MMS SEM - I
MASTERS IN MANAGEMENT STUDIES

BUSINESS STATISTICS

Dr. Suhas Pednekar
Vice Chancellor,
University of Mumbai

Dr. Dr. Prakash Mahanwar
Director, Institute of Distance and Open Learning,
University of Mumbai.

Dr. Madhura Kulkarni
Incharge Study Material Section,
Institute of Distance and Open Learning,
University of Mumbai.

Programme Co-ordinator :

Course Writer :

August 2021, MMS SEM - I, Masters in Management Studies, Business Statistics

Published by : Incharge Director
Institute of Distance and Open Learning ,
University of Mumbai,
Vidyanagari, Mumbai - 400 098.

DTP Composed : Ashwini Arts
Gurukripa Chawl, M.C. Chagla Marg, Bamanwada,
Vile Parle (E), Mumbai - 400 099.

Printed by :

MMS – Masters in Management Studies

Semester – I

Business Statistics

Syllabus

Semester	**:**	**I – Core**			
Title of the Subject / course	**:**	**Business Statistics**			
Course Code	**:**				
Credits	**:**	**4**	**Duration in Hrs.**	**:**	**40**

Learning Objectives

1	To know statistical techniques
2	To understand different statistical tools
3	To understand importance of decision support provided by analysis techniques
4	To appreciate and apply it in business situations using caselets, modeling, cases and projects
5	To understand Managerial applications of Statistics

Prerequisites if any	Basic Mathematics
Connections with Subjects in the current or Future courses	Operations Research, Economics, Research Methodology, Quantitative Techniques, Project Management, Financial Management, production and operations management,

Module

Sr. No.	Content	Activity	Learning outcomes
1	Revision of Data Representation, Central Tendency and Dispersion Kurtosis and Skewness	Problem solving, cases demonstrating typical uses of mean, mode median, Use of Microsoft Excel, available software	Learner will be able to apply these basic concepts in business situations, Analyse charts graphs to analyse business situations
2	Probability- Axioms, Addition and Multiplication rule, Types of probability, Independence of events, probability tree, Bayes' Theorem	Solving problems and Caselets, Writing short cases	Understand the uncertainty in business situations as probability
3	Concept of Random variable, Probability distribution, Expected value and variance of random variable, conditional expectation, Classical News Paper boys problem(EMV, EVPI)	Problem solving , Creating decision tree, cases	Understand decision under risk, use of conditional expectation as basis for comparison
4	Probability distributions Binomial, Poisson, Normal	Problem solving, Microsoft excel, cases	Use of distributions in Quality control, Six sigma and process control
5	Sampling distribution	Problem solving, Microsoft Excel	Importance of Central limit theorem
6	Estimation- Point estimation , Interval estimation	Problem solving, Microsoft Excel	Understand Confidence interval as way of hypothesis testing

7	Hypothesis testing- students t, Chi square, Z	Problem solving, Microsoft excel, cases	Use in research
8	Analysis of variance- one way, two way	Problem solving, Microsoft excel, cases	Use in research

Text books

1	Statistics for Management	Richard Levin , David Rubin, Prentice Hall of India
2	Statistics for Managers	Levine, Stephen, Krihbiel, Berenson, Pearson Education
3	Complete Business Statistics	Aczel Sounderpandian, Tata McGraw Hill

Reference books

1	Statistics for Business and Economics	Newbold, Carlson, Thorne, Pearson Education
2	Statistics for Business and Economics	Anderson, Sweeney, Williams, Cengage Learning
3	Data Analysis and Decision Making	Albright, Winston, Zappe, Thomson

Assessment

Internal	40%
Semester end	60%

1

INTRODUCTION TO STATISTICS

Unit Structure

1.1 MEANING

The word Statistics describes several concepts of importance to decision-maker. It is important for a beginner to have an understanding of these different concepts.

STATISTICAL METHODS V/S EXPERIMENTAL METHODS

We try to get the knowledge of any phenomenon through direct experiment. There may be many factors affecting a certain phenomenon simultaneously. If we want to study the effect of a particular factor, we keep other factors fixed and study the effect of only one factor. This is possible in all exact sciences like Physics, Chemistry etc. This method cannot be used in many sciences, where all the factors cannot be isolated and controlled. This difficulty is particularly encountered in social sciences, where we deal with human beings. No two persons are exactly alike. Besides the environment also changes and it has its effect on every human being and therefore it is not possible to study one factor keeping other conditions fixed. Here we use statistical methods. The results obtained by the use of this science will not be as accurate as those obtained by experimental methods. Even then they are of much use and they have a very important role to play in the modern World. Even in exact sciences some of the statistical methods are made use of.

The word Statistics is derived from the Latin word "statis' which means a political state. The word Statistics was originally applied to only such facts and figures that were required by the state for

official purposes. The earliest form of statistical data is related to census of population and property, through the collection of data for other purposes was not completelyruled out. The word has now acquired a wider meaning.

1.2 STATISTICAL METHODS

The word Statistics used in the second sense means the set of techniques and principles for dealing with data.

1. Suppose you have the data about production profits and sales for a number of years of a company. Statistics in this sense is concerned with questions such as
   (i) What is the best way to present these data for review?
   (ii) What processing is required to reveal more details about the data?
   (iii) What ratios should be obtained and reported?
2. A public agency wants to estimate the number of fish in a lake. Five hundred fish are captured in a net tagged and returned to the lake. One week later 1, 000 fish are captured from the same lake in nets and 40 are found to be with tags. Here Statistics in this second sense deals with questions such as:
   (i) What is a good estimate of the number of fish in the lake?
(ii) What is our confidence in it and how much error can be expected? and
   (iii) Can we have a method, which will make a better estimate?

Statisticians have defined this in various ways. Bowley says, "Statistics may rightly be called the science of averages." But this definition is not correct. Statistics has many refined techniques and it does much more than just averaging the data.

Kendall defines it as, "The branch of scientific methods that deals with the data obtained by counting or measuring the properties of population of natural phenomena." This definition does not give the idea about the functions of Statistics. It is rather vague.

Seligman defines it as, "The science which deals with the methods of collecting, classifying, presenting, comparing and interpreting numerical data collected to throw some light on any sphere of inquiry." Croxton, Cowden and Klein define it as, "The last two definitions can be considered to be proper which explain the utility of 'statistics'. We will examine the four procedures mentioned in the definition in brief.

Collection: The day may be collected from various published and unpublished sources, or the investigator can collect his own information. Collecting first hand information is a very difficult task. The usefulness of the data collected depends to a very great extent upon the manner in which they are collected. Though theoretical knowledge is necessary for the propercollection of data, much can be learnt through experience and observation.

Presentation: The data collected, to be understood, should be presented in a suitable form. Just a given mass of figures signifies nothing to a person and they can lead only to confusion. They are usually presented in a tabular form and represented by diagrams.

Analysis: Presentation of data in a tabular form is one elementary step in the analysis of the collected data. If we want to compare two series, a typical value for each series is to be calculated. If we want to study some characteristic of a big group, exhaustive study is not possible. We take a sample, study it and inferences are drawn on the basis of sample studies. Sometimes forecasting is necessary. The management of a firm may be interested in future sales. For that it has to analyse the past data. We are going to study some of these methods of analysing the data in this book.

Interpretation: This is the final step in an investigation. Based upon the analysis of the data, we

draw certain conclusions. While drawing these conclusions, we must consider that nature of the original data. Experts in the particular field of activity must make the final interpretation. The statistical methods are not like experimental methods, which are exact. For interpreting the analysis of the data dealing with some psychological problems, a psychologist is right person. (An economist, though well versed in statistical methods will notbe of any use there).

STATISTICAL MEASURES

Statistics also has a precise technical meaning. Measures derived from the sample data are referred to as Statistics. If only one measure is obtained it is called a Statistic.

A magazine takes a sample of 100 readers. 15 of them are over 30 years of age. The sample proportion of readers over 30 years of age is 0.15. This sample proportion is referred to as a statistic obtained by this survey.

The weekly sales for 5 weeks for a salesman are Rs. 2, 000, Rs. 2, 500, Rs. 15, 000, Rs. 3000 and Rs. 1, 800. As a measure of the spread of the values the difference between the smallest and the largest value (called the range) is calculated. This range is a statistic.

1.3 IMPORTANCE OF STATISTICS

Statistics is not studied for its own sake. It is employed as a tool to study the problems in various natural and social sciences. The analysis of data is used ultimately for forecasting, controlling and exploring.

Statistics is important because it makes the data comprehensible. Without its use the information collected will hardly be useful. To understand the economic condition of any country we must have different economic aspects quantitatively expressed and properly presented. If we want to compare any two countries, statistics is to be used. For studying relationship between two phenomena, we have to take the help of statistics, which explains the correlation between the two. People in business can study past data and forecast the condition of their business, so that they can be ready to handle the situations in future. Nowadays a businessman has to deal withthousands of employees under him and cannot have direct control over them. Therefore, he can judge them all and control their performance using statistical methods e.g., he can set up certain standards and see whether the final product conforms to them. He can find out the average production per worker and see whether any one is giving less, i.e., he is not working properly.

Business must be planned properly and the planning to be fruitful must be based on the right analysis of complex statistical data. A broker has to study the pattern in the demand for money by his clients, so that he will have correct amount of reserves ready.

Scientific research also uses statistical methods. While exploring new theories, the validity ofthe theory is to be tested only by using statistical methods. Even in business many new methods are introduced. Whether they are really an improvement over the previous ones, can be tested using statistical techniques.

We can see many more examples from almost all sciences, like biology, physics, economics, psychology and show that statistical methods are used in all sciences. The point here is that 'Statistics' is not an abstract subject. It is a practical science and it is very important in the modern World.

1.4 FUNCTIONS OF STATISTICS

1. **Statistics presents the data in numerical form:** Numbers give the exact idea about any phenomenon. We know that India is overpopulated. But only when we see the census figure, 548 millions, we have the real idea about the population problem. If we want to compare the speed of two workmen working in the same factory, with the same type of machine, we have to see the number of units they turn out every day. Only when we express the facts with the help of numbers, they are convincing.
2. **It simplifies the complex data:** The data collected are complex in nature. Just by looking at the figures no person can know the real nature of the problem under consideration. Statistical methods make the data easy to understand. When we have data about the students making use of the college library, we can divide the students according to the number of hours spent in the library. We can also see how many are studying and how many are sitting there for general reading.
3. **It facilitates comparison:** We can compare the wage conditions in two factories by comparing the average wages in the two factories. We can compare the increase in wages and corresponding increase in price level during that period. Such comparisons are very useful in many social sciences.
4. **It studies relationship between two factors:** The relationship between two factors, like, height and weight, food habits and health, smoking and occurrence of cancer can be studied using statistical techniques. We can estimate one factor given the other when there is some relationship established between two factors.
5. **It is useful for forecasting:** We are interested in forecasting using the past data. A shopkeeper may forecast the demand for the goods and store them when they are easily available at a reasonable price. He can store only the required amount and there will not be any problem of goods being wasted. A baker estimates the daily demand for bread, and bakes only that amount so that there will be no problem of leftovers.
6. **It helps the formulation of policies:** By studying the effect of policies employed so far by analysing them, using statistical methods, the future policies can be formulated. The requirements can be studied and policies can be determined accordingly. The import policy for food can be determined by studying the population figures, their food habits etc.

1.5 LIMITATIONS OF STATISTICS

Though Statistics is a very useful tool for the study of almost all types of data it has certain limitations.

1. **It studies only quantitative data:** A very serious drawback is that statistics cannot study qualitative data. Only when we have data expressed in numerical form we can apply statistical methods for analysing them. Characteristics like beauty, cruelty, honesty or intelligence cannot be studied with the help of statistics. But in some cases we can relate the characteristics to number and try to study them. Intelligence of students can be studied by the marks obtained by them in various tests, we can compare the intelligence of students or arrange them in order if we take marks as an indicator of intelligence. Culture of a society or the lack of it can be studied considering the number of charitable institutions, their sizes and number of crimes.

2. **It cannot be used for an individual:** The conclusions drawn from statistical data are true for a group of persons. They do not give us any knowledge about an individual. Though Statistics can estimate the number of machines in a certain factory that will fail after say, 5 years, it cannot tell exactly which machines will fail. One in 2, 000 patients may die in a particular operation. Statistically this proportion is very small and insignificant. But for the person who dies and his family, the loss is total. Statistics shows now sympathy for such a loss.
3. **It gives results only on an average:** Statistical methods are not exact. The results obtained are true only on an average in the long run. When we say that the average student studies for 2 hours daily there may not be a singly student studying for 2 hours, not only that, every day the average will not be 2 hours. In the long run, if we consider a number of students, the daily average will be 2 hours.
4. **The results can be biased:** The data collected may sometimes be biased which will make the whole investigation useless. Even while applying statistical methods the investigator has to be objective. His personal bias may unconsciously make him draw conclusions favourable in one way or the other.
5. **Statistics can be misused:** It is said that statistics can prove or disprove anything. It depends upon how the data are presented. The workers in a factory may accuse the management of not providing proper working conditions, by quoting the number of accidents. But the fact may be that most of the staff is inexperienced and therefore meet with an accident. Besides only the number of accidents does not tell us anything. Many of them may be minor accidents. With the help of the same data the management can prove that the working conditions are very good. It can compare the conditions with working conditions in other factories, which may be worse. People using statistics have to be very careful to see thatit is not misused.

