inference: $|Z|$ = 1.03

$Z\alpha$ = 1.96 at 5% level of significance and for two tailed test from standard normal tables

$|Z| < Z\alpha \Rightarrow H_0$ may be accepted

i.e., there is no significant difference between the sample standard deviations.

11) A die is thrown 900 times and a face of 3 or 5 is observed 335 times. Test whether the dice is unbiased

Solution:- Given that n = 900, x = 335

P = probability of getting 3 or 5 = $\frac{2}{6} = \frac{1}{3}$

q = 1 - $\frac{1}{3}$

= $\frac{2}{3}$

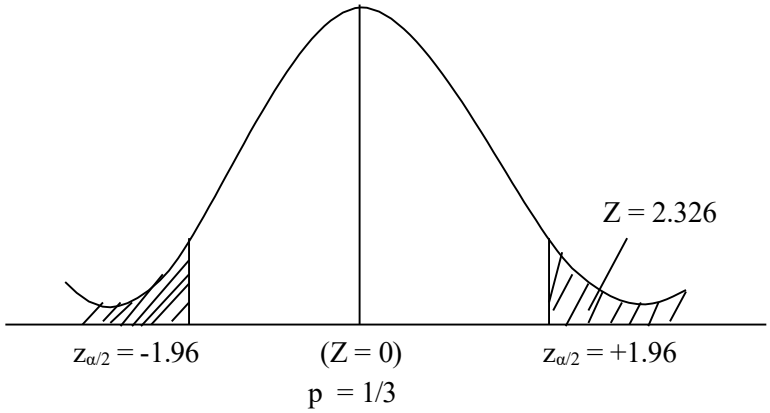
Null Hypothesis: H_0: P = $\frac{1}{3}$ i.e., dice is unbiased

Alternative Hypothesis: H_0:P≠ $\frac{1}{3}$ i.e., dice is not unbiased (two tailed test)

Test Statistic under H_0:

P = $\frac{x}{n} = \frac{335}{900}$

$Z = \frac{p - p_0}{\sqrt{\frac{P_0 Q_0}{n}}} \sim N(0,1)$



$= \frac{\frac{335}{900} - \frac{1}{3}}{\sqrt{\frac{\frac{1}{3}.\frac{2}{3}}{900}}}$

= 2.326

Inference = $|Z|$ = 2.326

Tabulated value (critical value) of Z at 5% level of significance for two tailed test is Zα = 1.96

$\therefore |Z| = Z\alpha$

$\Rightarrow H_0$ is rejected

i.e., dice is not unbiased

12) A coin was thrown 400 times and head resulted 240 times test whether coin is unbiased at 1% level of significance

Solution: Given that n = 400, x = 240

P = Probability of getting a head

$P = \frac{1}{2}$, $Q = \frac{1}{2}$,

$P = \frac{x}{n} = \frac{240}{400}$

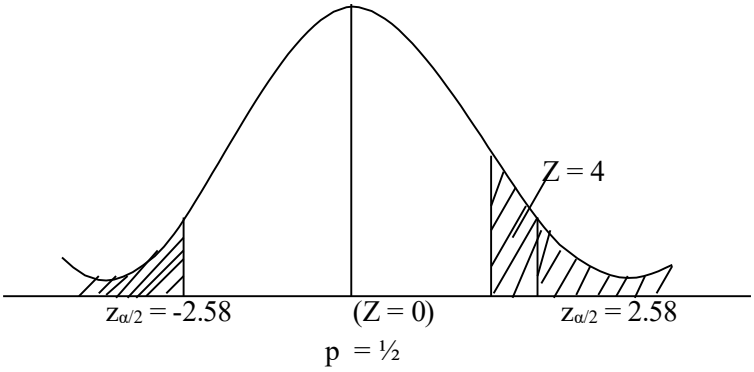
Null hypothesis: H_0: $P = \frac{1}{2}$,

i.e., Coin is unbiased

Alternative Hypothesis: H_1: $P \neq \frac{1}{2}$,

Test statistic under H_0:

$Z = \frac{p - p_0}{\sqrt{\frac{P_0 Q_0}{n}}} \sim N(0,1)$



$= \frac{\frac{240}{400} - \frac{1}{2}}{\sqrt{\frac{\frac{1}{2}.\frac{1}{2}}{400}}}$

Inference: $|Z| = 4$

$Z\alpha$ = 2.58 at 1% level of significance and for two tailed test from standard normal tables

$|Z| > Z\alpha \Rightarrow H_0$ is rejected

$|Z| > 3 \Rightarrow$ we always reject H_0

$\therefore$ The coin is not unbiased

13) In a sample of 1000 people in Maharashtra, 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in Maharashtra at 1% level of significance

Solution:- Given that n = 1000, x = 540

$P = \frac{540}{100} = 0.54$

P = Probability of rice eaters in Maharashtra = $\frac{1}{2} = 0.5$

$Q = 1 - P = 1 - \frac{1}{2} = \frac{1}{2} = 0.05$

Null Hypothesis:- $H_0 : P = 0.5$

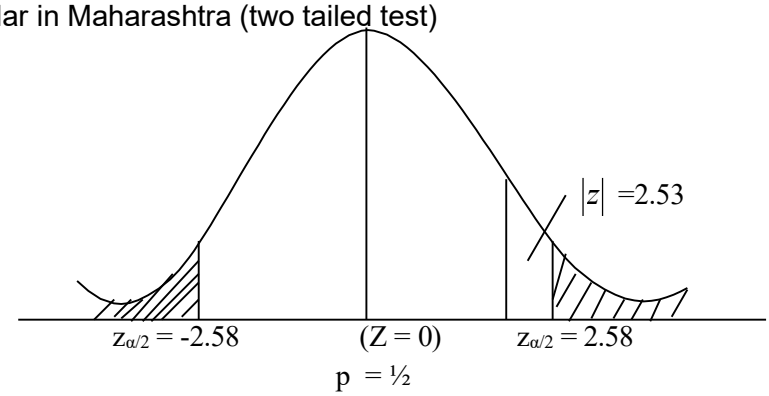
i.e., both rice and wheat are equally popular in Maharashtra

Alternative Hypothesis: $H_1 : P \neq 0.5$

i.e., rice and wheat are equally popular in Maharashtra (two tailed test)

Test Statistic under H_0:

$$Z = \frac{p - P}{\sqrt{\frac{P_0 Q_0}{n}}} \sim N(0,1)$$



$= \dfrac{0.54 - 0.5}{\sqrt{\dfrac{(0.5)(0.5)}{1000}}} = 2.53$

Inference: $|Z| = 2.53$

$Z\alpha = 2.53$ at 1% level of significance, for two tailed test from standard normal tables

$|Z| < Z\alpha \Rightarrow H_0$ is accepted

i.e., both rice and wheat are equally popular in Maharashtra.

(14) Random samples of 400 men and 600 women were asked whether they would like to have a flyover near their residence. 200 men and 325 women were infavor of the proposal. Test the hypothesis that proportions of men and women in favor of the proposal are some or not at 5% level of significance.

Solution:- Given that $n_1 = 400$, $n_2 = 600$

$X_1 = 200, \quad x_2 = 325$

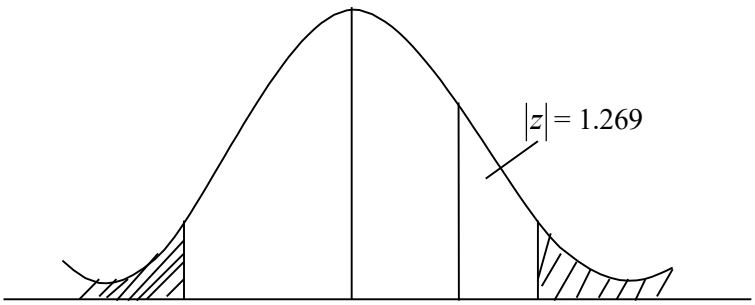
$P_1 = \dfrac{x_1}{n_1} = \dfrac{200}{400} = 0.5$

$P_2 = \dfrac{x_2}{n_2} = \dfrac{325}{600} = 0.5417$

P is not known

$\therefore \ \hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2} = \dfrac{200 + 325}{400 + 600}$

$= \dfrac{525}{1000} = 0.525$



$P_1 = P_2$

$\hat{Q} = 1 - \hat{P} = 1 - 0.525$

= 0.475

Null Hypothesis: H_0: $P_1 = P_2 = P$

i.e., there is no significance between the opinion of men and women about flyover

Alternative Hypothesis:

$H_1 : P_1 \neq P_2$ (Two tailed test)

i.e., the opinions of men and women about the flyover are not same

Test statistic under H_0:

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{0.5 - 0.5417}{\sqrt{(0.525)(0.475)\left(\frac{1}{400} + \frac{1}{600}\right)}}$$

= 1.269

Inference: $|Z|$ = 1.269

$Z\alpha$ = 1.96 at 5% level of significance and for two tailed test from standard normal tables

$\therefore |Z| < Z\alpha \Rightarrow$ H0 os accepted

i.e., Proportion of men and women about flyoer are same

15) In a random sample of 500 men form a particular district of U .P., 300 are found to be smokers. In one of 1000 men from another district, 550 are smokers. Do the data indicate that the two districts are significantly different with respect to the prevalence of smoling among men

Solution: Given that n_1 = 500, n_2=1000

x_1 = 300, x_2 = 550

$$P_1 = \frac{x_1}{n_1} = \frac{300}{500} = 0.6$$

$$P_2 = \frac{x_2}{n_2} = \frac{550}{1000} = 0.55$$

P is not known,

$$\therefore \ \hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{300 + 550}{500 + 1000}$$

$$= \frac{850}{1500} = 0.57$$

$$\hat{Q} = 1 - \hat{P} = 1 - 0.57$$

$$= 0.43$$

Null Hypothesis: H_0: $P_1 = P_2 = P$

i.e., there is no significant difference between the smokers of two districts in up

Alternative Hypothesis:

H_1:$P_1 \neq P_2$ (two tailed test)

i.e., there is a significant difference between the smokers of two districts in UP

Test statistic under H_0:

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

$$= \frac{0.6 - 0.55}{\sqrt{(0.57)(0.43)\left(\frac{1}{500} + \frac{1}{1000}\right)}}$$

$$= 1.84$$

Inference : $|Z|$ = 1.84

$Z\alpha$ = 1.96 at 5% level of significance and for two tailed test from standard normal tables

$\therefore\ |Z| < Z\alpha \Rightarrow$ H0 is accepted

i.e., there is no significant difference between the smokers of two districts in UP

(16) A machine produces 16 defective blots in a batch of 500 bolts. After the machine is overhauled, it produces 3 defective bolts in a batch of 100 bolts. Has the machine improved?

Solution: Given that n= = 500, n_2 = 5100

x_1=16, x_2 = 3

$$P_1 = \frac{x_1}{n_1} = \frac{16}{500} = 0.032$$

$$P_2 = \frac{x_2}{n_2} = \frac{3}{100} = 0.03$$

P is not known

$$\therefore\ \hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{16+3}{500+100}$$

$$= \frac{19}{600} = 0.0317$$

$\hat{Q} = 1 - \hat{P}$ = 1 – 0.0317

= 0.9686

Null Hypothesis: H_0: $P_1 = P_2 = P$

Machine has not improved (condition is same)

Alternative Hypothesis: H_1: $P_1 = P_2$ (one – tailed test)

i.e., Machine has improved

Test statistic under H_0:

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

$$= \frac{0.032 - 0.03}{\sqrt{(0.0317)(0.09683)\left(\frac{1}{500} + \frac{1}{100}\right)}}$$

$$= 0.1042$$

Conclusion: $|Z| = 0.1042$

$Z\alpha = 1.645$ at 5% level of significance and for one tailed test from standard normal tables

$\therefore |Z| < Z\alpha \Rightarrow H_0$ is accepted

i.e., the machine has not improved

18.11 Summary

In this chapter we have learned various aspects of testing of hypothesis. First, we dealt with hypothesis testing for one sample where we used test procedures for testing hypotheses about true mean, true variable, and true proportion. Then we discussed the comparison of two population through their true means, true variances and true proportions. For large samples z – statistics is used. Using the different types of Z –test we analyzed the quantitative data and qualitative data. The basic assumption for applying large samples test is the sample size should be more than 30. The critical values for one tailed and two tailed tests at various level of significances are specified.

18.12 Self assessment questions

1. Define large samples test. State the assumptions.

2. Explain the procedure of test for single mean.

3. Explain the procedure of test for two means.

4. Write the procedure of test for single proportion

5. Explain about the procedure of test for two proportions.

6. Sample of students were drawn from two college and from their weights in kgm; means and standard deviations are calculated. Make at large samples test to test the significance between the means.

	Mean	S.D	Sample Size
College A	55	10	400
College B	57	15	100

[Ans: Z = 1.2648: Net significant]

7. A sample of 400 individuals is found to home a mean height of 67.47 inches. Can it be reasonably regarded as a sample from a large population with mean height of 67.39 inchest and standard deviation 1.30 inches.

[Ans: Year, Z = 1.23]

8. A machine puts out 16 imperfect articles in a sample of 500. After machine is overhauled in puts out 3 imperfect articles in a batch of 100. Has the machine improved

(Ans: Z = 1.04, if is not significant at 5% level)

9. In a large city, 16 out of random sample 500 men were found to be tea drinkers. After the heavy increase in tax on intoxicants another random sample of 100 men in the same city included 3 drinkers. Was the disserved decrease in the proportion of drinkers significant after tax increase?

[Ans: Z = 1.04, Not significant]

10. The mean yield of two sets of plats and their variability are as given below:

	Set of 40 plats	Set of 60 plats
Mean	1258	1243
SD	34	28

(i) Is the difference in the mean yields of two sets significant.

(ii) Is the difference in the variability yields of two sets significant.

[Ans: (i) Z = 2.3 (ii) z=1.3}

11. Explain the procedure of test for two standard deviations.

12. Write the procedure of test for single standard deviation.

18.13 Reference Books

1. S.C. Gupta, Fundaments of Statistics, Himalaya Publishing House.

2. Digambar Patri., D.N.Patri, Quantitative Techniques, Kalyani publications.

3. P.N.Arora and S.Arora, Statistics for Management: S.Chand & Comp.Ltd.

4. G.V. Shenoy, Uma K.Srivastava, S.C.Sharma.: Business Statistics

5. B.M.Agarawal,: Business statistics

6. Gupta S.P: Statistical methods

Lesson Writer

Dr. J. Pratapa Reddy

19. Hypothesis testing – t test and Chi-square test

Objectives

After completion of this chapter, you should be able to:

- Understanding the assumptions of Small samples test;
- Apply the t – test for means;
- Applications and validity of Chi – square test.

Structure

19.1 Introduction

In this chapter, we consider inferential statistics involving the t-test for single mean, difference between the two means. Several inferential statistics are covered depending on whether the two samples are selected in an independent or dependent manner, and on whether the statistical assumptions are met. Two samples are dependent when the method of sample selection is such that those individuals selected for first sample do have a relationship to those individuals selected for second sample. In other words, the selections of individuals to be included in the two samples are related or correlated. The dependence condition leads us to consider the dependent samples t-test, alternatively known as the correlated samples t-test or the paired samples t-test.

Using chi-square distribution, along with sample data and frequency counts, we will be able to examine, whether a sample could have come from a given type of population distribution and whether the two qualitative variables could be independent of each other.

19. 2 t-test and conditions for the validity of t-test:

t-test: If $x_1, x_2, \ldots. x_n$ be the sample of size 'n' drawn from the normal population with mean 'μ' and variance 'σ^2' then the student's t-statistic is defined as

$$t = \frac{\bar{x} - \mu}{\sqrt{S^2/n}}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, is the sample mean and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, is an unbiased estimator of the population variance 'σ^2'.

Types of t-test: There are three types of t-tests.

(i) Single sample t-test (ii) Independent t-test (iii) Paired t-test

The single sample t-test, which is the most simple, determines whether there is any significant difference between population parameter (μ) and sample statistic ($\bar{x}$).

The independent t-test is used for comparing the means from two independent groups of individuals.

The paired t-test is used when comparing the means of two sets of observations from the same individuals or from pairs of individuals.

Conditions for the validity of t-test:

The t-test statistic can be used if the following conditions are exist:

(i) The observations are independent.
(ii) The parent population from which the sample is drawn is normal.
(iii) The population standard deviation(s) is (are) unknown.
(iv) The variables of interest are measured on at least an interval scale.
(v) The sample size(s) is (are) less than or equal to 30.

19.3 Applications of t-test:

Some of the applications of t-test are

(i) To test the sample mean ($\bar{x}$) is differ significantly from the hypothetical value (μ) of the population mean.
(ii) To test the significance of the difference between two sample means.
(iii) To test the significance of an observed sample correlation coefficient.
(iv) To test the significance of an observed partial correlation coefficient etc.

19.4 t-test for single mean

Suppose we want to test whether there is any significant difference between sample mean ($\bar{x}$) and population (μ) (or) the population mean has a specific value (say μ_0).

Step1: Null Hypothesis: There is no significant difference between sample mean ($\bar{x}$) and population (μ) (or) the population mean has a specific value (say μ_0).

i.e., $H_0 : \mu = \mu_0$

step2: Alternative Hypothesis: $H_1 : \mu \neq \mu_0$ (sometimes $\mu \geq \mu_0$ or $\mu \leq \mu_0$)

step3: If σ is unknown and the sample size is less than or equal to 30 then the required test statistic for testing the above null hypothesis is

$$t = \frac{\bar{x} - \mu_0}{\sqrt{S^2/n}} \sim t_{n-1} \qquad \text{(or)} \qquad t = \frac{\bar{x} - \mu_0}{\sqrt{S^2/n-1}} \sim t_{n-1}$$

which follows student t-distribution with (n-1) degrees of freedom.

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, n = sample size, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - x)^{-2}$ and $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - x)^{-2}$ is the sample variance.

Step4: Identify the tabulated value (critical value) of 't' for (n-1) degrees of freedom at α% level of significance.

Step5: We compare the computed value of 't' in step 4 with the significant value t_α at given level of significance, 'α'.

If calculated value of 't' or |t| is less than the tabulated value of 't' at α% level of significance we accept H_0 otherwise we reject H_0.

Note: The probable limits for the population mean are

$$\bar{x} \pm t_{\alpha/2}, (n-1)\sqrt{s^2/n-1}$$

19.5 t-test for two means:

Suppose we want to test whether there is any significant difference between two sample means ($\bar{x}$ - $\bar{y}$) (or) the population means are equal.

Working procedure:

Step1: Null Hypothesis: There is no significant difference between the two sample means or the population means are equal i.e., $H_0 : \mu_1 = \mu_2$

Step2: Alternative Hypothesis $H_1 : \mu_1 \neq \mu_2$ (sometimes $\mu_1 \geq \mu_2$ or $\mu_1 \leq \mu_2$)

Step3: If σ_1^2 and σ_2^2 are unknown and sample sizes or small then the required test statistic for testing the above null hypothesis is

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

where $\bar{x}$ = mean of the first sample n_1 = size of the first sample

$\bar{y}$ = mean of the second sample n2 = size of the second sample and

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum (x - \bar{x})^2 + \sum (y - \bar{y})^2\right] \quad \text{(or)}$$

If S_1^2 and S_2^2 are given then S^2 is obtained as $S^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

Step4: Identify the tabulated value (critical value) of 't' for (n_1+n_2-2) degrees of freedom at α% level of significance.