Thus, it can be seen that Statistics is a very important tool. But its usefulness depends to a great extent upon the user. If used properly, by an efficient and unbiased statistician, it will prove to be a wonderful tool.

1.6 BRANCHES IN STATISTICS

Statistics may be divided into two main branches:

1. **Descriptive Statistics:** In descriptive statistics, it deals with collection of data, its presentation in various forms, such as tables, graphs and diagrams and findings, averages and other measures which would describe the data.

For example, Industrial Statistics, population statistics, trade statistics etc....Such as businessmen make to use descriptive statistics in presenting their annual reports, final accounts and bank statements.

2. **Inferential Statistics:** In inferential statistics deals with techniques used for analysis of data, making the estimates and drawing conclusions from limited information taken onsample basis and testing the reliability of the estimates.

For example, suppose we want to have an idea about the percentage of illiterates in our country. We take a sample from the population and find the proportion of illiterates in the sample. This sample proportion with the help of probability enables us to make some inferences about the population proportion. This study belongs to inferential statistics.

1.7 CHARACTERISTICS OF STATISTICS

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

1.8 SOME BASIC DEFINITIONS IN STATISTICS

Constant: A quantity which can be assuming only one value is called a constant. It is usually denoted by the first letters of alphabets a, b, c.

For example value of $\pi = 22/7 = 3.14159....$ and value of $e = 2.71828....$

Variable: A quantity which can vary from one individual or object to and other is called a variable. It is usually denoted by the last letters of alphabets x, y, z.

For example, heights and weights of students, income, temperature, number of children in a family etc.

Continuous variable: A variable which can assume each and every value within a given range is called a continuous variable. It can occur in decimals.

For example, heights and weights of students, speed of a bus, the age of a shopkeeper, the life time of a T.V. etc.

Continuous Data: Data which can be described by a continuous variable is called continuous data.

For example: Weights of 50 students in a class.

Discrete Variable: A variable which can assume only some specific values within a given range is called discrete variable. It cannot occur in decimals. It can occur in whole numbers. For example: Number of students in a class, number of flowers on the tree, number of housesin a street, number of chairs in a room etc...

Discrete Data: Data which can be described by a discrete variable is called discrete data.

For example, Number of students in a College.

Quantitative Variable: A characteristic which varies only in magnitude from an individualto another is called quantitative variable. It can be measurable.

For example, wages, prices, heights, weights etc.

Qualitative Variable: A characteristic which varies only in quality from one individual to another is called qualitative variable. It cannot be measured.

For example, beauty, marital status, rich, poor, smell etc.

1.9 EXERCISE

1. Explain the meaning of statistics.
2. Give a definition of statistics and discuss it.
3. Explain the functions of statistics.
4. What are the limitations of statistics?
5. Define the term Statistics and discuss its characteristics.
6. Enumerate with example some terms of Statistics.
7. Discuss on the different branches of Statistics.

2

DATA: COLLECTION AND PRESENTATION

Unit Structure

2.1.1 STATISTICAL DATA

A sequence of observation made on a set of objects included in the sample drawn from population is known as statistical data.

1. **Ungrouped Data:** Data which have been arranged in a systematic order are called raw data or ungrouped data.
2. **Grouped Data:** Data presented in the form of frequency distribution is called grouped data.

2.1.2 COLLECTION OF DATA

The first step in any enquiry (investigation) is collection of data. The data may be collected for the whole population or for a sample only. It is mostly collected on sample basis. Collection of data is very difficult job. The enumerator or investigator is the well trained person who collects the statistical data. The respondents (information) are the persons whom the information is collected.

2.1.3 TYPES OF DATA

There are two types (sources) for the collection of data:

1. **Primary Data:** The primary data are the first hand information collected, compiled and published by organisation for some purpose. They are most original data in character and have not undergone any sort of statistical treatment.

For example, Population census reports are primary data because these are collected, compiled and published by the population census organisation.

2. **Secondary Data:** The secondary data are second hand information which are already collected by someone (organisation) for some purpose and are available for the present study. The secondary data are not pure in character and have undergone some treatment at least once.

For example, Economics survey of England is secondary data because these are collected by more than one organisation like Bureau of Statistics, Board of Revenue, the Banks etc.

2.1.4 (a) METHODS OF COLLECTING PRIMARY DATA

Primary data are collected by the following methods:

1. **Personal Investigation:** The researcher conducts the survey him/herself and collects data from it. The data collected in this way is usually accurate and reliable. This method of collecting data is only applicable in case of small research projects.
2. **Through Investigation:** Trained investigators are employed to collect the data. These investigators contact the individuals and fill in questionnaire after asking the required information. Most of the organisations implied this method.
3. **Collection through questionnaire:** The researchers get the data from local representation or agents that are based upon their own experience. This method is quick but gives only rough estimate.
4. **Through Telephone:** The researchers get information through telephone. This method is quick.

2.1.4 (b) METHODS OF COLLECTING SECONDARY DATA

The secondary data are collected by the following sources:

- Official: The publications of Statistical Division, Ministry of Finance, the Federal Bureaus of Statistics, Ministries of Food, Agriculture, Industry, Labour etc....
- Semi-Official: State Bank, Railway Board, Central Cotton Committee, Boards of Economic Enquiry etc....
- Publication of Trade Associations, Chambers of Commerce etc....
- Technical and Trade Journals and Newspapers.
- Research Organisations such as Universities and other Institutions.

2.1.4 (c) DIFFERENCE BETWEEN PRIMARY AND SECONDARY DATA

The difference between primary and secondary data is only a change of hand. The primary data are the first hand information which is directly collected from one source. They are most original data in character and have not undergone any sort of statistical treatment while the secondary data are obtained from some other sources or agencies. They are not pure in character and have undergone some treatment at least once.

For example, suppose we are interested to find the average age of MS students. We collect the age's data by two methods; either by directly collecting from each student himself personally or getting their ages from the University record. The data collected by the direct personal investigator is called primary data and the data obtained from the University record is called Secondary data.

EDITING OF DATA

After collecting the data either from primary or secondary source, the next step is its editing. Editing means the examination of collected data to discover any error before presenting it. It has to be decided before hand what degree of accuracy is wanted and what extent of errorscan be tolerated in the inquiry. The editing of secondary data is simpler than that of primary data.

2.2 CLASSIFICATION OF DATA

The process of arranging data into homogenous group or classes according to some common characteristics present in the data is called classification.

For example, the process of sorting letters in a post office, the letters are classified according to the cities and further arranged according to streets.

2.2.1 BASES OF CLASSIFICATION

There are four important bases of classification:

1. Qualitative Base
2. Quantitative Base
3. Geographical Base
4. Chronological or Temporal Base

1. **Qualitative Base:** When the data are classified according to some quality or attributes such as sex, religion, literacy, intelligence etc...
2. **Quantitative Base:** When the data are classified by quantitative characteristics likeheights, weights, ages, income etc..
3. **Geographical Base:** When the data are classified by geographical regions or location, like states, provinces, cities, countries etc.
4. **Chronological or Temporal Base:** When the data are classified or arranged by their timeof occurrence, such as years, months, weeks, days etc .....For example, Time Series Data.

2.2.2 TYPES OF CLASSIFICATION

1. **One-way classification:** If we classify observed data keeping in view single characteristic, this type of classification is known as one-way classification.

For example, the population of world may be classified by religion as Muslim, Christian etc.

2. **Two-way classification:** If we consider two characteristics at a time in order to classify the observed data then we are doing two-way classification.

For example, the population of world may be classified by Religion and Sex.

3. **Multi-way classification:** We may consider more than two characteristics at a time to classify given data or observed data. In this way we deal in multi-way classification.

For example, the population of world may be classified by Religion, Sex and Literacy.

2.3 TABULATION OF DATA

The process of placing classified data into tabular form is known as tabulation. A table is a symmetric arrangement of statistical data in rows and columns. Rows are horizontal arrangements whereas columns are vertical arrangements. It may be simple, double or complex depending upon the type of classification.

2.3.1 TYPES OF TABULATION

1. **Simple Tabulation or One-way tabulation:** When the data are tabulated to one characteristic, it is said to be simple tabulation or one-way tabulation.

For example, tabulation of data on population of world classified by one characteristic like Religion is example of simple tabulation.

2. **Double Tabulation or Two-way tabulation:** When the data are tabulated according to two characteristics at a time. It is said to be double tabulation or two-way tabulation.

For example, tabulation of data on population of world classified by two characteristics like religion and sex is example of double tabulation.

3. **Complex Tabulation:** When the data are tabulated according to many characteristics, it is said to be complex tabulation.

For example, tabulation of data on population of world classified by two characteristics like Religion, Sex and Literacy etc..... is example of complex tabulation.

DIFFERENCES BETWEEN CLASSIFICATION AND TABULATION

1. First the data are classified and then they are presented in tables, the classification and tabulation in fact goes together. So classification is the basis for tabulation.
2. Tabulation is a mechanical function of classification because in tabulation classified dataare placed in row and columns.
3. Classification is a process of statistical analysis whereas tabulation is a process ofpresenting the data in suitable form.

2.4 FREQUENCY DISTRIBUTION

A frequency distribution is a tabular arrangement of data into classes according to the size or magnitude along with corresponding class frequencies (the number of values fall in each class).

Ungrouped data or Raw Data

Data which have not been arranged in a systemic order is called ungrouped or raw data.

Grouped Data

Data presented in the form of frequency distribution is called grouped data.

Array

The numerical raw data is arranged in ascending or descending order is called an array.

Example

Array the following data in ascending or descending order 6, 4, 13, 7, 10, 16, 19.

Solution

Array in ascending order is 4, 6, 7, 10, 13, 16 and 19.

Array in descending order is 19, 16, 13, 10, 7, 6, and 4.

CLASS LIMITS

The variant values of the classes or groups are called the class limits. The smaller value of theclass is called lower class limit and larger value of the class is called upper class limit. Class limits are also called inclusive classes.

For example, let us take class 10-19, the smaller value 10 is lower class limit and larger value19 is called upper class limit.

CLASS BOUNDARIES

The true values, which describes the actual class limits of a class, are called class boundaries. The smaller true value is called the lower class boundary and the larger true value is calledthe upper class boundary of the class. It is important to note that the upper class boundary ofa class coincides with the lower class boundary of the next class. Class boundaries are also known as exclusive classes.

For example,

Weights in Kg	**Number of Students**
60-65	8
65-70	12
70-75	5
	25

A student whose weights are between 60 kg and 64.5 kg would be included in the 60-65 class. A student whose weight is 65 kg would be included in next class 65-70.

A class has either no lower class limit or no upper class limit in a frequency table is called an open-end class. We do not like to use open-end classes in practice, because they create problems in calculation.

For example,

Weights (Pounds)	**Number of Persons**
Below - 110	6
110-120	12
120-130	20
130-140	10
140-above	2

Class Mark or Mid Point

The class marks or mid point is the mean of lower and upper class limits or boundaries. So it divides the class into two equal parts. It is obtained by dividing the sum of lower and upper- class limit or class boundaries of a class by 2.

For example, The class mark or mid-point of the class 60-69 is 60+69/2 = 64.5

Size of Class Interval

The difference between the upper and lower class boundaries (not between class limits) of a class or the difference between two successive mid points is called size of class interval.

2.4.1 CONSTRUCTION OF FREQUENCY DISTRIBUTION

Following steps are involved in the construction of a frequency distribution.

1. **Find the range of the data:** The range is the difference between the largest and the smallest values.
2. **Decide the approximate number of classes:** Which the data are to be grouped. There are no hard and first rules for number of classes. Most of the cases we have 5 to 20 classes. H. A. Sturges has given a formula for determining the approximation number of classes.

K = 1 + 3.322 log N
where K = Number of classes
where log N = Logarithm of the total number of observations
For example, if the total number of observations is 50, the number of classes would be:K = 1 + 3.322 log N
K = 1 + 3.322 log 50
K = 1 + 3.322 (1.69897)
K = 1 + 5.644
K = 6.644 or 7 classes approximately.

3. **Determine the approximate class interval size:** The size of class interval is obtained by dividing the range of data by number of classes and denoted by h class interval size

(h) = Range/Number of Classes
In case of fractional results, the next higher whole number is taken as the size of the class interval.

Decide the starting Point: The lower class limits or class boundary should cover the smallest value in the raw data. It is a multiple of class interval.
For example, 0, 5, 10, 15, 20 etc... are commonly used.

4. **Determine the remaining class limits (boundary):** When the lowest class boundary of the lowest class has been decided, then by adding the class interval size to the lower class boundary, compute the upper class boundary. The remaining lower and upper class limits may be determined by adding the class interval size repeatedly till the largest value of the data is observed in the class.
5. **Distribute the data into respective classes:** All the observations are marked into respective classes by using Tally Bars (Tally Marks) methods which is suitable for tabulating the observations into respective classes. The number of tally bars is counted to get the frequency against each class. The frequency of all the classes is noted to get grouped data or frequency distribution of the data. The total of the frequency columns must be equal to the number of observations.