Step5: We compare the computed value of 't' in step4 with the significant value t_α at given level of significance, 'α'.

If calculated value of 't' or |t| is less than the tabulated value of 't' at α% level of significance we accept H_0 otherwise we reject H_0.

Note: The probable limits for the population means are

$$\bar{x} - \bar{y} \pm t_{\alpha/2},\ (n_1+n_2-2)\ \sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

19.6 Paired t-test

PAIRED t-TEST: The conditions to apply paired t-test are

(i) The sample sizes are equal i.e., $n_1 = n_2 = n$.

(ii) The two samples are not independent but the sample observations are paired together.

(iii) The two populations are normal.

Suppose a marketing researcher wants to know if the advertisement is really effective in promoting sales of a product or not. If $x_1, x_2, \ldots x_n$ be the sales of the product in 'n' departmental

stores for a certain period before advertisement campaign and $y_1, y_2, \ldots y_n$ be the corresponding sales of the same product in same 'n' departmental stores after advertisement campaign. Let $d_i = x_i - y_i$ (i=1,2,…n) be the difference in the observations for the i^{th} unit. Now, the null hypothesis is

H_0: The increments are just by chance and not due to the advertisement campaign (or) H_0: There is no significant difference between the sales of the product before and after advertisement campaign.

i.e., $H_0 : \mu_x = \mu y$

The required test statistic for testing the above hypothesis is

$t = \frac{\bar{d}\sqrt{n}}{S} \sim t_{n-1}$, which follows t-distribution with (n-1) degrees of freedom. where

$\bar{d} = \frac{\sum d}{n}$ and $S^2 = \frac{1}{n-1}\left[\sum d^2 - \frac{(\sum d)^2}{n}\right]$

Identify the tabulated value (critical value) of 't' for (n-1) degrees of freedom at α% level of significance.

If calculated value of 't' or |t| is less than the tabulated value of 't' for (n-1) at α% level of significance we accept H_0, otherwise we reject H_0.

19.7 Steps of Chi-square Test:

To determine the value of χ^2 and to draw conclusions the following steps are required:

1. Calculate the expected frequencies $E_1, E_2, \ldots E_n$ corresponding to observed frequencies $O_1, O_2, \ldots O_n$ under some theory of hypothesis.
2. Compute the deviations (O – E) for each frequency and then square these deviations to obtain $(O - E)^2$ $\frac{(O-E)^2}{E}$
3. Divide the square deviations i.e., $(O - E)^2$ by the corresponding expected frequency to obtain.
4. Obtain the sum of all values computed in the step (iii) to compute

$$\chi^2 = \sum \left[\frac{(O - E)^2}{E} \right]$$

This gives the value of χ^2, if it is zero if multiples that there is no discrepancy between the observed and expected frequencies. They coincide completely. The greater the value of χ^2, the greater will be discrepancy between the observed and expected frequencies.

5. Under the null Hypothesis the theory fits the data well, the above statistic follow χ^2 distribution with v (produced as nu) = n-1, degrees of freedom.
6. Compare the calculated value of χ^2 with table value, if it is less than the table value, the difference between theory and observation is not considered as significant. Such difference is regarded on account of sampling fluctuations and is ignored. On the other hand, if calculated value of χ^2 exceeds the table value, the difference between theory and observation is considered significant. In other words, the discrepancy between theory and observation cannot be attributed to chance and we reject the null Hypothesis and conclude that experiment does not support the theory.

19.8 Conditions for the validity of Chi-Square test:

The chi-square test statistic can be used only if the following conditions are satisfied:

1. N, the total number of frequencies should be reasonably large, say greater than 50.
2. The sample observations should be independent. This implies that no individual item should be included twice or more in the sample.
3. No theoretical cell frequency should be small. If expected frequencies are small, than the value of χ^2 would be over – estimated. This will result rejection of many null hypothesis. Small is a relative term. Preferably each theoretical frequency should be larger than 10, but in any case not less than 5. If any theoretical frequency is less than 5 then we cannot apply χ^2 test as such. In that case we use the technique of pooling which consists of adding the frequencies which are less than 5 with preceding or succeeding frequency, so that the resulting sum is greater than 5 and adjust the degree of freedom accordingly.
4. The given distribution should not be replaced by relative frequencies of proportions but data should be given in original units.
5. In contingency table (or) independent of attributes expected frequencies are obtained as

$$E = \frac{Row\ \text{Total} \times \text{Column total}}{\text{Grand Total}}$$

Degree of freedom = (r-1) (c-1)

R = no. of rows

C = No. of columns

19.9 Applications of chi-square Test:

1. Chi- Square test of goodness of gift
2. Chi-Square test for independence of attributes and
3. Chi-Square test as a test of homogeneity.

19.10 Chi – Square Test of goodness of fit:

Chi-Square test can be used to find out how well the theoretical distribution fit with the empirical distribution of observed distribution obtained from sample data. When Chi – Square test is used as a test of goodness of fit, you will be taking the following steps:

1. Set up null and alternative hypothesis
2. Decide level of significance for rejection of null hypothesis.
3. Draw a random sample of observations from the relevant population.
4. Derive theoretical distribution under the assumption that null hypothesis is true.
5. Compare observed frequencies with expected frequencies with the help of test.
6. If computed χ^2 is less than the table value at a certain level of significance, the fit is considered to be good. On the other hand, if the calculated value exceeds the table value, the fit is considered poor.

19.11 χ^2 – test FOR SINGLE POPULATION VARIANCE:

Suppose we want to test if a random sample of size 'n' drawn from the normal population with a specified $\sigma^2 = \sigma_0^2$ (say).

The null hypothesis is H_0: $\sigma^2 = \sigma_0^2$ (say)

The required test statistic for testing the above hypothesis is

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(x - \bar{x})^2}{\sigma_0^2} \right] = \frac{ns^2}{\sigma_0^2}$$, which follows Chi - square

Distribution with (n-1) degrees of freedom.

Conclusion: If the calculated value of χ^2 is less than tabulated of χ^2 for (n-1) degrees of freedom at 5% level of significance, we accept H_0, otherwise we reject H_0 otherwise we reject H_0

19.12 SOLVED PROBLEMS

1. A soap manufacturing company was distributing a particular brand of soap through a large number of retail shops. Before a heavy advertisement campaign, the mean sales was found to be 147 dozens with standard deviation 16. Can you consider the advertisement effective?

Sol: We are given: n=26, $\bar{x}$ = 147 dozens, s = 16 dozens Null Hypothesis. H_0 : μ = 140 dozens, i.e., the deviation between $\bar{x}$ and μ is just due to fluctuations of sampling. In other words, advertisement is not effective.

Alternative hypothesis. H_1: μ > 140 (Right –tail)

Test statistic: under the null hypothesis H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \sim t_{n-1} = t_{25}$$

$$t = \frac{147 - 140}{16/\sqrt{25}} = \frac{7 \times 5}{160} = 2.19$$

Tabulated value of t for 25 d.f. at 5% level of significance for single (right) tail test is 1.798 i.e: t_{25} (0.05) = 1.708. Since calculated value of t is greater than the tabulated value, it is significant and we reject H_0 at 5% level of significance, hence, the increase in sales can not be attributed to

fluctuations of sampling and we conclude that advertisement is certainly effective in increasing the sales.

2. Certain pesticide is packed into bags by a machine. A random sample of 10 bags is drawn and their contents are found to weigh (in kg.) as follows:

50, 49, 52, 44, 45, 48, 46, 45, 49, 45

Test is the average packing can be taken to be 50 kg.

Sol: Null Hypothesis, H_0: $\mu \neq 50$ kgs. i.e., the average packing is 50 kgs. Alternative Hypothesis, H_1: $\mu \neq 50$ kgs (two tailed)

Calculations for sample mean and S. D.

X	50	49	52	44	45	48	46	45	49	45	Total
d = x - 48	2	1	4	-4	-3	0	-2	-3	1	-3	-7
d2	4	1	16	16	9	0	4	9	1	9	69

$\bar{x}$ = A + $\frac{\sum d}{n}$ = 48 + $\frac{-7}{10}$ = 48 -0.7 = 47.3

$S^2 = \frac{1}{n-1}\left(\sum d^2 - \frac{(\sum d)^2}{n}\right) = \frac{1}{9}\left(69 - \frac{(-7)^2}{10}\right) = \left(\frac{69 - 4.9}{9}\right) = 7.12$

Under H_0, the test statistic is:

$t = \left(\frac{\bar{x} - \mu}{\sqrt{s^2/n}}\right) = \left(\frac{47.3 - 50.0}{\sqrt{7.12/10}}\right) = \left(\frac{-2.7}{\frac{0.712}{\sqrt{}}}\right) = \frac{-2.7}{0.8438} = -3.2$

Which follows student's t - distribution with 10-1= 9 df.

Tabulated $t_{0.05}$ for 9 d.f = 2.262. Since calculated $|t|$ is greater than tabulated t. it is significance. Hence H_0 is rejected at 5% level of significance and we conclude that the average packing cannot be taken to be 50 kgs.

3. A group of 5 patients treated with medicine ‘A’ weight 42, 39,48,60 and 41 kgs. Seconds group of 7 patients from the same hospital treated with medicine ‘B’ weight 38, 42, 56, 64, 68, 69 and 62 kgs. Do you agree with the claim that medicine ‘B’ increases the weight significantly? (The value of ‘t’ at 5% level of significance for 10 degrees of freedom is 2.228)

Solution: Let the weight (in kgs) of the patients treated with medicines A and B be denoted variables X and Y respectively.

Null Hypothesis: H_0: $\mu_x = \mu_y$ i.e., there is no significant difference between the medicines A and B as regard their effect on increase in weight.

Alternative Hypothesis: H_1 : $\mu_x < \mu_y$ (left – tailed) i.e., medicine B increases the weight significantly.

Computation of sample means and S.D’S

Medicine A			Medicine B		
X	$x - \bar{x} = x - 46$	$(x - \bar{x})^2$	y	$y - \bar{y} = y - 57$	$(y - \bar{y})^2$
42	-4	16	38	-19	361
39	-7	49	42	-15	225
48	2	4	56	-1	1
60	14	196	64	7	49
41	-5	25	68	11	121
			69	12	144
			62	5	25
Total 230	0	290	Total 399	0	926

Under null Hypothesis H_0, the test statistic is:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} = t_{10}$$

Here $n_1 = 5$, $\Sigma x = 230$, $\sum(x - \bar{x})^2 = 290$, $n_2 = 7$, $\Sigma y = 399$, $\sum(y - \bar{y})^2 = 926$

$$\bar{x} = \frac{\sum x}{n_1} = \frac{230}{5} = 46; \qquad \bar{y} = \frac{\sum y}{n_2} = \frac{399}{7} = 57$$

and $S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum (x - \bar{x})^2 + \sum (y - \bar{y})^2\right] = \left(\frac{290 + 926}{10}\right) = 121.6$

$$t = \frac{46 - 57}{\sqrt{121.6 \times (\frac{1}{5} + \frac{1}{7})}} = \frac{-11}{\sqrt{121.6 \times \frac{12}{35}}} = \frac{-11}{\sqrt{41.69}} = \frac{-11}{6.457} = -1.7$$

Tabulated value of t for 10 d.f. at 5% level of significance for the left tailed test is -1.81.

Since calculated |t| is less than tabulated value of t, it is not significant. Hence, H_0 may be accepted and we may conclude that the medicines A and B do not differ significantly as regards their effect on increases in weight.

4. A random sample of 20 daily workers of state A was found to have average daily earning of Rs.44 with sample variance 900. Another sample of 20 daily workers from state B was found to earn on an average Rs.30 per day with sample variance 400. Test whether the workers in state A are earning more than those in state B.

Sol: Let the daily earnings (in Rs.) of workers in the states A and B be denoted by the variables x and y respectively. Then we are given:

Here $n_1 = 20$, $\bar{x} = 44$, $S_x^2 = 900$, $n_2 = 20$, $\bar{y} = 30$, $S_y^2 = 400$

Null Hypothesis, H_0: $\mu_x = \mu_y$ i.e., there is no significant difference in the average daily earnings of the workers in state A and B. Alternative Hypothesis. H_1: $\mu_x > \mu_y$ (Right – tailed).

Test statistic. Under the test – statistic is:

$$t = \frac{x - y}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} = t_{38}$$

where

$$S^2 = \left(\frac{n_1 S_x^2 + n_2 S_y^2}{n_1 + n_2 - 2}\right) = \left(\frac{20x900 + 20x400}{38}\right) = \left(\frac{18000 + 8000}{38}\right) = \frac{26000}{38} = 648.21$$

$$t = \frac{44 - 30}{\sqrt{648.21\left(\frac{1}{20} + \frac{1}{20}\right)}} = \frac{14}{\sqrt{648.21 \times \frac{1}{10}}} = \frac{14}{\sqrt{64.821}} = \frac{14}{8.0511} = 1.7389$$

Tabulated $t_{0.05}$ for d.f. = $n_1+n_2-2 = 38$, for tight tailed test is 1.645. (For d.f. > 30, the significant values of t are same as those of z for the normal test).

Since calculated t is greater than the tabulated t, it is significant at 5% level of significance. Hence H_0 is rejected (H_1 is accepted) at 5% level of significance and we conclude with 95% confidence that the workers in the state A are earning more than those in state B.

5. The means of two random samples of size 9 and 7 are 196.42 and 198.82 respectively. The sum of squares of the deviations from the mean are 26.94 and 18.73 respectively. Can the samples be considered to have been drawn from the same normal population? (Apply t – test)

Sol: We are given:

Here $n_1 = 9$, $\bar{x} = 196.42$, $\sum(x - \bar{x})^2 = 26.94$; and $n_2 = 7$, $\bar{y} = 198.82$, $\sum(y - \bar{y})^2 = 18.73$

Null Hypothesis: The samples have been drawn from the same normal population i.e., H_0: $\mu_x = \mu_y$

Alternative Hypothesis H_1: $\mu_x \neq \mu_y$ (Two – tailed)

Under the null hypothesis H_0, the test statistic is

$$t = \frac{x - y}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} = t_{14}$$

We have

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum (x - \bar{x})^2 + \sum (y - \bar{y})^2\right] = \left(\frac{26.94 + 18.73}{9 + 7 - 2}\right) = \left(\frac{45.67}{14}\right) 3.26$$

$$\therefore t = \frac{196.42 - 198.82}{\sqrt{3.26\left(\frac{1}{9} + \frac{1}{7}\right)}} = \frac{-2.40}{\sqrt{3.26 \times 0.254}} = \frac{-2.40}{\sqrt{0.828}} = \frac{-2.40}{0.9099} = -2.64$$

Tabulated value for $t_{0.05}$ 14 d.f. for two tailed test is 2.15. Since calculated |t| is greater than tabulated t, it is significant. Hence H_0 is rejected and, therefore, the samples be considered to have come from the same normal population.

6. The sales data of an item in six shops before and after a special promotional campaign are as under:

Shops	A	B	C	D	E	F
Before campaign	53	28	31	48	50	42
After campaign	58	29	30	55	56	45

Can the campaign be judged to be a success? Test at 5% level of significance.

Sol: Here the sales data before campaign (x) and after campaign (y) are not independent but paired together for 6 shops A to F. Hence we shall apply paired t – test.

Null Hypothesis. H_0: $\mu_x = \mu_y$ i.e., the average sales before campaign and after campaign are same. In other words, there is no significant change in sales after the special promotional campaign.

Alternative Hypothesis: H_1: $\mu_x < \mu_y$ (Left – tailed) i.e., the special promotional campaign increase the sales.

Test statistic. Under H_0, the test statistic is

$$t = \frac{\bar{d}}{S/\sqrt{n}} = \frac{\bar{d}}{\sqrt{S^2/n}} \sim t_{n-1} = t_5$$

Shop	A	B	C	D	E	F	Total
X	53	28	31	48	50	42	
Y	58	29	30	55	56	45	
d = x - y	-5	-1	1	-7	-6	-3	$\Sigma d = -21$
d^2	25	1	1	49	36	¤ 9	$\Sigma d^2 = 121$

$$\therefore \bar{d} = \frac{\sum d}{n} = \frac{-21}{6} = -3.5$$

$$S^2 = \frac{1}{n-1}\left[\sum d^2 - \frac{(\sum d)^2}{n}\right] = \frac{1}{5}\left[121 - \frac{441}{6}\right] = 9.5$$

$$\therefore t = \frac{-3.5}{\sqrt{9.5/6}} = \frac{-3.5}{\sqrt{1.2583}} = -2.78$$

The tabulated value of t for 5 d.f. at 5% level of significance for a single tailed test is 2.02 i.e., $t_{5(0.05)} = 2.02$

Since |t| = 2.78 is greater than 2.02, it is significant. Hence H_0 is rejected at 5% level of significance and we conclude that the special promotional campaign has been effective in increasing sales.

7) In a certain sample of 2,000 families1,400 families are consumers of tea. Out of 1,800 Hindu families, 1,236 families consume tea. Use χ^2-test and state whether there is any significant difference between consumption of tea among Hindu and non-Hindu families.

Sol: - The above data can be arranged in the form of a 2×2

Contingency table as given below:

Number of ↓	Hindu	Non-Hindu	Total
Families consuming tea	1236	164	1400
Families not consuming tea	564	36	600
Total	1800	200	2000

We set up the null hypothesis that the two attributes viz., 'consumption of tea' and the 'community' are independent. In other words, there is no significant difference between the consumption of tea among Hindu and non - Hindu families.

Under the null hypothesis of independence, $E(1236) = \frac{1800 \times 1400}{2000} = 1260$

The complete table of expected frequencies is given in table 18.21.

TABLE18.21 : Expected frequencies

1260	1400-1260 = 140	1400
1800-1260 = 540	200-140 = 60	600
1800	200	200

TABLE 18.22 : Computation of χ^2

o	E	o-E	$(o-E)^2$
1236	1260	-24	576
564	540	24	576
164	140	24	576
36	60	-24	576

CHI – SQUARE TEST

$$\therefore\ \chi^2 = \sum\left[\frac{(0-E)^2}{E}\right] = 576\left[\frac{1}{1260} + \frac{1}{540} + \frac{1}{140} + \frac{1}{60}\right]$$

= 576 [0.000794 + 0.001852 + 0.007143 + 0.096667]

= 576 × 0.026 456

= 15.2387.

d.f. = (2-1) (2-1) = 1 ; Tabulated $\chi^2_{0.05}$ for 1 d.f. = 3.841.

Since the calculated value of χ^2, viz., 15.24 is much greater than the tabulated value of χ^2 at 5% level of significance, it is highly significant and the null hypothesis is rejected at 5% level of significance. Hence we conclude with 95% confidence that the two communities (Hindus and Non - Hindus) differ significantly as regards the consumption of tea among them.