Example, **Construction of Frequency Distribution**
Construct a frequency distribution with suitable class interval size of marks obtained by 50 students of a class are given below:
23, 50, 38, 42, 63, 75, 12, 33, 26, 39, 35, 47, 43, 52, 56, 59, 64, 77, 15, 21, 51, 54, 72, 68, 36, 65, 52, 60, 27, 34, 47, 48, 55, 58, 59, 62, 51, 48, 50, 41, 57, 65, 54, 43, 56, 44, 30, 46, 67, 53.
Solution
Arrange the marks in ascending order as:
12, 15, 21, 23, 26, 27, 30, 33, 34, 35, 36, 38, 39, 41, 42, 43, 43, 44, 46, 47, 47, 48, 48, 50, 50, 51, 51, 52, 52, 53, 54, 54, 55, 56, 56, 57, 58, 59, 59, 60, 62, 63, 64, 65, 65, 67, 68, 72, 75, 77.
Minimum value = 12; Maximum value = 77
Range = Maximum value - Minimum value = 77 - 12 = 65
Number of classes = 1 + 3.322 log N
= 1 + 3.322 log 50
= 1 + 3.322 (1.69897)
= 1 + 5.64 = 6.64 or 7 approximate
class interval size (h) = Range/No. of classes = 65/7 = 9.3 or 10.

Marks Class Limits C.L.	Tally Marks	Number of Students *f*	Class Boundary C.B.	Class Marks x
10-19	II	2	9.5-19.5	10 + 19/2 = 14.5
20-29	IIII	4	19.5-29.5	20 + 29/2 = 24.5
30-39	~~IIII~~ II	7	29.5-39.5	30 + 39/2 = 34.5
40-49	~~IIII IIII~~	10	39.5-49.5	40 + 49/2 = 44.5
50-59	~~IIII IIII IIII~~ I	16	49.5-59.5	50 + 59/2 = 54.5
60-69	~~IIII~~ III	8	59.5-69.5	60 + 69/2 = 64.5
70-79	III	3	69.5-79.5	70 + 79/2 = 74.5
		50		

Note: For finding the class boundaries, we take half of the difference between lower class limit of the 2nd class and upper class limit of the 1st class 20 - 19/2 = 1/2 = 0.5 This value is subtracted from lower class limit and added in upper class limit to get the required class boundaries.

Frequency Distribution by Exclusive Method

Class Boundary C.B.	Tally Marks	Frequency *f*
10 - 19	II	2
20 - 29	IIII	4
30 - 39	~~IIII~~ II	7
40 - 49	~~IIII~~ ~~IIII~~	10
50-59	~~IIII~~ ~~IIII IIII~~ I	16
60 - 69	~~IIII~~ III	8
70 - 79	III	3
		50

2.4.2 CUMULATIVE FREQUENCY DISTRIBUTION

The total frequency of all classes less than the upper class boundary of a given class is called the cumulative frequency of the class. "A table showing the cumulative frequencies is calleda cumulative frequency distribution". There are two types of cumulative frequency distribution.

Less than cumulative frequency distribution

It is obtained by adding successively the frequencies of all the previous classes including the class against which it is written. The cumulate is started from the lowest to the highest size.

More than cumulative frequency distribution

It is obtained by finding the cumulative total of frequencies starting from the highest to the lowest class. The less than cumulative frequency distribution and more than cumulative frequency distribution for the frequency distribution given below are:

			Less than C.F.		More than C.F.	
Class Limit	***f***	**C.B.**	**Marks**	**C.F**	**Marks**	**C.F.**
10 - 19	2	9.5 - 19.5	Less than 19.5	2	9.5 or more	48 + 2 = 50
20 - 29	4	19.5 - 29.5	Less than 29.5	2 + 4 = 6	19.5 or more	44 + 4 = 48
30 – 39	7	29.5 - 39.5	Less than 39.5	6 + 7 = 13	29.5 or more	37 + 7 = 44
40 – 49	10	39.5 - 49.5	Less than 49.5	13 + 10 = 23	39.5 or more	27 + 10 = 37

50 – 59	16	49.5 - 59.5	Less than 59.5	23 + 16 = 39	49.5 or more	11 + 16 = 27
60 – 69	8	59.5 - 69.5	Less than 69.5	39 + 8 = 47	59.5 or more	3 + 8 = 11
70 – 79	3	69.5 - 79.5	Less than 79.5	47 + 3 = 50	69.5 or more	3

DIAGRAMS AND GRAPHS OF STATISTICAL DATA

We have discussed the techniques of classification and tabulation that help us in organising the collected data in a meaningful fashion. However, this way of presentation of statistical data does not always prove to be interesting to a layman. Too many figures are often confusingand fail to convey the message effectively.

One of the most effective and interesting alternative way in which a statistical data may be presented is through diagrams and graphs. There are several ways in which statistical data may be displayed pictorially such as different types of graphs and diagrams. The commonly used diagrams and graphs to be discussed in subsequent paragraphs are given as under:

2.5 TYPES OF DIAGRAMS/CHARTS

1. Simple Bar Chart
2. Multiple Bar Chart or Cluster Chart
3. Staked Bar Chart or Sub-Divided Bar Chart or Component Bar Chart

a. Simple Component Bar Chart

b. Percentage Component Bar Chart

c. Sub-Divided Rectangular Bar Chart

d. Pie Chart

4. Histogram
5. Frequency Curve and Polygon
6. Lorens Curve
7. Historigram

1. SIMPLE BAR CHART

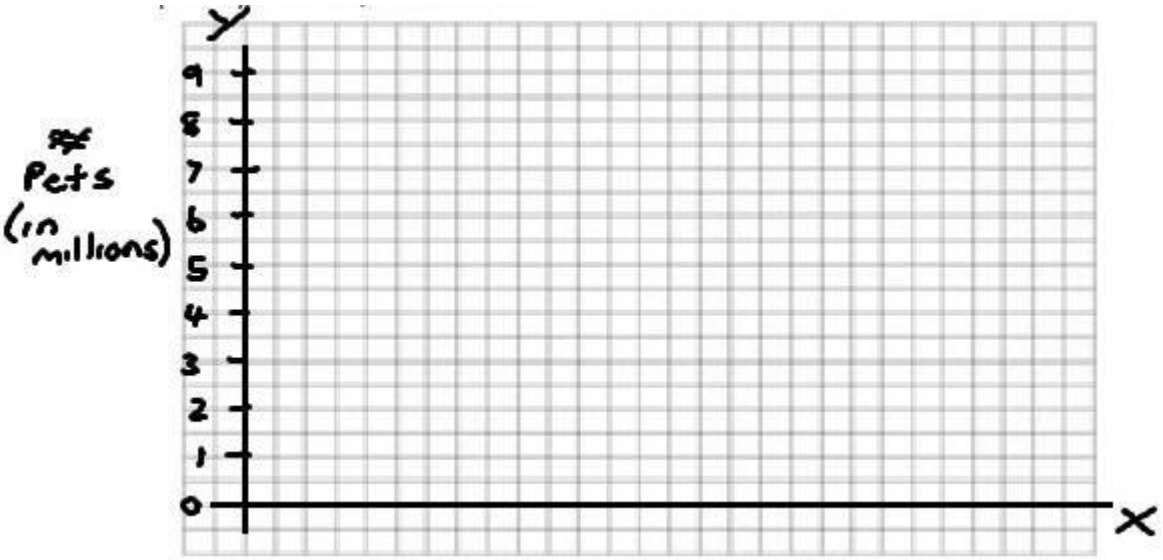
A simple bar chart is used to represent data involving only one variable classified on spatial, quantitative or temporal basis. In simple bar chart, we make bars of equal width but variable length, i.e. the magnitude of a quantity is represented by the height or length of the bars. Following steps are undertaken in drawing a simple bar diagram:

- Draw two perpendicular lines one horizontally and the other vertically at an appropriate place of the paper.
- Take the basis of classification along horizontal line (X-axis) and the observed variable along vertical line (Y-axis) or vice versa.
- Mark signs of equal breadth for each class and leave equal or not less than half breadth in between two classes.
- Finally mark the values of the given variable to prepare required bars.

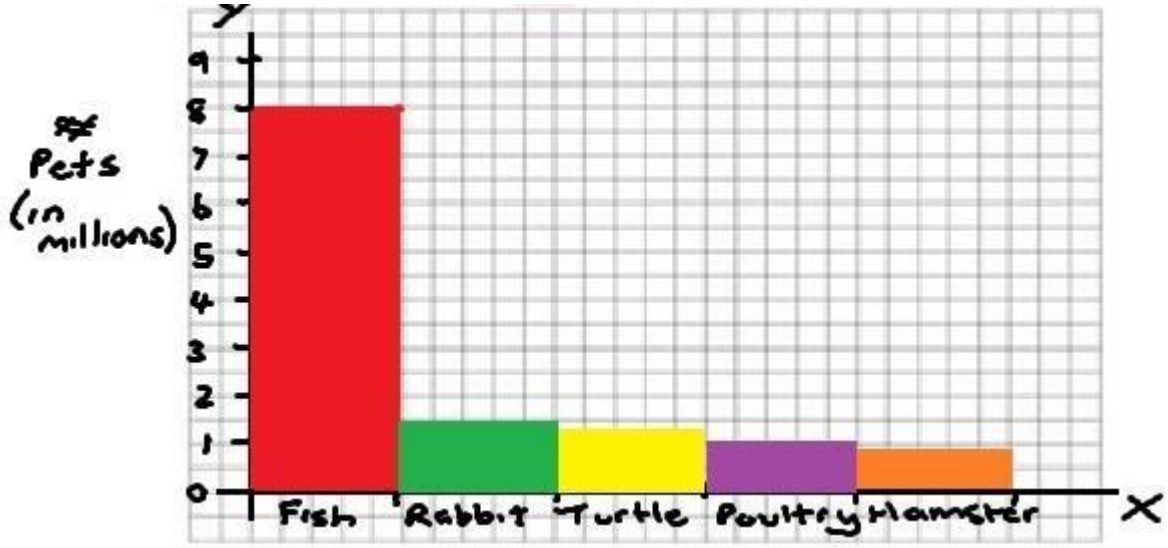
Sample problem: *Make a bar graph that represents exotic pet ownership in the United States. There are 8,000,000 fish, 1,500,000 rabbits, 1,300,000 turtles, 1,000,000 poultry and*

900,000 hamsters.

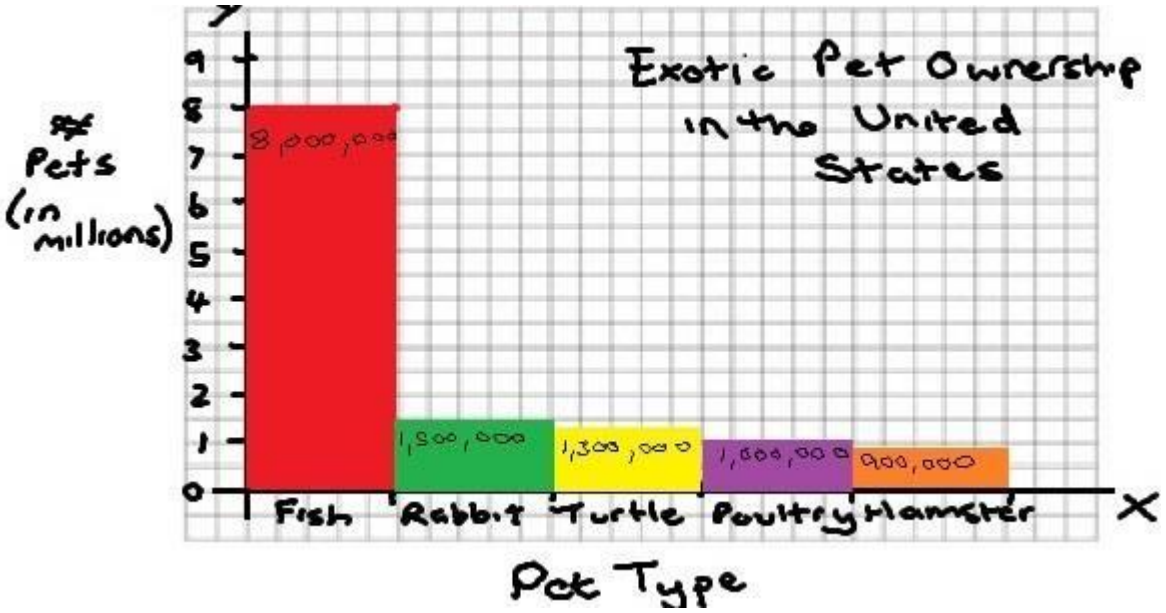
Step 1: **Number the Y-axis** with the dependent variable. The dependent variable is the one being tested in an experiment. In this sample question, the study wanted to know how many pets were in U.S. households. So the number of pets is the dependent variable. The highest number in the study is 8,000,000 and the lowest is 1,000,000 so it makes sense to label the Y- axis from 0 to 8.



Step 2: **Draw your bars.** The height of the bar should be even with the correct number on theY-axis. Don't forget to label each bar under the x-axis.



Step 3: **Label the X-axis** with what the bars represent. For this sample problem, label the x- axis "Pet Types" and then label the Y-axis with what the Y-axis represents: "Number of pets (per 1,000 households)." Finally, give your graph a name. For this sample problem, call the graph "Pet ownership (per 1,000 households).



Optional: In the above graph, I chose to write the actual numbers on the bars themselves. You don't have to do this, but if you have numbers than don't fall on a line (i.e. 900,000), then it can help make the graph clearer for a viewer.