8. In a survey of 200 boys of which 75 were intelligent, 40 had educated fathers, while 85 of the unintelligent boys had uneducated fathers. Do these figures support the hypothesis that educated fathers have intelligent boys? (value of χ^2 for 1 d.f. is 3.841)

Sol : - The given data can be arranged in the form of a 2 × 2 contingency table as given below:

	Intelligent boys	Un-intelligent boys	Total
Educated fathers	40	125-85 = 40	40+40 = 80
Un-educated fathers	120-85 = 35	85	200-80 = 120
Total	75	200-75 = 125	200

Note. Bold figures are the given values

Null Hypothesis. We set up the null hypothesis that the two attributes viz., 'education of fathers' and 'intelligence of boys', are independent. In other words, educated fathers do not have intelligent boys i.e., the education of fathers does not have any effect on the intelligence of the boys.

Under the null hypothesis of independence, various expected frequencies are calculated as follows.

$$E(40) = \frac{75 \times 80}{200} = 30 \qquad : \qquad E(35) = \frac{75 \times 120}{200} = 45$$

$E(40) = \frac{125 \times 80}{200} = 50$: $E(85) = \frac{125 \times 120}{200} = 75$

$\chi^2 = E\left[(0-E)^2 / E\right]$

= 3.33 + 2.22 +200 + 1.33

= 8.88

O	E	O -E	$(O-E)^2$	$(O-E)^2/E$
40	30	10	100	3.33
35	45	-10	100	2.33
40	50	-10	100	2.00
85	75	10	100	1.33

d.f. = (2-1) (2-1) = 1

Tabulated $\chi^2_{0.05}$ for 1 df = 3.841

Since the calculated value of χ^2 viz., 8.88 is greater than the tabulated value of χ^2 at 5% level of significance, it is significant and the null hypothesis is rejected. Hence we conclude that the education of the fathers has a significant effect on the intelligence of the boys.

9. A sample of 400 students of under – graduate and 400 students of post – graduate classes was taken to know their opinion about autonomous colleges. 290 of the under – graduate and 310 of the post-graduate students favored the autonomous status. Present these facts in the form of a table and test, at 5% level, that the opinion regarding autonomous status of colleges is independent of the level of classes of students. (Table value of χ^2 at 5% level is 3.84 for 1 d.f.).

Sol : - Null Hypothesis. H_0 : Option about autonomous colleges is independent of the level of classes.

The given information can be presented in a 2×2 contingency as given in Table as :

Table : observed frequencies

Class	Number of students		Total
	Favoring	Opposing	
Under graduate	290	400-29 = 110	400
Post graduate	310	400 – 310 = 90	400
Total	290 + 310 = 600	110 + 90 = 200	400+400 = 800

Note: The figures in the bold are the given figures. Under the null hypothesis, the expected frequencies are calculated as follows:

$$E(290)=\frac{600\times 400}{800}=300 \quad : \quad E(310)=\frac{600\times 400}{800}=300$$

$$E(110)=\frac{200\times 400}{800}=100 \quad : \quad E(90)=\frac{200\times 400}{800}=100$$

Test statistic:

$$\chi^2=\sum\left[\frac{(0-E)^2}{E}\right]=\frac{(290-300)^2}{300}+\frac{(310+300)^2}{300}+\frac{(110-100)^2}{100}+\frac{(90-100)^2}{100}$$

= 0.33 + 0.33 + 1.00 +1.00 = 2.66

d.f. = (2 -1) (2-1) = 1

The critical value (tabulated value) of Chi-square for 1 d.f. at 5% level of significance is 3.84.

Since the calculated value of chi-square (2.66) is less than the tabulated value (3.84), it is not significant.

Hence, we fail to reject H_o and conclude that the option about autonomous colleges may be regarded to be independent of the level of classes of the students.

10. Two researchers adopted different sampling techniques while investigating the same group of students to find the number if students falling in different intelligence levels. The results are as follows:

Research	No. of students in each level				Total
	Below Average	Average	Above Average	Genius	
X	86	60	44	10	200
Y	40	33	25	2	100
Total	126	93	69	12	300

Would you bay that the sampling techniques adopted by the two researchers are significantly different?

(Given 5% values of χ^2 for 3 d.f. and 4 d.f. are 7.82 and 9.49 respectively).

Sol: - We set up the null hypothesis H_0 that the data obtained are independent of the sampling techniques adopted by the two researchers. In other words, the null hypothesis is that there is no significant difference between the sampling techniques used by the two researchers for collecting the required data.

Here we have a 4×2 contingency table and d.f. = $(4 -1) \times (2 - 1) = 3 \times 1 = 3$. Hence we need to compute only 3 independent expected frequencies and the remaining expected frequencies can be obtained by subtraction from the marginal totals.

Under the null hypothesis of independence we have :

$$E(86) = \frac{126 \times 200}{300} = 84 \quad ; \quad E(60) = \frac{93 \times 200}{300} = 62 \quad ; \quad E(44) = \frac{69 \times 200}{300} = 46.$$

The table of expected frequencies can now be completed, as given in the following table.

Table: Expected frequencies

Research	No. of students in each level				Total
	Below Average	Average	Above Average	Genius	
X	84	62	46	200-192=8	200
Y	126-84=42	93-62=31	69-46=23	12-8=4	100
Total	126	93	69	12	300

Table: Computation of Chi-square

O	E	(O-E)	$(O-E)^2$	$(O-E)^2/E$
86	84	2	4	0.048
60	62	-2	4	0.064
44	46	-2	4	0.087
10	8	2	4	0.500
40	42	-2	4	0.095
33	31	2	4	0.129
25	23	2	4	0.174
2	4	-2	4	1.000

Total 300	300	0		2.097

$$\therefore \chi^2 = \sum\left[\frac{(0-E)^2}{E}\right] = 2.097$$

d.f. = (4 -1) × (2-1) = 3×1 = 3 and $\chi^2_{0.05}$ for 3.df = 7.82

Since the calculated value of χ^2 is less than the tabulated value, it is not significant. Hence, null hypothesis may be accepted at 5% level of significance and we may conclude that the sampling techniques adopted by the two investigators do not differ significantly.

Important Remark.

In fact in this question we should have also been given the tabulated value of χ^2 for 2 d.f. at 5% level of significance. The computation of χ^2 as given in the above table is wrong since we can not apply the χ^2 -test as the last expected frequency is less than 5. We should use the technique of pooling in this case as given in Table below

Table : Computation of Chi-Square

O	E	(O-E)	$(O-E)^2$	$(O-E)^2/E$
86	84	2	4	0.048
60	62	-2	4	0.064
44	46	-2	4	0.087
10	8	2	4	0.500
40	42	2	4	0.095
33	31	-2	4	0.129
27	27	0	0	0
Total: 300	300	0		0.923

Chi-Square Test is

After pooling, $$\chi^2 = \sum\left\lfloor\frac{(0-E)^2}{E}\right\rfloor = 0.923.$$

and the df. = [(4-1) × (2-1)] -1 = 3-1 = 2, since 1 d.f. is lost in the method of pooling the theoretical cell frequencies which are less than 5.

Tabulated value of χ^2 for 2 d.f. at 5% level of significance is 5.991. Since calculated value is less than the tabulated value. Null hypothesis may be a accepted at 5% level of significance and we may conclude that there is no significance difference in the sampling techniques used by the researchers.

11. The following table gives for a sample of married women, the level of education and marriage adjustment score:

		Marriage Adjustment Score			
		Very low	low	High	Very High
Level of Education	College	24	97	62	58
	High school	22	28	30	41
	Middle school	32	10	11	20

Can you conclude from the above, the higher the level of education, the greater is the degree of adjustment in marriage?

Sol: - We set up the null Hypothesis H_0 : the level of education is independent of 't' The degree of adjustment in marriage'. Since we are given data in the form of a 3×4 contingency table, the d.f. = $(3-1) \times (4-1) = 2 \times 3 = 6$. Hence it will suffice to compute 6 (independent) expected frequencies and the remaining expected frequencies can be obtained easily by subtraction from marginal row and column totals.

Under the null hypothesis of independence, the expected frequencies are obtained below.

$$E(24) = \frac{121 \times 78}{435} = 43.21 \quad ; \quad E(97) = \frac{241 \times 135}{435} = 74.79 \quad ; \quad E(62) = \frac{241 \times 103}{435} = 57.06;$$

$$E(22) = \frac{121 \times 78}{435} = 21.69 \quad ; \quad E(28) = \frac{121 \times 135}{435} = 37.55 \; ; \quad E(30) = \frac{121 \times 103}{435} = 28.65$$

The corresponding expected frequencies are given in Table below:

Table corresponding expected frequencies

			Marriage Adjustment Score		
Level of Education	Very low	low	High	Very High	Total

College	43.21	74.79	57.06	241-175.06=65.94	241
High school	21.69	37.55	28.65	121-87.89=33.11	121
Middle school	78-64.90=13.10	135-112.34=22.66	103-58.71=17.29	73-53.05=19.95	73
Total	78	135	103	119	435

Under H_0, the test statistic is:

$$\chi^2 = \sum \left| \frac{(0-E)^2}{E} \right|$$

= 8.54 + 6.60 + 0.42 + 0.95 + 0.00 + 2.43 + 0.06 + 1.88 + 27.27 + 7.07 +2.29 + 0.00

(Row wise in order)

= 57.51

Df = (3-1) × (4-1) = 2 × 3 = 6; $\chi^2_{0.05}$ for 6 df. = 12.59

Since calculated value of χ^2 is much greater than tabulated value, it is highly significant at 5% level of significance. Hence, we reject the null hypothesis at % level of significance. Hence, we reject the null hypothesis at 5% level of significance and conclude that 'the higher the level of education, the greater is the degree of adjustment in marriage.'

12.

1) The following figures show the distribution of digits in numbers chosen at random from a telephone directory:

Digit:	0	1	2	3	4	5	6	7	8	9	total
Frequency:	1,026	1,107	997	996	1,075	933	1,107	972	964	853	10,000

Test whether the digits may be taken to occur equally frequently in the directory. (The table value of χ^2 for 9 dF at 5% level of significance is 16.92)

Sol: Null Hypothesis. Set up the null hypothesis that the digits 0, 1, 2, 3, …. 9 in the numbers in the telephone directory are uniformly distributed, i.e., all the digits occur equally frequently in the directory.

Then under the null hypothesis, the expected frequency for each of the digits 0, 1, 2, 3, …. 9 is

$10{,}000 \div 10 = 1000$

Computation of χ^2

Digits	Observed frequency (O)	Expected frequency (E)	(O – E)	$(O - E)^2$	$(O - E)^2/E$
0	1,026	1000	26	676	0.676
1	1,107	1000	107	11,449	11.449
2	997	1000	-3	9	0.009
3	996	1000	34	1,156	1.156
4	1,075	1000	-75	5,625	5.625
5	933	1000	-67	4,489	4.489
6	1,107	1000	107	11,449	11.449
7	972	1000	-28	784	0.784
8	964	1000	-36	1,296	1.296
9	853	1000	-147	21,609	21.609
	10,000	10,000	0		58.542

$$\therefore \chi^2 = \sum \left[\frac{(O - E)^2}{E} \right]$$

$= 58.542$

Since we are given 10 frequencies subjected to only one linear constraint $\Sigma O = \Sigma E = 10{,}000$

$df = 10 - 1$

$= 9$

Tabulated value of χ^2 for 9 df at 5% level of significance is 16.919. Since calculated value of $\chi^2 = 58.542$ is much greater than the tabulated value 16.919, it is highly significant and null hypothesis is rejected at 5% L.O.S. Hence, we conclude that the digits 0, 1, 2, ….. 9 cannot be regarded to the distributed uniformly in the numbers in the directory.

13. The numbers of scooter accidents per month in a certain town were as follows:

12, 8, 20, 2, 14, 10, 15, 6, 9, 4

Are these frequencies in agreement with the belief that accident conditions were the same during this 10 month period?

Sol: Null Hypothesis: Set up the null hypothesis that the given frequencies (of number of accidents per month in a certain town) are consistent with the belief that the accident conditions were same during the 10 – month period.

Since the total number of accidents over the 10 months are:

12+8+20+2+14+10+15+6+9+4 = 100

Under the null hypothesis, these accidents should be uniformly distributed over the 10 – month period and hence the expected number of accidents for each 10 months are (100/10) = 10

Computation of χ^2

Digits	Observed no. of accidents (O)	Expected no. of accidents (E)	(O – E)	$(O – E)^2$	$(O – E)^2/E$
1	12	10	2	4	0.4
2	8	10	-2	4	0.4
3	20	10	10	100	10.0
4	2	10	-8	64	6.4
5	14	10	4	16	1.6
6	10	10	0	0	0.0
7	15	10	5	25	2.5
8	6	10	-4	16	1.6
9	9	10	-1	1	0.1
10	4	10	-6	36	3.6
	100	100	0		26.6

$$\therefore \chi^2 = \sum \left[\frac{(O - E)^2}{E} \right]$$

$= 26.6$

$df = 10 - 1$

$= 9$

Tabulated value of $\chi^2_{0.05}$ for 9 df = 16.919

Since the calculated value of $\chi^2 = 26.6$ is greater than the tabulated value of, viz., 16.919, it is significant and the null hypothesis is rejected at 5% level of significance. Hence we conclude that the accident conditions are certainly not uniform (same) over the 10 – month period.

14. Records taken of the number of male and female births in 800 families having four children are given

Number of births		**Frequency**
Male	**Female**	
0	4	32
1	3	178
2	2	290
3	1	236
4	0	64

Test whether the data are consistent with the hypothesis that the binomial law holds and the chance of a male birth is equal to that of a female birth.

Sol: Let us set up the null hypothesis that the data are consistent with the binomial law of equal probability for male and female births, so that under the null hypothesis we have p = q = ½, where p denotes the probability of a male birth.

THEORETICAL BINOMIAL FREQUENCIES

No. of male births (r)	**Expected frequency f(r)**
0	$50 \times {}^4c_0 = 50 \times 1 = 50$
1	$50 \times {}^4c_1 = 50 \times 4 = 200$
2	$50 \times {}^4c_2 = 50 \times 6 = 300$
3	$50 \times {}^4c_3 = 50 \times 4 = 200$
4	$50 \times {}^4c_4 = 50 \times 1 = 50$
Total	800
	N

In the usual notations we are given n = 4, N = 800

According to binomial probability law, the frequency of r male births is given by

$$f(r) = N.p(r) = N \times {}^nC_r\, p^r\, q^{n-r}$$

$$= 800 \times {}^4C_r\, (1/2)^r\, (1/2)^{4-r}$$

$$= 800 \times {}^4C_r\, (1/2)^4$$

$$= 50 \times {}^4C_r \text{ i } (r = 0, 1, 2, 3, 4) \text{ ---------- (*)}$$

Thus, substituting r = 0, 1, 2, 3, 4 successively in (*), we get the theoretical (Binomial) frequencies as given in the table.

TESTING GOODNESS OF FIT

No. of male births	Observed frequency (O)	Expected frequency (E)	(O – E)	$(O - E)^2$	$(O - E)^2/E$
0	32	50	-18	324	6.48
1	178	200	-22	484	2.42
2	290	300	-10	100	0.33

3	236	200	36	1296	6.48
4	64	50	14	196	3.92
Total	800	800	0		19.63

$$\chi^2 = \sum \left[\frac{(O - E)^2}{E} \right]$$

= 19.63

Here we are given 5 frequencies.

Hence the degrees of freedom are: d.f = 5 – 1 = 4

One d.f is reduced because of the linear constraint $\Sigma O = \Sigma E = 800$

Tabulated (critical) value of χ^2 for 9d.f at 5% level of significance is 9.488

Since, calculated value of $\chi^2 = 19.63$ is greater than the tabulated value, it is significant. Thus, the difference between observed and expected frequencies is significant and cannot be attributed to chance fluctuations. Hence, we reject the null hypothesis s at 5% level of significance and conclude that the hypothesis of male and female births is wrong. Hence, the binomial distribution with p = q = 1/2, is not a good fit to the given data.

15. The following mistakes per page were observed in a book.

No. of mistakes per page	0	1	2	3	4	total
No. of pages	211	90	19	5	0	325

Fit a Poisson distribution and test the goodness of fit.

Sol: If the random variable x denotes the number of mistakes per page then the given distribution is given in table.

Computation of mean

x	0	1	2	3	4	total
f	211	90	19	5	0	$N = \Sigma f = 325$
fx	0	90	38	15	0	$\Sigma fx = 143$

Mean of the given distribution of mistakes per page is:

$$\bar{x} = \frac{\sum fx}{N} = \frac{143}{325}$$

$= 0.44$

In order to fit a Poisson distribution to the given data, we take the mean (parameter) ^ of the Poisson distribution equal to the mean of the given distribution i.e., we take

$m = \bar{x} = 0.44$

The frequency of r mistakes per page is given by the Poisson law as:

$$f(r) = N.p(r) = 325 \times \frac{e^{-0.44}(0.44)^r}{r!} \; ; r = 0, 1, \ldots.4$$

$f(0) = 325 \times e^{-0.44} = 325 \times \text{Antilog}\,[-0.44 \log_{10} e]$

$= 325 \times \text{Antilog}\,[-0.44 \log_{10} 2.7183]$ [□e = 2.7183]

$= 325 \times \text{Antilog}\,[-0.44 \times 0.4343]$

$= 325 \times \text{Antilog}\,[-0.1911]$

$= 325 \times \text{Antilog}\,[\bar{1}.8089]$

$= 325 \times 0.6440$

$= 209.30$

$f(1) = m \times f(0) = 0.44 \times 209.3 = 92.092$

$f(2) = \frac{m}{2} \times f(1) = 0.22 \times 92.092 = 20.36$

$f(3) = \frac{m}{3} \times f(2) = \frac{0.44}{2} \times 20.26 = 2.97$

$f(4) = \frac{m}{2} \times f(3) = 0.11 \times 2.97 = 0.3267$

Hence the theoretical Poisson frequencies correct to one decimal place are as given below:

x	0	1	2	3	4	Total
Expected frequency	209.3	92.1	20.3	3.0	0.3	325

Testing goodness of fit

Observed frequency (O)	Expected frequency (E)	(O – E)	$(O – E)^2$	$(O – E)^2/E$
211	209.3	1.7	2.89	0.01381
90	92.1	-2.1	4.41	0.04788
19	20.3	0.4	0.16	0.00678
5 (24)	3.0 (23.6)			
0	0.3			
325	325.0			0.06847

$$\chi^2 = \sum \left[\frac{(O - E)^2}{E} \right]$$

= 0.06847

Degrees of freedom [c.f., δ18.7] we have n = 5

Required df are:

v = 5 – 1 – 1 – 2

= 1

One df begin lost because of the linear constraint $\Sigma 0 = \Sigma E$; 1 df is lost because the parameter m has been estimated from the given data and is then used for computing the expected frequencies; 2 df are lost because of pooling the last three cell frequencies which are less than five.

Tabulated value of χ^2 for 1 df at 5% level of significance is 3.841. Since calculated value of χ^2, viz., 0.068 is less than 3.841, it is not significant. Thus we conclude that the difference between observed and theoretical frequencies is just due to chance and the Poisson distribution is a good fit to the given data.