Tips:

1. Line the numbers up on the lines of the graph paper, not the spaces.
2. Make all your bars the same width.

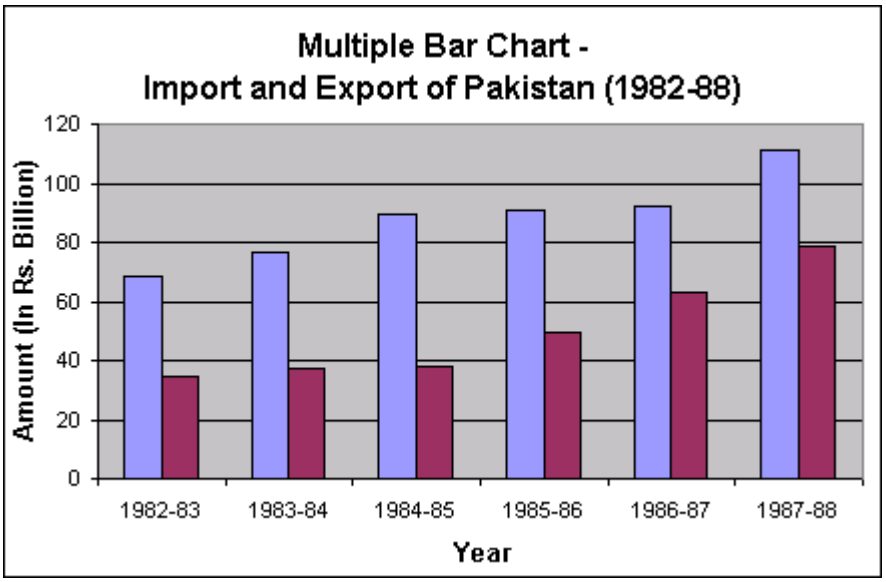
2. MULTIPLE BAR CHART

By multiple bars diagram two or more sets of inter related data are represented (multiple bar diagram facilities comparison between more than one phenomena). The technique of simple bar chart is used to draw this diagram but the difference is that we use different shades, colours or dots to distinguish between different phenomena. We use to draw multiple bar charts if the total of different phenomena is meaningless.

Sample Example

Draw a multiple bar chart to represent the import and export of Pakistan for the years 1982- 1988.

Years	Imports Rs. (billion)	Exports Rs. (billion)
1982-83	68.15	34.44
1983-84	76.71	37.33
1984-85	89.78	37.98
1985-86	90.95	49.59
1986-87	92.43	63.35
1987-88	111.38	78.44



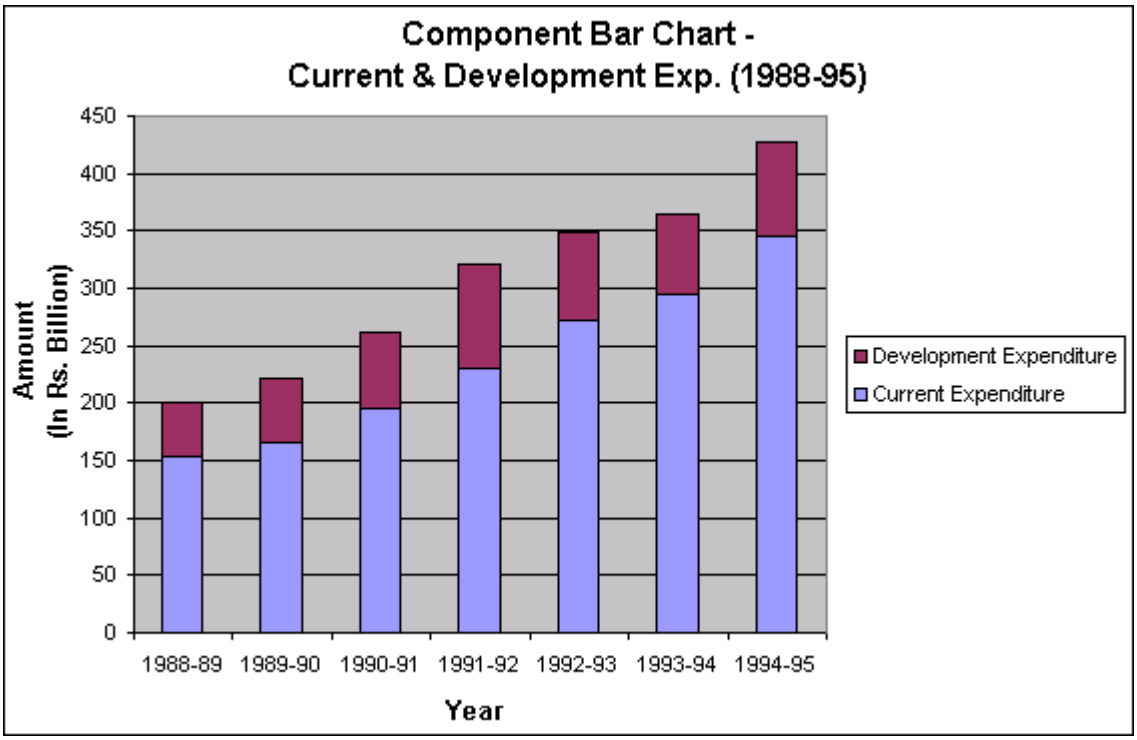
3. COMPONENT BAR CHART

Sub-divided or component bar chart is used to represent data in which the total magnitude is divided into different components.

In this diagram, first we make simple bars for each class taking total magnitude in that class and then divide these simple bars into parts in the ratio of various components. This type of diagram shows the variation in different components without each class as well as between different classes. Sub-divided bar diagram is also known as component bar chart or staked chart.

Current and Development Expenditure – Pakistan (All figures in Rs. Billion)

Years	Current Expenditure	Development Expenditure	Total Expenditure
1988-89	153	48	201
1989-90	166	56	222
1990-91	196	65	261
1991-92	230	91	321
1992-93	272	76	348
1993-94	294	71	365
1994-95	346	82	428



3. b. PERCENTAGE COMPONENT BAR CHART

Sub-divided bar chart may be drawn on percentage basis. to draw sub-divided bar chart on percentage basis, we express each component as the percentage of its respective total. In drawing percentage bar chart, bars of length equal to 100 for each class are drawn at first stepand sub-divided in the proportion of the percentage of their component in the second step. The diagram so obtained is called percentage component bar chart or percentage staked bar chart. This type of chart is useful to make comparison in components holding the difference of total constant.

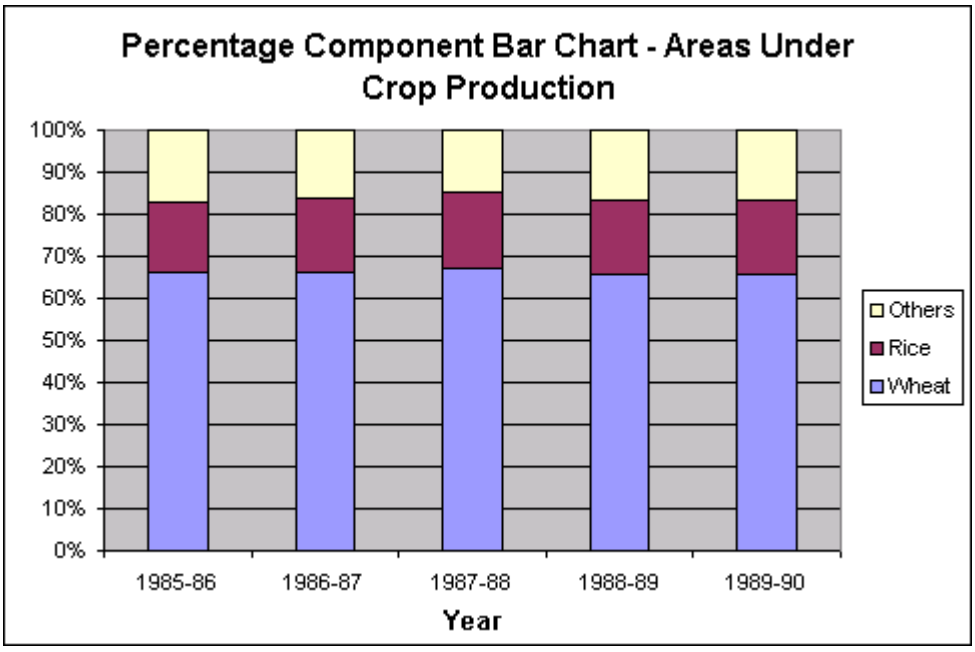
Areas Under Crop Production (1985-90)

('000 hectors)

Year	Wheat	Rice	Others	Total
1985-86	7403	1863	1926	11192
1986-87	7706	2066	1906	11678
1987-88	7308	1963	1612	10883
1988-89	7730	2042	1966	11738
1989-90	7759	2107	1970	11836

Percentage Areas Under Production

Year	Wheat	Rice	Others	Total
1985-86	66.2%	16.6%	17.2%	100%
1986-87	66.0	17.7	16.3	100
1987-88	67.2	18.0	14.8	100
1988-89	65.9	17.4	16.7	100
1989-90	65.6	17.8	16.6	100



3. d. PIE-CHART

Pie chart can be used to compare the relation between the whole and its components. Pie chart is a circular diagram and the area of the sector of a circle is used in pie chart. Circles aredrawn with radii proportional to the square root of the quantities because the area of a circleis $A = 2\pi r^2$

To construct a pie chart (sector diagram), we draw a circle with radius (square root of the total). The total angle of the circle is 360°. The angles of each component are calculated by the formula:

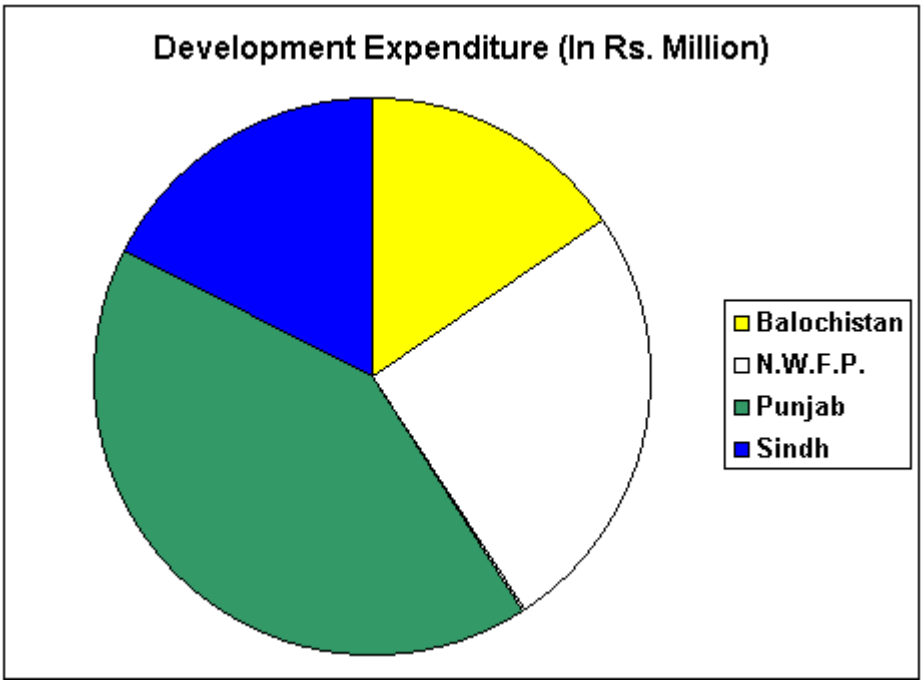
Angle of Sector = $\frac{Component\ Part}{Total} \times 360°$

These angles are made in the circle by means of a protractor to show differentcomponents. The arrangement of the sectors is usually anti-clock wise.

Example

Development Expenditure (1994-95)

Provinces	Development Expenditure (In Rs. Million)	Angles of Sectors (In Degrees)	Cumulative Angle
Balochistan	4874	$\frac{4874}{31189}\times 360^0 = 56^0$	56º
N.W.F.P.	7861	$\frac{7861}{31189}\times 360^0 = 91^0$	147º
Punjab	12954	$\frac{12954}{31189}\times 360^0 = 150^0$	297º
Sindh	5500	$\frac{5500}{31189}\times 360^0 = 63^0$	360º
Total	31189	360º	



2.6 EXERCISES

1. Draw a histogram of the following data:

Weekly Wages	1 - 10	11 - 20	21 - 30	31 - 40	41 - 50
No. of Workers	14	28	36	12	10

2. The following table shows the temperature for the consecutive five days in a particular week. Draw range graph.

Day	M	T	W	Th	F
High° C	40	35	50	60	25
Low° C	25	20	40	55	15

3. The following is the distribution of total house hold expenditure (in Rs.) of 202 workers in a city.

Expenditure in Rs.	100 - 150	150 - 200	200 - 250	250 - 300
No. of Workers	25	40	33	28
Expenditure in Rs.	300 - 350	350 - 400	400 - 450	450 - 500
No. of Workers	30	22	16	8

3

MEASURES OF DISPERSION

Unit Structure

3.1 INTRODUCTION TO MEASURE OF DISPERSION

A modern student of statistics is mainly interested in the study of variability and uncertainty. We live in a changing world. Changes are taking place in every sphere of life. A man of Statistics does not show much interest in those things which are constant. The total area ofthe earth may not be very important to a research minded person but the area under different crops, areas covered by forests, area covered by residential and commercial buildings are figures of great importance because these figures keep on changing from time to time and from place to place. Very large number of experts is engaged in the study of changing phenomenon. Experts working in different countries of the world keep a watch on forces which are responsible for bringing changes in the fields of human interest. The agricultural, industrial and mineral production and their transportation from one part to the other parts of the world are the matters of great interest to the economists, statisticians and other experts. The changes in human population, the changes in standard of living, and changes in literacy rate and the changes in price attract the experts to make detailed studies about them and then correlate these changes with the human life. Thus variability or variation is somethingconnected with human life and study is very important for mankind.

3.1.1 DISPERSION

The word dispersion has a technical meaning in Statistics. The average measures the centre ofthe data. It is one aspect observations. Another feature of the observations is as to how the observations are spread about the centre. The observation may be close to the centre or they may be spread away from the centre. If the observation are close to the centre (usually the arithmetic mean or median), we say that dispersion, scatter or variation is small. If the observations are spread away from the centre, we say dispersion is large. Suppose we have three groups of students who have obtained the following marks in a test. The arithmetic means of the three groups are also given below:

Group A: 46, 48, 50, 52, 54 $\quad \bar{X}_A = 50$
Group B: 30, 40, 50, 60, 70 $\quad \bar{X}_B = 50$
Group C; 40, 50, 60, 70, 80 $\quad \bar{X}_C = 60$

In a group A and B arithmetic means are equal i.e. $\bar{X}_A = \bar{X}_B = 50$. But in group A the observations are concentrated on the centre. All students of group A have almost the same level of performance. We say that there is consistence in the observations in group A. In group B the mean is 50 but the observations are not close to the centre. One observation is as small as 30 and one observation is as large as 70. Thus, there is greater dispersion in group B.In group C the mean is 60 but the spread of the observations with respect to the centre 60 is the same as the spread of the observations in group B with respect to their own centre whichis 50. Thus in group B and C the means are different but their dispersion is the same. In groupA and C the means are different and their dispersions are also different. Dispersion is an important feature of the observations and it is measured with the help of the measures of dispersion, scatter or variation. The word variability is also used for this idea of dispersion.

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

1. The degree to which numerical data tend to spread about an average value is called the dispersion or variation of the data.
2. Dispersion or variation may be defined as a statistics signifying the extent of the scattered items around a measure of central tendency.
3. Dispersion or variation is the measurement of the scattered size of the items of a series about the average.

For the study of dispersion, we need some measures which show whether the dispersion is small or large. There are two types of measures of dispersion, which are:

a. Absolute Measure of Dispersion
b. Relative Measure of Dispersion.

3.2 ABSOLUTE MEASURE OF DISPERSION

These measures give us an idea about the amount of dispersion in a set of observations. They give the answers in the same units as the units of the original observations. When the observations are in kilograms, the absolute measure is also in kilograms. If we have two sets of observations, we cannot always use the absolute measures to compare their dispersion. We shall explain later as to when the absolute measures can be used for comparison of dispersions in two or more than two sets of data. The absolute measures which are commonlyused are:

1. The Range
2. The Quartile Deviation
3. The Mean Deviation
4. The Standard Deviation and Variance

3.3 RELATIVE MEASURE OF DISPERSION

These measures are calculated for the comparison of dispersion in two or more than two sets of observations. These measures are free of the units in which the original data is measured.If the original data is in dollar or kilometers, we do not use these units with relative measure of dispersion. These measures are a sort of ratio and are called coefficients. Each absolute measure of dispersion can be converted into its relative measure.

Thus, the relative measures of dispersion are:

1. Coefficient of Range or Coefficient of Dispersion.
2. Coefficient of Quartile Deviation or Quartile Coefficient of Dispersion.
3. Coefficient of Mean Deviation or Mean Deviation of Dispersion.
4. Coefficient of Standard Deviation or Standard Coefficient of Dispersion.
5. Coefficient of Variation (a special case of Standard Coefficient of Dispersion).

3.4 RANGE AND COEFFICIENT OF RANGE

The Range

Range is defined as the difference between the maximum and the minimum observation of the given data. If X_m denotes the maximum observation X_o denotes the minimum observation then the range is defined as Range = X_m - X_o.

In case of grouped data, the range is the difference between the upper boundary of the highest class and the lower boundary of the lowest class. It is also calculated by using the difference between the mid points of the highest class and the lowest class. it is the simplest measure of dispersion. It gives a general idea about the total spread of the observations. It does not enjoy any prominent place in statistical theory. But it has its application and utility in quality control methods which are used to maintain the quality of the products produced in factories. The quality of products is to be kept within certain range of values.

The range is based on the two extreme observations. It gives no weight to the central values of the data. It is a poor measure of dispersion and does not give a good picture of the overall spread of the observations with respect to the centre of the observations. Let us consider three groups of the data which have the same range:

Group A: 30, 40, 40, 40, 40, 40, 50

Group B: 30, 30, 30, 40, 50, 50, 50

Group C: 30, 35, 40, 40, 40, 45, 50

In all the three groups the range is 50 - 30 = 20. In group A there is concentration of observations in the centre. In group B the observations are friendly with the extreme corner and in group C the observations are almost equally distributed in the interval from 30 to 50. The range fails to explain these differences in the three groups of data. This defect in range cannot be removed even if we calculate the coefficient of range which is a relative measure of dispersion. If we calculate the range of a sample, we cannot draw any inferences about the range of the population.

Coefficient of Range

It is relative measure of dispersion and is based on the value of range. It is also called range coefficient of dispersion. It is defined as:

Coefficient of Range = $\frac{X_m - X_o}{X_m + X_o}$.

The range X_m - X_o is standardised by the total X_m + X_o.

Let us take two sets of observations. Set A contains marks of five students in Mathematics out of 25 marks and group B contains marks of the same student in English out of 100 marks. Set A: 10, 15, 18, 20, 20

Set B: 30, 35, 40, 45, 50

The values of range and coefficient of range are calculated as

	Range	Coefficient of Range
Set A: (Mathematics)	20 - 10 = 10	$\frac{20-10}{20+10} = 0.33$
Set B: (English)	50 - 30 = 20	$\frac{50-30}{50+30} = 0.25$

In set A the range is 10 and in set B the range is 20. Apparently it seems as if there is greater dispersion in set B. But this is not true. The range of 20 in set B is for large observations and the range of 10 in set A is for small observations. Thus 20 and 10 cannot be compared

directly. Their base is not the same. Marks in Mathematics are out of 25 and marks of Englishare out of 100. Thus, it makes no sense to compare 10 with 20. When we convert these two values into coefficient of range, we see that coefficient of range for set A is greater than that of set B. Thus, there is greater dispersion or variation in set A. The marks of students in English are more stable than their marks in Mathematics.

Example

Following are the wages of 8 workers of a factory. Find the range and coefficient of range.
Wages in ($) 1400, 1450, 1520, 1380, 1485, 1495, 1575, 1440.

Solution

Here Largest Value = X_m = 1575 and Smallest Value = X_o = 1380

Range = X_m - X_o = 1575 - 1380 = 195.

Coefficient of Range = $\frac{X_m - X_o}{X_m + X_o} = \frac{1575 - 1380}{1575 + 1380} = \frac{195}{2955} = 0.66$

Example

The following distribution gives the numbers of houses and the number of persons per house.

Number of Persons	1	2	3	4	5	6	7	8	9	10
Number of Houses	26	113	120	95	60	42	21	14	5	4

Calculate the range and coefficient of range.

Solution

Here Largest Value = X_m = 10 and Smallest Value = X_o = 1Range =

X_m - X_o = 10 - 1 = 9.

Coefficient of Range = $\frac{X_m - X_o}{X_m + X_o} = \frac{10 - 1}{10 + 1} = \frac{9}{11} = 0.818$

Example

Find the range of the weight of the students of a University.

Weights (Kg)	60-62	63-65	66-68	69-71	72-74
Number of Students	5	18	42	27	8

Calculate the range and coefficient of range.

Solution

Weights (Kg)	Class Boundaries	Mid Value	No. of Students
60-62	59.5 - 62.5	61	5
63-65	62.5 - 65.5	64	18
66-68	65.5 - 68.5	67	42
69-71	68.5 - 71.5	70	27
72-74	71.5 - 74.5	73	8

Method 1

Here X_m = Upper class boundary of the highest class = 74.5; X_o = Lower Class Boundary ofthe lowest class = 59.5

Range = X_m - X_o = 74.5 - 59.5 = 15 Kilogram.

Coefficient of Range = $\frac{X_m - X_o}{X_m + X_o} = \frac{74.5 - 59.5}{74.5 + 59.5} = \frac{15}{134} = 0.1119.$

Method 2

Here X_m = Mid value of the highest class = 73; X_o = Mid Value of the lowest class = 61Range = $X_m - X_o = 73 - 61 = 12$ Kilogram.

Coefficient of Range = $\frac{Xm - Xo}{Xm + Xo} = \frac{73 - 61}{73 + 61} = \frac{12}{134}$ = 0.0895.

3.5 QUARTILE DEVIATION AND ITS COEFFICIENT

Quartile Deviation

It is based on the lower Quartile Q1 and the upper quartile Q3. The difference Q3 - Q1 is called the inter quartile range. The difference Q3 - Q1 divided by 2 is called semi-inter- quartile range or the quartile deviation. Thus

Quartile Deviation (Q.D) = $\frac{Q3-Q1}{2}$. The quartile deviation is a slightly better measure of absolute dispersion than the range. But it ignores the observation on the tails. If we take different samples from a population and calculate their quartile deviations, their values are quite likely to be sufficiently different. This is called sampling fluctuation. It is not a popular measure of dispersion. The quartile deviation calculated from the sample data does not help us to draw any conclusion (inference) about the quartile deviation in the population.

Coefficient of Quartile Deviation

A relative measure of dispersion based on the quartile deviation is called the coefficient of quartile deviation. It is defined as

Coefficient of Quartile Deviation = $\frac{\frac{Q3-Q1}{2}}{\frac{Q3+Q1}{2}} = \frac{Q3-Q1}{Q3+Q1}$.

It is pure number free of any units of measurement. It can be sued for comparing the dispersion in two or more than two sets of data.

Example

The Wheat production (in Kg) of 20 acres is given as: 1120, 1240, 1320, 1040, 1080, 1200, 1440, 1360, 1680, 1730, 1785, 1342, 1960, 1880, 1755, 1600, 1470, 1750 and 1885. Find the quartile deviation and coefficient of quartile deviation.

Solution

After arranging the observation in ascending order, we get, 1040, 1080, 1120, 1200, 1240, 1320, 1342, 1360, 1440, 1470, 1600, 1680, 1720, 1730, 1750, 1755, 1785, 1880, 1885, 1960.

Q1 = Value of $\left(\frac{n+1}{4}\right)$th item

= Value of $\left(\frac{20+1}{4}\right)$ th item

= Value of (5.25)th item

= 5th item + 0.25 (6th item - 5th item) = 1240 + 0.25 (1320 - 1240)Q1

= 1240 + 20 = 1260

Q3 = Value of $\frac{3\,(n+1)}{4}$ th item

= Value of $\frac{3\,(20+1)}{4}$ th item

= Value od (15.75) th item

15th item + 0.75 (16th item - 15th item) = 1750 + 0.75 (1755 - 1750)

Q3 = 1750 + 3.75 = 1753.75

Quartile Deviation (QD) = $\frac{Q3-Q1}{2} = \frac{1753.75-1260}{2} = \frac{492.75}{2} = 246.875$

Coefficient of Quartile Deviation = $\frac{Q3-Q1}{Q3+Q1} = \frac{1753.75-1260}{1753.75+1260} = 0.164.$

Example

Calculate the quartile deviation and coefficient of quartile deviation from the data given below:

Maximum Load (Short tons)	Number of Cables
9.3 - 9.7	2
9.8 - 10.2	5
10.3 - 10.7	12
10.8 - 11.2	17
11.3 - 11.7	14
11.8 - 12.2	6
12.3 - 12.7	3
12.8 - 13.2	1

Solution

The necessary calculations are given below:

Maximum Load (Short Tons)	Number of Cables F	Class Boundaries	Cumulative Frequencies
9.3 - 9.7	2	9.25 - 9.75	2
9.8 - 10.2	5	9.75 - 10.25	2 + 5 = 7
10.3 - 10.7	12	10.25 - 10.75	7 + 12 = 19
10.8 - 11.2	17	10.75 - 11.25	19 + 17 = 36
11.3 - 11.7	14	11.25 - 11.75	36 + 14 = 50
11.8 - 12.2	6	11.75 - 12. 25	50 + 6 = 56
12.3 - 12.7	3	12.25 - 12. 75	56 + 3 = 59
12.8 - 13.2	1	12. 75 - 13.25	59 + 1 = 60

Q_1 = Value of [ ] th item
4

= Value of [60] th item
4

= 15th item

Q_1 lies in the class 10.25 - 10.75

$\therefore Q_1 = 1 + \frac{h}{f}\left[\frac{n}{4} - c\right]$

Where 1 = 10.25, h = 0.5, f = 12, n/4 = 15 and c = 7

$Q_1 = 10.25 + \frac{0.25}{12}(15 - 7)$

= 10.25 + 0.33

= 10.58

Q_1 = Value of [$\frac{3n}{4}$] th item

= value of [$\frac{3 \times 60}{4}$] th item

= 45th item

Q_3 lies in the class 11.25 - 11.75

$\therefore Q_3 = 1 + \frac{h}{f}\left[\frac{3n}{4} - c\right]$

where 1 = 11.25, h = 0.5, f = 14, 3n/4 = 45 and c = 36

$\therefore Q_1 = 11.25 + \frac{0.5}{12}(45 - 36)$

= 11.25 + 0.32

= 11.57

Quartile Deviation (Q.D) = $\frac{Q3-Q1}{2}$

$= \frac{11.57-10.58}{2}$

$= \frac{0.99}{2} = 0.495$

Coefficient of Quartile Deviation = $\frac{Q3-Q1}{Q3+Q1}$

$= \frac{11.57-10.58}{11.57+10.58} = \frac{0.99}{22.15} = 0.045$

3.6 THE MEAN DEVIATION

The mean deviation or the average deviation is defined as the mean of the absolute deviations of observations from some suitable average which may be arithmetic mean, the median or the mode. The difference (X - average) is called deviation and when we ignore the negative sign, this deviation is written as $|X - average|$ and is read as mod deviations. The mean of these more or absolute deviations is called the mean deviation or the mean absolute deviation. Thus for sample data in which the suitable average is the $\bar{X}$, the mean deviation (M.D) is given by the relation

$$M.D = \frac{\Sigma|X-\bar{X}|}{n}$$

For frequency distribution, the mean deviation is given by

$$M.D = \frac{\Sigma f|X-\bar{X}|}{\Sigma f}$$

When the mean deviation is calculated about the median, the formula becomes

$$M.D. \text{ (about median)} = \frac{\Sigma f|X-Median|}{\Sigma f}$$

The mean deviation about the mode is

$$M.D \text{ (about mode)} = \frac{\Sigma f|X-Median|}{\Sigma f}$$

For a population data the mean deviation about the population mean μ is

$$M.D = \frac{\Sigma f|X-\mu|}{\Sigma f}$$

The mean deviation is a better measure of absolute dispersion than the range and the quartile deviation.