19. 13 Summary

If the sample size is small i.e., below 30 and independent and population standard deviation is unknown, t – statistic can be used to test the hypotheses for the difference two population means. This technique is based on the assumption that the characteristic being

studied is normality distributed for both the populations. The t – test can also be applied for dependent samples or related samples, which is also called paired t – test. In this test the observations in sample I are related to the observations in sample II. Chi – square test for independent of attributes and goodness of fit are powerful methods explained to deal qualitative data. In Chi – square test for goodness of fit, various types of distributions like uniform, Binomial and Poisson distributions are fitted for different types of real data experiments.

19.14 Self assessment problems

1. The mean life time of a sample of 25 bulbs is found to be 1550 hours with the standard deviation of 120 hours. The company manufacturing the bulbs claims that the average life of their bulbs is 1600 hours. Is this claim is accepted at 5% level of significance.

(Ans: $|t| = 2.04$, H_0 is rejected)

2. A random sample of 16 values from a normal population showed a mean of 103.75 and sum of squares of deviations from the mean is 843.75. Test the assumptions of mean for the population is 108.75 is reasonable.

(Ans: $t = -2.67$, H_0 is rejected)

3. The mean height and the standard deviation of 8 randomly chosen soldiers are 166.9 and 8.29 cm respectively. The corresponding values of 6 randomly chosen sailors are 170.3 and 8.50 cm respectively. Based on this information, can we conclude that soldiers are shorter than sailors? (Ans: $|t| = 0.695$, H_0 is accepted)

4. Explain the t – test to test significant difference between two groups.

5. Define t – statistic. State the applications of t – test.

6. Explain about the paired t – test.

7. Define χ^2 – test. State the applications of Chi – square test.

8. Explain about the Chi – square test for goodness of fit.

9. Eleven school boys were given a test in statistics. They were given a coaching and a second test was conducted. Test whether the students have benefited by the extra coaching.

Boys	1	2	3	4	5	6	7	8	9	10	11
Marks in I test	23	20	19	21	18	20	18	17	23	16	19
Marks in II test	24	19	22	18	20	22	20	20	23	20	18

(Ans: $t = 1.483$, H_0 is accepted)

10. Explain about the procedure of t – test for single mean.

11. The following table gives the number of accidents occurred during the 6 days of a week. Find whether the accidents are uniformly distributed over the week.

Days	MON	TUE	WED	THUR	FRI	SAT
No. of accidents	14	18	12	11	15	14

(Ans: $\chi^2 = 2.143$, H_0 is accepted)

12. Fit a Poisson distribution and test the goodness of fit.

x	0	1	2	3	4
f	123	59	14	3	1

(Ans: $\chi^2 = 0.99$, H_0 is accepted)

13. The following data was collected on literacy and smoking.

	Smokers	Non – smokers
Literates	141	312
Illiterates	192	72

Based on the above data, can we conclude that smoking depends on literacy.

(Ans: $\chi^2 = 113.6$, H_0 is rejected.

19.15 Reference Books:

1. S. C. Gupta, Fundamentals of Statistics, Himalaya Publishing House.

2. Digambar Patri., D. N. Patri, Quantitative Techniques, Kalyani publications.

3. P. N. Arora and S. Arora, Statistics for Management: S. Chand & Comp. Ltd.

4. G. V. Shenoy, Uma K. Srivastava, S. C. Sharma: Business Statistics

5. B. M. Agarwal: Business statistics

6. Gupta S. P.: Statistical Methods

Lesson Writer

Dr. J. Pratapa Reddy

20. Hypothesis testing – F test

Objectives

After completion of this chapter, you should be able to:

- Understanding the assumptions of Small samples test;
- Discuss the applications of F – test;
- Know the meaning of two types of errors.

Structure

20.1 Introduction

As indicated in previous chapters, one of the assumptions made in test concerning difference between means when the size of the samples are small in that both the populations from which samples are drawn have equal variances. If the population variances are not equal, the validity of such a test is subject to equation. Accordingly before using the t-test for judging whether the population means of the two groups are equal, we should consider whether the two population variances are equal or not. This is called a hypothesis test for equality of the two population variances or whether there is any significant difference between the two population variances. Obviously, such a test has to be two-tailed due to its vary nature.

The inductive inference consists in arriving at a decision to accept or reject a null hypothesis after inspection only based on the sample from it. As such, an element of risk – the risk of taking wrong decisions involved. These risks in testing of hypothesis are called type-I and Type-II errors.

20.2 Two types of errors:

The main objective of the sampling theory is to draw the valid conclusions or inferences about the population parameters on the basis of the sample results. The conclusions drawn on the basis of a particular sample may not always be true in respect of population. The four possible situations that arise in any test procedure are shown in the following table:

True statement	Decision from sample	
	Reject H_0	Accept H_0
H_0 is true	Wrong (Type I error)	Correct
H_0 is false (H_1 is true)	Correct	Wrong (Type II error)

From the above table,

Type I error :- Rejecting H_0 , when H_0 true.

Type II error : - Accepting H_0 , when H_0 false.

20.3 Level of significance, confidence coefficient and power of the test:

From the above table,

α = Level of significance or size of critical region

= Probability of type I error = Prob {Rejecting H_0, when H_0 is true}

β = Probability of type II error = Prob {Accepting H_0, when H_0 is false}

1-α = Confidence coefficient

= Prob {Accepting H_0, when H_0 is true}

1-β =Prob {Rejecting H_0, when H_0 is false}

= Power of the test

The probability of type I error is known as **level of significance.** Generally, the community used levels of significance are 5% (0.05) and 1% (0.01). A critical value or level of significance is the probability set by the researcher during the test.

It is denoted by **α**. If 'α' is the critical value then '1 – α' is known as **confidence coefficient.**

20.4 Applications of F-test:

F-test has a number of applications in statistics. Some of the applications of F-test are,

1. F-test is used to test equality of several means.
2. F-test is used to test the significant difference between the two variances.
3. F-test is used for testing the significance of an observed sample correlation ratio.
4. F-test is used for testing the significance of an observed sample multiple correlation

20.5. Assumptions of F-test:

The basic assumptions to apply F-test in any test of significance are

i) The samples are independent of each other.

ii) The samples are simple random samples.

iii) The parent population from which the samples are drawn, are Normal.

iv) The total variance of the various sources of variance should be additive.

v) Since F – is always formed by a ratio of squared values. It can have be a negative number. (larger variance / smaller variance).

20.6 F – test for equality of population variances:

Suppose we want to test whether there is any significant difference between two sample variances (or) the population variances are equal.

Working procedure:

Step 1: Null Hypothesis: There is no significant difference between the two sample variances or the population variances are equal i.e., H_0: $\sigma_1^2 = \sigma_2^2$

Step 2: Alternative Hypothesis: H_1: $\sigma_1^2 \neq \sigma_2^2$

Step 3: The required test statistic for testing the above null hypothesis is

$$F = \frac{S_1^2}{S_2^2} \sim F_{(n_1 - 1, n_2 - 1)} \text{ if } S_1^2 > S_2^2$$

or $$F = \frac{S_1^2}{S_1^2} \sim F_{(n_2 - 1, n_1 - 1)} \text{ if } S_2^2 > S_1^2$$

n_1 = size of the first sample

n_2 = size of the second sample and

$$S_1^2 = \frac{1}{n_1 - 1}\left[\sum (x - \bar{x})^2\right] \text{ and } \qquad S_2^2 = \frac{1}{n_2 - 1}\left[\sum (y - \bar{y})^2\right]$$

Step 4: Identify the tabulated value (critical value) of 'F' for (n_1 – 1, n_2 - 1) or (n_2 – 1, n_1 -1) degrees of freedom at α% level of significance.

Step 5: We compare the computed value of 'F' in step 4 with the significant value 'F' at given level of significance, 'α'.

If calculated value of 'F' or $|F|$ is less than the tabulated value of 'F' at α% level of significance we accept H_0 otherwise we reject H_0.

20.7 : Solved Problems:

Problem 1: -

Two random sample sizes 9 and 12 have the standard deviations 2.9 and 2.6 drawn from two normal populations. Test the significance difference between the sample variances.

1) Null Hypothesis :-

$H_0 : \sigma_1^2 = \sigma_2^2$

i.e., there is no significant difference between the sample variances.

2) Alternative Hypothesis:-

$H_1 : \sigma_1^2 \neq \sigma_2^2$

i.e., there is a significant difference between the sample means.

3) Test statistic under H_0 :

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{9 \times (2.9)^2}{9-1} = 9.46$$

$$S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{12 \times (2.6)^2}{12-1} = 7.37$$

Since $S_1^2 > S_2^2$

$$F = \frac{S_1^2}{S_2^2} \sim F_{(n_1 - 1, n_2 - 1)} = \frac{9.46}{7.37} = 1.28.$$

$F_{(\alpha,(n_1-1),(n_2-1))} = F_{(5\%,(8,11))}$ =2.95 at 5% level and at (8 , 11) of from F-tables

$\therefore F < F_{(5\%,(8,11))} \Rightarrow H_0$ may be accepted.

i.e., There is no significant difference between the sample variances.

Problem 2: -

In one sample of 8 observations, we sum of the squares of the deviation of the sample values from the sample mean was 84.4 and in the other sample of 10 observations it was 102.6.Test whether this difference is significant at 1% level.

Solution : -

Given that $n_1 = 8, n_2 = 10$

$$\sum_{i=1}^{n_1} (x_i - \bar{x})^2 = 84.4$$

$$\sum_{i=1}^{n_2} (y_i - \bar{y})^2 = 102.6$$

1. Null Hypothesis

$H_0 : \sigma_1^2 = \sigma_2^2$

i.e., the difference is not significant.