A drawback in the mean deviation is that we use the absolute deviations $|X - average|$ which does not seem logical. The reason for this is that Σ (X - $\bar{X}$) is always equal to zero. Even if we use median or more in place of $\bar{X}$ even then the summation Σ (X - median) or Σ (X - mode) will be zero or approximately zero with the result that the mean deviation would
always be better either zero or close to zero. Thus, the very definition of the mean deviationis possible only on the absolute deviations.

The mean deviation is based on all the observations, a property which is not possessed by the range and the quartile deviation. The formula of the mean deviation gives a mathematical impression that it is a better way of measuring the variation in the data. Any suitable average among the mean, median or more can be used in its calculation but the value of the mean deviation is minimum if the deviations are taken from the median. A drawback of the mean deviation is that it cannot be used in statistical inference.

Coefficient of the Mean Deviation

A relative measure of dispersion based on the mean deviation is called the coefficient of the mean deviation or the coefficient of dispersion. It is defined as the ratio of the mean deviationto the average used in the calculation of the mean deviation.
Thus,
Coefficient of M.D (about mean) = Mean Deviation from Mean/Mean
Coefficient of M.D (about median) = Mean Deviation from Median/Median
Coefficient of M.D (about mode) = Mean Deviation from Mode/Mode

Example

Calculate the mean deviation from (1) Arithmetic Mean (2) Median (3) Mode in respect of the marks obtained by nine students given below and show that the mean deviation from median is minimum.
Marks out of 25: 7, 4, 10, 9, 15, 12, 7, 9, 7

Solution

After arranging the observations in ascending order, we getMarks 4,
7, 7, 7, 9, 9, 10, 12, 15
Mean = $\frac{\Sigma X}{n} = \frac{80}{9} = 8.89$
Median = Value of $(\frac{n+1}{2})$ th item
= Value of $(\frac{9+1}{2})$ th item
= Value of (5) the item = 9
Mode = 7 (Since 7 is repeated maximum number of times)

Marks X	$\|X - \bar{X}\|$	$\|X - median\|$	$\|X - mode\|$
4	4.89	5	3
7	1.89	2	0
7	1.89	2	0
7	1.89	2	0
9	0.11	0	2
9	0.11	0	2
10	1.11	1	3
12	3.11	3	5
15	6.11	6	8
Total	21.11	21	23

M.D from mean = $\frac{\Sigma f|X-\bar{X}|}{n}$

$= \frac{21.11}{9} = 2.35$

M.D from Median = $\frac{\Sigma|X-Median\ | = 21}{n} = \frac{}{9} 2.33$

M.D from Mode = $\frac{\Sigma f|X-Mode,| = 23}{n} = \frac{}{9} 2.56$

From the above calculations, it is clear that the mean deviation from the median has the least value.

Example

Calculate the mean deviation from mean and its coefficients from the following data:

Size of items	3 - 4	4 - 5	5 - 6	6 - 7	7 - 8	8 - 9	9 - 10
Frequency	3	7	22	60	85	32	8

Solution

The necessary calculation is given below:

Size of Items	X	F	fX	$\|X - \bar{X}\|$	f $\|X - \bar{X}\|$
3 - 4	3.5	3	10.5	3.59	10.77
4 - 5	4.5	7	31.5	2.59	18.13
5 - 6	5.5	22	121.0	1.59	34.98
6- 7	6.5	60	390.0	0.59	35.40
7 - 8	7.5	85	637.5	0.41	34.85
8 - 9	8.5	32	272.0	1.41	45.12
9 - 10	9.5	8	76.0	2.41	19.28
Total		217	1538.5		198.53

Mean = $\bar{X} = \frac{\Sigma fX}{\Sigma f} \frac{1538.5}{217} = 7.09$

M.D from Mean = $\frac{\Sigma|X-\bar{X}| = 198.53}{n} \frac{}{217} = 0.915$

Coefficient of M.D (Mean) = $\frac{M.D\ from\ mean\ = 0.915}{Mean} \frac{}{7.09} = 0.129$

3.7 Standard Deviation

The standard deviation is defined as the positive square root of the mean of the square deviations taken from arithmetic mean of the data.

For the sample data the standard deviation is denoted by S and is defined as

$$S_n = \sqrt{\frac{\Sigma (X-\bar{X})^2}{n}}$$

For a population data the standard deviation is denoted by σ (sigma) and is defined as:

$$\sigma_N = \sqrt{\frac{\Sigma (X-\mu)^2}{N}}$$

For frequency distribution the formulas become

$$S_{\Sigma f} = \sqrt{\frac{\Sigma f (X-\bar{X})^2}{\Sigma f}} \quad \text{or} \quad \sigma = \sqrt{\frac{\Sigma f (X-\mu)^2}{\Sigma f}}$$

The standard deviation is in the same units as the units of the original observations. If the original observations are in grams, the value of the standard deviation will also be in grams.

The standard deviation plays a dominating role for the study of variation in the data. It is a very widely used measure of dispersion. It stands like a tower among measure of dispersion. As far as the important statistical tools are concerned, the first important tool is the mean $\bar{X}$ and the second important tool is the standard deviation S. It is based on the observations and is subject to mathematical treatment. It is of great importance for the analysis of data and for the various statistical inferences.

However, some alternative methods are also available to compute standard deviation. The alternative methods simplify the computation. Moreover in discussing, these methods we will confirm ourselves only to sample data because sample data rather than whole population confront mostly a statistician.

Actual Mean Method

In applying this method first of all we compute arithmetic mean of the given data either ungroup or grouped data. Then take the deviation from the actual mean. This method is already is defined above. The following formulas are applied:

For Ungrouped Data	For Grouped Data
$S_n = \sqrt{\frac{\Sigma (X-\bar{X})^2}{n}}$	$S_{\Sigma f} = \sqrt{\frac{\Sigma f (X-\bar{X})^2}{\Sigma f}}$

This method is also known as **direct method.**

Assumed Mean Method

a. We use the following formulas to calculate standard deviation:

For Ungrouped Data	For Grouped Data
$S_n = \sqrt{\frac{\Sigma D^2}{n} - \left(\frac{\Sigma D}{n}\right)^2}$	$S_{\Sigma f} = \sqrt{\frac{\Sigma f D^2}{\Sigma f} - \left(\frac{\Sigma f D}{\Sigma f}\right)^2}$

where D = X - A and A is any assumed mean other than zero. This method is also known as short-cut method.

b. If A is considered to be zero then the above formulas are reduced to the following formulas:

For Ungrouped Data	For Grouped Data
$S_n = \sqrt{\frac{\Sigma X^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$	$S_{\Sigma f} = \sqrt{\frac{\Sigma f X^2}{\Sigma f} - \left(\frac{\Sigma f X}{\Sigma f}\right)^2}$

c. If we are in a position to simplify the calculation by taking some common factor or divisor from the given data the formulas for computing standard deviation are:

For Ungrouped Data	For Grouped Data
$S = \sqrt{\frac{\Sigma u2}{n}} - (\frac{\Sigma u}{n})2$ X c	$S = \sqrt{\frac{\Sigma fu2}{\Sigma f}} - (\frac{\Sigma fu}{\Sigma f})$ 2 X c or h

Where $u = \frac{X-A}{h \text{ or } c} = \frac{D}{h \text{ or } c}$; h = Class Interval and c = Common Divisor. This method is also called method of step-deviation.

Examples of Standard Deviation

This tutorial is about some examples of standard deviation using all methods which are discussed in the previous tutorial.

Example

Calculate the standard deviation for the following sample data using all methods: 2, 4, 8, 6,10 and 12.

Solution:

Method - 1 Actual mean Method

X	**$(X - \bar{X})^2$**
2	$(2 - 7)^2 = 25$
4	$(4 - 7)^2 = 9$
8	$(8 - 7)^2 = 1$
6	$(6 - 7)^2 = 1$
10	$(10 - 7)^2 = 9$
12	$(12 - 7)^2 = 25$
$\Sigma X = 42$	**$\Sigma(X - \bar{X})^2 = 70$**

$\bar{X} = \frac{\Sigma X}{n} = \frac{42}{6} = 7$

$S = \sqrt{\frac{\Sigma (X-\bar{X})2}{n}}$

$S = \sqrt{\frac{70}{6}} = \sqrt{35} = 3.42$

Method 2: Taking assumed mean as 6.

X	**D = (X - 6)**	**D^2**
2	- 4	16
4	- 2	4
8	2	4
6	0	0
10	4	16
12	6	36
Total	**$\Sigma D = 6$**	**$\Sigma D^2 = 76$**

$S_n = \sqrt{\frac{\Sigma D2}{n} - \left(\frac{\Sigma D}{n}\right)^2}$

$S_6 = \sqrt{\frac{76}{6} - \left(\frac{6}{6}\right)^2} = \sqrt{\frac{70}{6}} = \frac{\sqrt{35}}{3} = 3.42$

Method 3: Taking Assumed Mean as Zero

X	**X^2**
2	4
4	16
8	64
6	36
10	100
12	144
ΣX = 42	**Σ X^2= 364**

$S_n = \sqrt{\frac{\Sigma X2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$

$S_6 = \sqrt{\frac{364}{6} - \left(\frac{42}{6}\right)^2}$

$S_6 = \sqrt{70} = \sqrt{35} = 3.42$

Method 4: Taking 2 as common divisor or factor

X	**u = (X - 4)/2**	**u^2**
2	- 1	1
4	0	0
8	2	4
6	1	1
10	3	9
12	4	16
Total	**Σu = 9**	**Σ u^2 = 31**

$S_n = \sqrt{\frac{\Sigma u2}{n} - \left(\frac{\Sigma u}{n}\right)2} \text{ X c}$

$S_6 = \sqrt{\frac{31}{6} - \left(\frac{9}{6}\right)2} \text{ X 2}$

$S = \sqrt{2.92} \text{ X 2} = 3.42.$

Example

Calculate standard deviation from the following distribution of marks by using all themethods:

Marks	**No. of Students**
1 - 3	40
3 - 5	30
5 - 7	20
7 - 9	10

Solution

Method 1: Actual mean method

Marks	f	X	fX	$(X - \bar{X})^2$	$f(X - \bar{X})^2$
1 - 3	40	2	80	4	160
3 - 5	30	4	120	0	0
5 - 7	20	6	120	4	80
7 - 9	10	8	80	16	160
Total	**100**		**400**		**400**

$\bar{X} = \frac{\Sigma}{\Sigma f} = \frac{400}{100} = 4$

$S_{\Sigma f} = \sqrt{\frac{\Sigma f (X - \bar{X})^2}{}}$

$S_{100} = \sqrt{\frac{400}{100}} = \sqrt{4} = 2$ Marks

Method 2: Taking assumed mean as 2

Marks	f	X	D = (X - 2)	fD	fD^2
1 - 3	40	2	0	0	0
3 - 5	30	4	2	60	120
5 - 7	20	6	4	80	320
7 - 9	10	8	6	60	160
Total	**100**			**200**	**800**

$S_{\Sigma f} = \sqrt{\frac{\Sigma fD^2}{} - \left(\frac{\Sigma fD}{\Sigma f}\right)^2}$

$S_{100} = \sqrt{\frac{800}{} - \left(\frac{200}{100}\right)^2}$

$S = \sqrt{8 - 4} = \sqrt{4} = 2$ Marks

Method 3: Using Assumed Mean as Zero

Marks	f	X	fX	fX^2
1 - 3	40	2	80	160
3 - 5	30	4	120	480
5 - 7	20	6	120	720
7 - 9	10	8	80	640
Total	**100**		**400**	**2000**

$S_{\Sigma f} = \sqrt{\frac{\Sigma fX^2}{} - \left(\frac{\Sigma fX}{\Sigma f}\right)^2}$

$S_{100} = \sqrt{\frac{2000}{} - \left(\frac{400}{100}\right)^2}$

$S = \sqrt{20 - 16} = \sqrt{4} = 2$ marks.

Method 4: By taking 2 as the Common Divisor

Marks	f	X	u = (X - 2)/2	Fu	fu^2
1 - 3	40	2	- 2	- 80	160
3 - 5	30	4	- 1	- 30	30
5 - 7	20	6	0	0	0
7 - 9	10	8	1	10	10
Total	**100**			**- 100**	**200**

$S = \sqrt{\frac{\Sigma fu2}{\Sigma f} - \left(\frac{\Sigma fu}{\Sigma f}\right)^2} X\ h$

$S = \sqrt{\frac{200}{100} - \left(\frac{-100}{100}\right)^2} X\ 2$

$S = \sqrt{2 - 1}\ X\ 2 = \sqrt{1}\ X\ 2 = 2$ marks.

3.7. 1 Coefficient of Standard Deviation

The standard deviation is the absolute measure of dispersion. Its relative measure is called standard coefficient of dispersion or coefficient of standard deviation, It is defined as

Coefficient of Standard Deviation = $\frac{S}{\bar{X}}$

3.8 Coefficient of Variation

The most important of all the relative measure of dispersion is the coefficient of variation. This word is variation and not variance. There is no such thing as coefficient of variance. Thecoefficient of variation (CV) is defined as

Coefficient of Variation (C.V) = $\frac{S}{\bar{X}}$ x 100

Thus C.V is the value of S when $\bar{X}$is assumed equal to 100. It is a pure number and the unit of observations is not mentioned with its value. It is written in percentage form like 20% or 25%. When its value is 20%, it means that when the mean of the observation is assumed equal to 100, their standard deviation will be 20. The C.V is used to compare the dispersionin different sets of data particularly the data which differ in their means or differ in the unitsof measurement. The wages of workers may be in dollars and the consumption of meat in their families may be in kilograms. The standard deviation of wages in dollars cannot be compared with the standard deviation of amount of meat in kilograms. Both the standard deviations need to be converted into coefficient of variation for comparison. Suppose the value of C.V for wages is 10% and the values of C.V for kilograms of meat is 25%. This means that the wages of workers are consistent. Their wages are close to the overall average of their wages. But the families consume meat in quite different quantities. Some families usevery small quantities of meat and some others use large quantities of meat. We say that there is greater variation in their consumption of meat. The observations about the quantity of meatare more dispersed or more variant.

Example

Calculate the coefficient of standard deviation and coefficient of variation for the following sample data: 2, 4, 8, 6, 10 and 12.

Solution

X	**$(X - \bar{X})^2$**
2	$(2 - 7)^2 = 25$
4	$(4 - 7)^2 = 9$
8	$(8 - 7)^2 = 1$
6	$(6 - 7)^2 = 1$
10	$(10 - 7)^2 = 9$
12	$(12 - 7)^2 = 25$
$\Sigma X = 42$	**$\Sigma(X - \bar{X})^2 = 70$**

$\bar{X} = \frac{\Sigma 42}{n} = \frac{}{6} = 7$

$S = \sqrt{\Sigma(X - \bar{X})^2/n}$

$S = \sqrt{\frac{70}{6}} = \sqrt{35} = 3.42$

Coefficient of Standard Deviation = $\frac{S}{\bar{X}} = \frac{3.42}{7} = 0.49$

Coefficient of Variation (C.V) = $\frac{S}{\bar{X}} \times 100 = \frac{3.42}{7} \times 100 = 48.86\%$

3.8.1 USES OF COEFFICIENT OF VARIATION

- Coefficient of variation is used to know the consistency of the data. By consistency we mean the uniformity in the values of the data/distribution from arithmetic mean of the data/distribution. A distribution with smaller C.V than the other is taken as more consistent than the other.
- C.V is also very useful when comparing two or more sets of data that are measured in different units of measurement.

3.9 THE VARIANCE

Variance is another absolute measure of dispersion. It is defined as the average of the squared difference between each of the observation in a set of data and the mean. For a sample data the variance is denoted by S^2 and the population variance is denoted by σ^2 (sigma square).

The sample variance S^2 has the formula

$$S^2 = \frac{\Sigma (X - \bar{X})^2}{n}$$

where $\bar{X}$ is sample mean and n is the number of observations in the sample.

The population variance σ^2 is defined as

$$\sigma^2 = \frac{\Sigma (X - \mu)^2}{N}$$

where μ is the mean of the population and N is the number of observations in the data. It may be remembered that the population variance σ^2 is usually not calculated. The sample variance S^2 is calculated and if need be, this S^2 is used to make inference about the populationvariance.

The term $\Sigma (X - \bar{X})^2$ is positive, therefore S^2 is always positive. If the original observations are in centimetre, the value of the variance will be (Centimetre) 2. Thus the unit of S^2 is the square of the units of the original measurement.

For a frequency distribution the sample variance S^2 is defined as S^2

$$= \frac{\Sigma f(X - \bar{X})^2}{\Sigma f}$$

For a frequency distribution the population variance σ^2 is defined as

$$\sigma^2 = \frac{\Sigma (X - \mu)^2}{\Sigma f}$$

In simple words we can say that variance is the square root of standard deviation.Variance = (Standard Deviation)^2

Example

Calculate variance from the following distribution of marks:

Marks	**No. of Students**
1 - 3	40
3 - 5	30
5 - 7	20
7 - 9	10

Solution

Marks	**F**	**X**	**fX**	**$(X - \bar{X})^2$**	**$f (X - \bar{X})^2$**
1 -3	40	2	80	4	160
3 - 5	30	4	120	0	0
5- 7	20	6	120	4	80
7 - 9	10	8	80	16	160
Total	**100**		**400**		**400**

$\bar{X} = \frac{\Sigma}{\Sigma f} = \frac{400}{100} = 4$

$S^2 = \frac{\Sigma f (X - \bar{X})^2}{\Sigma f} = \frac{400}{100} = 4$

Variance $S^2 = 4$.

3.10 SKEWNESS AND KURTOSIS

1) SKEWNESS

Skewness is the absence of symmetry in a distribution. Though averages and measures of dispersion are useful in studying the data, the shape of the frequency curve may also be equally important to the statistician. If we are studying a certain phenomenon over a period oftime, the average may remain the same, but the structure of the distribution may change. Twodistributions may have identical averages, yet one my tail off towards the higher values and the other towards the lower values.

To study the distribution we need a measure of this tendency which will give us the direction and degree of this tendency which is called skewness.

A fundamental task in many statistical analyses is to characterize the *location* and *variability* of a data set. A further characterization of the dataincludes skewness and kurtosis.

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

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

The histogram is an effective graphical technique for showing both the skewness and kurtosis of data set.

For univariate data Y_1, Y_2, ..., Y_N, the formula for skewness is:
g1=∑Ni=1(Yi−Y¯)3/Ns3

where Y¯ is the mean, ***s*** is the standard deviation, and *N* is the number of data points. Note that in computing the skewness, the *s* is computed with *N* in the denominator rather than *N* - 1.

The above formula for skewness is referred to as the Fisher-Pearson coefficient of skewness. Many software programs actually compute the adjusted Fisher-Pearson coefficient of skewness

G1=N(N−1)−−−−−−−−−√N−2∑Ni=1(Yi−Y¯)3/Ns3

This is an adjustment for sample size. The adjustment approaches 1 as *N* gets large. For reference, the adjustment factor is 1.49 for *N* = 5, 1.19 for *N* = 10, 1.08 for *N* = 20, 1.05 for *N* = 30, and 1.02 for *N* = 100.

The skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right. By skewed left, we mean that the left tail is long relative to the right tail. Similarly, skewed right means that the right tail is long relative to the left tail. If the data are multi-modal, then this may affect the sign of the skewness.

Some measurements have a lower bound and are skewed right. For example, in reliability studies, failure times cannot be negative.

It should be noted that there are alternative definitions of skewness in the literature. For example, the Galton skewness (also known as Bowley's skewness) is defined as

Galton skewness=Q1+Q3−2Q2Q3−Q1

where Q_1 is the lower quartile, Q_3 is the upper quartile, and Q_2 is the median.The

Pearson 2 skewness coefficient is defined as

Sk2=3(Y¯−Y~)s

where Y~ is the sample median.

There are many other definitions for skewness that will not be discussed here.

2) KURTOSIS

For univariate data Y1, Y2, ..., YN, the formula for kurtosis is:

kurtosis=∑Ni=1(Yi−Y¯)4/Ns4

where $\bar{Y}$ is the mean, s is the standard deviation, and N is the number of data points. Note that in computing the kurtosis, the standard deviation iscomputed using N in the denominator rather than N - 1.

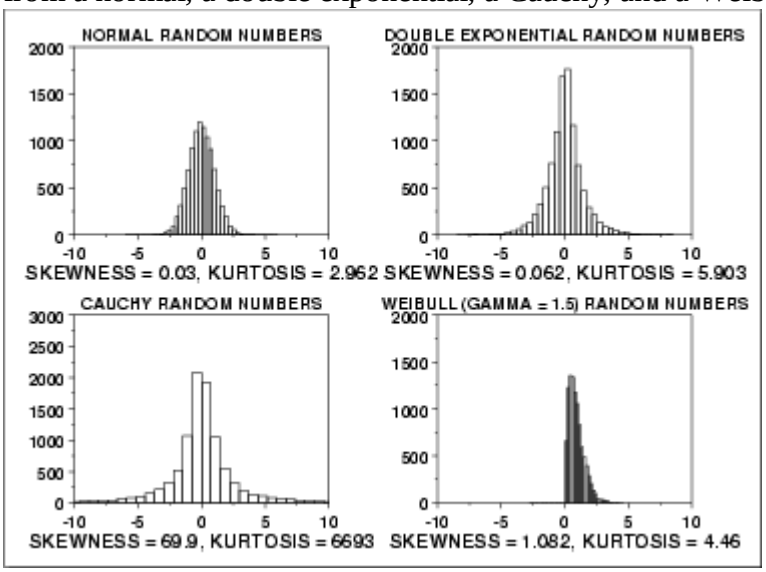
The kurtosis for a standard normal distribution is three. For this reason, some sources use the following definition of kurtosis (often referred to as "excess kurtosis"):

$$\text{kurtosis}=\sum_{i=1}^{N}(Y_i-\bar{Y})^4/Ns^4-3$$

This definition is used so that the standard normal distribution has a kurtosis of zero. In addition, with the second definition positive kurtosis indicates a "heavy-tailed" distribution and negative kurtosis indicates a "light tailed" distribution.

Which definition of kurtosis is used is a matter of convention (this handbook uses the original definition). When using software to compute the sample kurtosis, you need to be aware of which convention is being followed. Many sources use the term kurtosis when they are actually computing "excess kurtosis", so it may not always be clear.

The following example shows histograms for 10,000 random numbers generated from a normal, a double exponential, a Cauchy, and a Weibull distribution.



The first histogram is a sample from a normal distribution. The normal distribution is a symmetric distribution with well-behaved tails. This is indicated by the skewness of 0.03. The kurtosis of 2.96 is near the expected value of 3. The histogram verifies the symmetry.

The second histogram is a sample from a double exponential distribution. The double exponential is a symmetric distribution. Compared to the normal, ithas a stronger peak, more rapid decay, and heavier tails. That is, we would

expect a skewness near zero and a kurtosis higher than 3. The skewness is 0.06 and the kurtosis is 5.9.

The fourth histogram is a sample from a Weibull distribution with shape parameter 1.5. The Weibull distribution is a skewed distribution with the amount of skewness depending on the value of the shape parameter. The degree of decay as we move away from the center also depends on the value of the shape parameter. For this data set, the skewness is 1.08 and the kurtosisis 4.46, which indicates moderate skewness and kurtosis.

Many classical statistical tests and intervals depend on normality assumptions. Significant skewness and kurtosis clearly indicate that data are not normal. If a data set exhibits significant skewness or kurtosis (as indicatedby a histogram or the numerical measures), what can we do about it?

One approach is to apply some type of transformation to try to make the data normal, or more nearly normal. The Box-Cox transformation is a useful technique for trying to normalize a data set. In particular, taking the log or square root of a data set is often useful for data that exhibit moderate right skewness.

Another approach is to use techniques based on distributions other than the normal. For example, in reliability studies, the exponential, Weibull, and lognormal distributions are typically used as a basis for modeling rather than using the normal distribution. The probability plot correlation coefficient plot and the probability plot are useful tools for determining a good distributional model for the data.

3.11 EXERCISES

Q1. Calculate the range and quartile deviation for wages: Also calculate coefficient ofquartile deviation:

Wages	30 - 32	32 - 34	34 - 36	36 - 38	38 - 40	40 - 42	42 - 44
Labourers	12	18	16	14	12	8	6

Hint: Coefficient of Q.D. = $\frac{Q3-Q1}{Q3+Q1} = 0.081$

Q2. Calculate the standard deviation from the following:

Marks	10	20	30	40	50	60
No. of Students	8	12	20	10	7	3

Hint: $\sigma = \sqrt{\sqrt{\frac{\Sigma fd12}{N} - (\frac{\Sigma fd1}{N})^2}} X\ C$

$= \sqrt{\sqrt{\frac{109}{60} - (\frac{5}{60})^2}} X\ 10 = 13.5$

Q3. Find the mean and standard deviation of the following observations:X:

1 2 4 6 8 9

Transform the above observation such that the mean of transformed observations becomes double the mean of X, standard deviation remain unchanged.