2. Alternative Hypothesis

$H_1 : \sigma_1^2 \neq \sigma_2^2$

i.e., the difference is significant

3. Test statistic under H_0 :

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(x_i - \bar{x}\right)^2 = \frac{84.4}{8-1} = 12.06$$

$$S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}\left(y_i - \bar{y}\right)^2 = \frac{102.6}{10-1} = 11.4$$

Since $S_1^2 > S_2^2$,

$$F = \frac{S_1^2}{S_2^2} \sim F_{(n_1 - 1,\, n_2 - 1)} = \frac{12.06}{11.4} = 1.06$$

Inference: -

F = 1.06

$F_{(\alpha,(n_1-1,n_2-1))}$ = $F_{(1\%,(7,9))}$ = 5.62 at 1% level, (7, 9) d.f. from F-tables

$\therefore F < F_{(1\%,(7,9))} \Rightarrow H_0$ may be accepted.

i.e., The difference of the two samples are not significant.

Problem 3: -

Two random samples of size 10 and 12 are drawn and given below. Test the equality of population variances.

Sample I	10	6	16	17	13	12	8	15	9	14		
Sample II	7	13	22	15	12	14	18	8	21	23	10	7

Solution : - Given that $n_1 = 10$, $n_2 = 12$.

x_i	x_1^2	y_i	y_1^2
10	100	7	49
6	36	13	169
16	256	22	484
17	289	15	225
13	169	12	144
12	144	14	196
8	64	18	324
15	225	8	64

9	81	21	441
14	196	23	529
		10	100
		7	49
$\sum_{i=1}^{n_1} x_i = 120$	$\sum_{i=1}^{n_1} x_i^2 = 1560$	$\sum_{j=1}^{n_2} y_j = 170$	$\sum_{j=1}^{n_2} y_j^2 = 2774$

$$\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} xi = \frac{120}{10} = 12$$

$$\bar{y} = \frac{1}{n_2}\sum_{j=1}^{n_2} y_j = \frac{170}{12} = 14.17$$

$$s_1^2 = \frac{1}{n_1}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2 = \frac{1560}{10} - (12)^2 = 12$$

$$s_2^2 = \frac{1}{n_2}\sum_{j=1}^{n} y_j^2 - (\bar{y})^2 = \frac{2774}{12} - (14.17)^2 = 30.398$$

$$s_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{10 \times 12}{9} = 13.33$$

$$s_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{12 \times 30.98}{11} = 33.14$$

Another way of calculation of s_1^2 and s_2^2

$$s_1^2 = \frac{1}{n_1 - 1}\left[\sum_{i=1}^{n} x_i^2 - n_1 \bar{x}^2\right] = \frac{1}{9}\left[1560 - 10 \times 12^2\right] = 13.33$$

$$s_2^2 = \frac{1}{n_2 - 1}\left[\sum_{j=1}^{n_2} y_j^2 - n_2 \bar{y}^2\right] = \frac{1}{11}\left[2774 - 12 \times (14.17)^2\right] = 33.14$$

1) Null Hypothesis:

H_0: $\sigma_1^2 = \sigma_2^2$

Two population variances are equal.

2) Alternative Hypothesis:

H_1: $\sigma_1^2 \neq \sigma_2^2$

Two population variance are not equal.

3) Test statistic under H_0 :

Since $s_2^2 > s_1^2$

$$F = \frac{s_2^2}{s_1^2} \sim F_{(n_2-1, n_1-1)} = \frac{33.14}{13.33} = 2.49$$

4) Inference: -

F = 2.49

$F_{(\alpha,(n_2-1,n_1-1))} = F_{5\%,(11,9)} = 3.10\ at\ 5\%$ level of, (11 , 9) df from F-tables

$\therefore F < F_{(5\%,(11,9))} \Rightarrow H_0$ may be accepted.

i.e population variances are equal.

Problem 4: -

In random samples of sizes 10 and 15, the unbiased estimators of variances are found to be 5 and 9 respectively. Can we reasonably conclude that the population variances are equal.

Solution: -

Given that $n_1 = 10$, $n_2 = 15$

$s_1^2 = 5$, $s_2^2 = 9$

1. Null hypothesis: -

H_0: $\sigma_1^2 = \sigma_2^2$

Two population variances are equal.

2. Alternative Hypothesis: -

H_1: $\sigma_1^2 \neq \sigma_2^2$

Two population variances are not equal.

3. Test statistic under H_0:

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{10 \times 5}{9} = 5.56$$

$$S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{15 \times 9}{14} = 9.64$$

Since $S_2^2 > S_1^2$

$$F = \frac{S_2^2}{S_1^2} \sim F_{(n_2-1, n_1-1)} = \frac{9.64}{5.56} = 1.73$$

Inference: -

F = 1.73

$F_{(\alpha,(n_2-1,n_1-1))}$ = $F_{(5\%,(14,9))}$ = 3.02 at 5% level, (14, 9) d.f from F-tables

$\therefore F < F_{(5\%,(14,9))} \Rightarrow H_0$ may be accepted.

i.e., it can be reasonably conclude that population variances are equal.

Problem 5: -

The time taken by workers in performing a job by method I and II is given below :

Method I :	20	16	26	27	23	22	
Method II :	27	33	42	35	32	34	38

Do the data show that the variances of time distribution from population from which these samples are drawn do not differ significantly?

Solution: -

Null Hypothesis. $H_0\ \sigma_1^2 = \sigma_1^2$, i.e., there is no significant difference between the variances of the time distribution by the workers in performing a job by Method I and Method II.

Computation of sample variances

x	D = x - 22	d^2
20	-2	4
16	-6	36
26	4	16
27	5	25
23	1	1
22	0	0
Total	$\sum d = 2$	$\sum d^2 = 82$

y	D = y - 35	D^2
27	-8	64
33	-2	4
42	7	49
35	0	0
32	-3	9
34	-1	1
38	3	9
Total	$\sum D = -4$	$\sum D^2 = 136$

$$S_1^2 = \frac{1}{n_1-1}\sum(x-\bar{x})^2 = \frac{1}{n_1-1}\left(\sum d^2 - \frac{(\sum d)^2}{n_1}\right)$$

$$= \frac{1}{5}\left(82 - \frac{4}{6}\right) = \frac{1}{5}(82-0.67) = 16.266$$

$$= S_2^2 = \frac{1}{n_2 - 1}\sum (y - \bar{y})^2 = \frac{1}{n_2 - 1}\left(\sum D^2 - \frac{(\sum D)^2}{n_2}\right)$$

$$= \frac{1}{6}\left(136 - \frac{16}{7}\right) = \frac{1}{6}(136 - 2.286) = 22.286$$

Since $S_2^2 > S_1^2$, under H_0, the test statistic is

$$F = \frac{S_2^2}{S_1^2} \sim F_{(n_2 - 1, n_1 - 1)} = F(6,5)$$

Alternative Hypothesis : $H_1 : \sigma_2^2 > \sigma_1^2$

$$F = \frac{22.286}{16.266} = 1.37$$

Tabulated $F_{0.05}(6, 5) = 4.95$.

Since calculated F is less than tabulated F, it is not significant. Hence H_0 may be accepted at 5% level of significance and we conclude that variability of the time distribution in the two populations is same.

Problem No.: 6

It is known that the mean diameters of rivets produced by two firms A and B are practically the same but the standard deviations may differ. For 22 rivets produced by firm A. The standard deviation is 2.9 mm. while for 16 rivets manufactured by firm B, the standard deviation is 3.8 mm. compute the statistic you would use to test whether the product B of firm A have the same variability as those of firm of firm B and test its significance,

Sol:- We are given

$n_1 = 22$, $s_1 = 2.9$ mm; $n_2 = 16$, $s_2 = 3.8$ mm.

Null Hypothesis. $H_0 = \sigma_1^2 = \sigma_2^2$, i.e., the products of both the firms A and B have the same variability.

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{22 \times (2.9)^2}{21} = 8.805;$$

$$S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{16 \times (3.8)^2}{15} = 15.393,$$

Since $s_2^2 > s_1^2$ under H_0, the test statistic is

$$F = \frac{S_2^2}{S_1^2} = \frac{15.393}{8.805} = 1.74(\text{approx})$$

which follows F-distribution wit (15, 21) d.f.

Tabulate: $F_{0.05}$ (15 , 21) = 2.20 (approx)

Since calculated F is less than tabulated F, it is not significant at 5% level of significance and the hypothesis of equal variability may be accepted.

20.8 Summary

In this chapter, we shall begin with considering tests of hypothesis about variances where a two samples are involved. We deal with those situations where variances of two samples need to be compared, whether one is higher than the other or they are just equal. This test is done by the another sampling distribution known as the F-distribution. In this context, F-test are used to test the equality of two population variances. In addition that, two types of errors are play key role while taking the decisions in testing of hypothesis.

20.9 Self assessment questions :

1. Define two types of errors
2. State the applications and assumptions of F-test
3. Explain the procedure of F-test, for testing equality of two population variances.
4. Two random samples of sizes 13 and 15 gave the population variances 3.0 and 2.5 respectively. Can we assume that both the samples be regarded from same normal population.

   (Ans: F = 1.2, H_0 is accepted)
5. Two random samples of sizes 9 and 8 gave the squares of deviations from their respective means equal to 160 and 91 respectively. Can they be regarded as drawn from same normal populations with respect to variance.

   F(= 1.54, H_0 is accepted)
6. Define (i) Level of significance

   (ii) Power of the test

   (iii) Confidence coefficient

20.10 Reference Books

1. S.C. Gupta, Fundamentals of Statistics, Himalaya Publishing House.
2. Digambar Patri., Quantitative Techniques, kalyani publications.
3. P.N. Arora and S. Arora, Statistics for Management: S. Chand & Comp, Ltd.
4. G.V. Shenoy, Uma K. Srivastava, S.C. Sharma: Business Statistics
5. B. M.Agarwal,: Business statistics
6. Gupta S.P.: Statistical Methods

Lesson Writer

Dr. J. Pratapa Reddy