Hint: Mean = $\frac{\Sigma X}{N} = 30/6 = 5$ Let d = X - 5. Then

$\frac{\Sigma d^2}{N} = 52$. $\sigma = \sqrt{\frac{\Sigma d2}{N} - (\frac{\Sigma d}{N})^2} = \sqrt{\frac{52}{6}} = 2.94$.

Q4. Explain positive and negative skewness with the help of sketches.Q5.

Write short notes on skewness and kurtosis.

4

PROBABILITY AND PROBABILITY DISTRIBUTION

PERMUTATIONS AND COMBINATIONS

UNIT STRUCTURE

4.1.1 (i) COUNTING TECHNIQUES

Fundamental Principle of Counting

If the first operation can be performed in any one of the m ways and then a second operation can be performed in any one of the n ways, then both can be performed together in any one of the m × n ways. This rule can be generalised. If first operation can be performed in any one of the n1 ways, second operation in any one of the n2 ways, ... kth operation in any one of the nk ways, then together these can be performed in any one of the n1 × n2 × ..... × nk ways.

Factorial Notation:

The product of the first natural number is called factorial n, denoted by n!.

Thus **n! =1x2x3……..x(n-1) x n Note: 0!=1, n!=n (n-1)!**
e.g. 4! =1x2x3x4=24; n! = n(n-1)(n-2)……..(n-r+1)(n-r)!;r ≤ n

(A) Permutations:

Each of the different **arrangements,** which can be made out of a given number of things by taking some or all of them at a time is called a **permutation.**
Note: In Permutations, the order in which the things are arranged is important.

(a) Permutations of n objects: The total number of permutations of n distinct objects is n! Using symbols, we can write $^{n}P_{n}$= n!, (where n denotes the permutations of n objects, all taken together).

Let us assume there are n persons to be seated on n chairs. The first chair can be occupied by any one of the n persons and hence, there are n ways in which it can be occupied. Similarly, the second chair can be occupied in n - 1 ways and so on. Using the fundamental principle of counting, the total number of ways in which n chairs can be occupied **by n** persons or the permutations of n objects taking all at a time is given by:

$^{n}P_{n}$= n(n – 1)(n – 2) ..... 3.2.1 = n!

(b) Permutations of n objects taking r at a time:

e.g. If there are three things a, b, c then the number of permutations i.e. arrangements of these three things taken two at a time is denoted by $^3P_2$ and is given by $^3P_2=3(3-1)=3.2=6$ as follows (a, b); (a ,c); (b ,c); (b, a); (c, a); (c, b)

In terms of the example, considered above, now we have n persons to be seated on r chairs, where r ≤ n.

The number of permutations of n persons to be seated on r chairs, where r ≤ n.

, ^nP_r== n(n – 1)(n – 2) ..... [n – (r – 1)] = n(n – 1)(n – 2) ..... (n – r + 1).

On multiplication and division of the R.H.S. by (n - r)!, we get

$$^nP_r = \frac{n(n-1)(n-2)(n-3)\ldots\ldots\ldots(n-r+1)(n-r)!}{(n-r)!} = \frac{n!}{(n-r)!}, \; r \le n$$

e.g. If there are three things a, b, c then the number of permutations i.e. arrangements of these three things taken two at a time is denoted by $^3P_2$ and is given by $^3P_2=3(3-1)=3.2=6$ as follows (a, b); (a ,c); (b ,c); (b, a); (c, a); (c, b)

(c) <u>Permutations of n objects taking r at a time when any object may be repeated any number of times:</u>

Here, each of the r places can be filled in n ways. Therefore, total number of permutations is n^r.

- **Examples:**

[1] What are the total number of ways of simultaneous throwing of (i) 3 coins, (ii) 2 dice and (iii) 2 coins and a die?

<u>Solution:</u>

(i) Each coin can be thrown in any one of the two ways, i.e., a head or a tail,

Therefore, the number of ways of simultaneous throwing of 3 coins = $2^3 = 8$.

(ii) Each die can be thrown in any one of the 6 ways,

Therefore the total number of ways of simultaneous throwing of two dice =$6^2 = 36$.

(iii) The total number of ways of simultaneous throwing of 2 coins and a die
= $2^2 \times 6 = 4x6=24$.

[2] In how many ways 5 passengers can sit in a compartment having 15 vacant seats?

<u>Solution:</u>

No. of ways of 5 passengers can sit in a compartment having 15 vacant seats=$^{15}P_5$

$$= \frac{15!}{(15-5)!} = \frac{15!}{10!} = \frac{15x14x13x12x11x10!}{10!} = 15x14x13x12x11$$

[3] In how many ways can the letters of the word MATHEMATICS be arranged?

Solution:

The word **MATHEMATICS** has 11 letters in which there are 2M 's, 2T's, 2A's,1H,1E,1I,1Cand 1S. Thus, the required number of permutations $\frac{11!}{2!\,2!\,2!\,1!\,1!\,1!\,1!\,1!}$ =4989600

[4] (a) In how many ways 4 men and 3 women can be seated in a row such that women occupy the even places?

(b) In how many ways 4 men and 4 women can be seated such that men and women occupy alternative places?

Solution:

(a) 4 men can be seated in 4! ways and 3 women can be seated in 3! ways. Since each arrangement of men is associated with each arrangement of women, therefore, the required number of permutations = 4! 3! = 144.

(b) There are two ways in which 4 men and 3 women can be seated

The required number of permutations = 2 .4! 3! = 288

[5] How many five digits numbers can be made by using the digits 1, 2, 3,4,5,6 and 7? When repetition is not allowed? How many of these will be greater than 50,000.

Solution:

There are 7 given digits 1, 2, 3,4,5,6 and 7.

No. of ways to form five digit number out of given digits=$^{7}P_{5}$

A five digit number out of given digit is greater than 50,000 if it begin with 5, 6 and 7.

Therefore digit at ten thousand's place can be chosen in 3 ways. The other four digits can be chosen out of the remaining 6 digits in $^{6}P_{4}$ ways.

Therefore, required no. of five digit numbers =3x$^{6}P_{4}$=3x$\frac{6!}{(6-4)!}$ =$3x\frac{6x5x4x3x2!}{2!}$ =3x6x5x4x3=1080

[**6] How many four digits numbers can be made by using the digits 0, 1, 2 and 4? When (i) repetition is allowed and (ii) repetition is not allowed.**

Solution:

(i)Repetition is allowed:

To make four digit number we cannot take 0 at thousand' place.

Therefore digit at thousand's place can be chosen in 3 ways. The digit at hundred's place can be chosen in 4 ways. The digit at ten's place can be chosen in 4 ways. The digit at unit's place can be chosen in 4 ways.

Therefore, four digit number can be made by 3x4x4x4=192 ways.

(ii) Repetition is not allowed.

To make four digit number we cannot take 0 at thousand' place.

Therefore digit at thousand's place can be chosen in 3 ways. The digit at hundred's place can be chosen in 3 ways. The digit at ten's place can be chosen in 2 ways. The digit at unit's place can be chosen in 1way.

Therefore, four-digit number can be made by 3x3x2x1=18 ways.

[7] In how many ways can 5 men, 4 women and 3 children be arranged for photographs so that (i) all men are together and so are all women and children. (ii) **all men are together and so are all women. (iii) only all men are together and so are all women and children.**

Solution:

(i) **all men are together and so are all women and children.**

Now, 5 men forms one group, 4 women forms one group and 3 children also forms one group. These three groups with all men, women and children are together can be arranged in 3! Ways. Further, 5 men among themselves in 5! ways, 4 women among themselves in 4! ways, 3 children among themselves in 3! ways.

Therefore, 5 men, 4 women and 3 children be arranged for photographs so that all men are together and so are all women and children in 3!5!4!3! ways.

(ii) **all men are together and so are all women.**

Now, 5 men forms one group, 4 women forms one group and 3 children can be arranged in 5! Ways. . Further, 5 men among themselves in 5! Ways, 4 women among themselves in 4! ways.

Therefore, 5 men, 4 women and 3 children be arranged for photographs so that all men are together and so are all women in 5!5!4! ways.

(iii) only all men are together.

Now, 5 men forms one group, 4 women and 3 children can be arranged in 8! ways.

Further, 5 men among themselves in 5! Ways.

Therefore, 5 men, 4 women and 3 children be arranged for photographs so that only all men are together in 8!5! ways.

[8] How many different 7-place license plates are possible if the first 3 places are occupied by letters and the final 4 by numbers under the assumption that no letter or number can be repeated in a single license plate?

Solution:

7-places license plates with the first 3 places are occupied by letters selecting from the A to Z (26 letters) and the final 4 by numbers selecting from the numbers 0, 1, 2……9 (10 numbers)

If no letter or number can be repeated in a single license plate, the first three places are occupied out of 26 letters in ${}^{26}p_3$ ways and the final 4 numbers are occupied from 10 numbers in ${}^{10}p_4$ ways.

Therefore, different 7-place license plates are possible if the first 3 places are occupied by letters and the final 4 by numbers in $^{26}p_3 x^{10}p_4 = 26x25x24x10x9x8x7 = 78624000$ ways.

[9] When Dr. Shah arrives in his dispensary, he finds 12 patients waiting to see him. If he can see only one patient at a time, find the number of ways, he can schedule his patients (a) if they all want their turn, and (b) if 3 leave in disgust before Dr. Shah get around to seeing them.

Solution:

(a) if they all want their turn, there are 12 patients and all wait to see the doctor.

Therefore, number of ways Dr. Shahcan schedule his patients

$= {}^{12}P_{12} = 12! = 479{,}001{,}600$

(b) if 3 leave in disgust before Dr. Shah get around to seeing them.

There are 12-3=9 patients. They can be seen in $^{12}P_9$ ways = 79,833,600 ways.

(B) Combination:

When no attention is given to the order of arrangement of the selected objects, we get a combination. We know that the number of permutations of n objects taking r at a time is, $^{n}P_r$. Since r objects can be arranged in r! ways, therefore, there are r! Permutations corresponding to one combination. Thus, the number of combinations of n objects taking r at a time, denoted by, $^{n}C_r$, can be obtained by dividing, $^{n}P_r$ by r!,

i.e. $^{n}C_r = \frac{n!}{(n-r)!\ r!} = {}^{n}P_r\ r!\ ,\ {}^{n}P_r$

Note:

(a) Since, $^{n}C_r = {}^{n}C_{n-r}$, therefore, $^{n}C_r$ is also equal to the combinations of n objects taking (n - r) at a time.

(b) (b) The total number of combinations of n distinct objects taking 1, 2, ... n respectively, at a time is $^{n}C_1 + {}^{n}C_2 + {}^{n}C_3 + {}^{n}C_4 \ldots\ldots + {}^{n}C_n = 2^{n-1}$.

- **Examples:**

[1] **In how many ways two balls can be selected from 8 balls?**

Solution:

2 balls can be selected from 8 balls in $^{8}C_2$ ways.

$^{8}C_2 = \frac{8!}{2!\ 6!} = \frac{8x7x6!}{2x1x6!} = \frac{4x7}{1} = 28$

[2] **In how many ways can 5 students be selected for a student's committee out of 7 students?**

Solution:

No. of ways 5 students can be selected for a student's committee out of 7 students= $^{7}C_5$

$$= \frac{7!}{5!\,(7-5)!} = \frac{7!}{5!\,2!} = \frac{\mathbf{7x6x5!}}{\mathbf{5!\,x2x1}} = \frac{\mathbf{7x6}}{\mathbf{2x1}} = 21$$

[3] **In how many ways can 2 boys and 2 girls be selected from a group of 6 boys and 5 girls?**

Solution:

No. of ways 2 boys can be selected from a group of 6 boys= ${}^{6}C_{2}$ and No. of ways 2 girls can be selected from a group of 5 girls= ${}^{5}C_{2}$

Therefore,

No. of ways 2 boys and 2 girls can be selected from a group of 6 boys and 5 girls

$$= {}^{6}C_{2} \times {}^{5}C_{2}$$

$$= \frac{6!}{2!\,(6-2)!} \times \frac{5!}{2!\,(5-2)!} = \frac{6!}{2!\,4!} \times \frac{5!}{2!\,3!} = \frac{\mathbf{6x5x4!}}{\mathbf{2x1x4!}} \times \frac{\mathbf{5x4x3!}}{\mathbf{2x1x3!}} = 15x10 = 150$$

[4] The staff of a department consists of a manager, an officer and 10 clerks. A committee of 4 is to be selected from the department. Find the number of ways in which this can be done so as to always include (1) the manager, (2) the manager but not the officer, (3) neither the manager nor the officer.

Solution:

(1) The number of ways in which this can be done so as to always include the manager:

The manager can be selected in only one way and the remaining three members of the committee can be selected from amongst the 11 members of the department in ${}^{11}C_{3}$ ways.

Therefore, number of ways in which the committee can be selected so as to include the manager=1 x ${}^{11}C_{3} = \frac{\mathbf{11x10x9}}{\mathbf{3x2x1}} = 165$

(2) The number of ways in which this can be done so as to always include the manager but not the officer:

The manager can be selected in only one way. Since officer is not to be included in the committee, the remaining three members can be selected amongst 10 clerks in ${}^{10}C_{3}$ ways.

Therefore, number of ways in which committee can be selected so as to include the manager but not the officer= 1x ${}^{10}C_{3} = \frac{\mathbf{10x9x8}}{\mathbf{3x2x1}} = 120$

(3) The number of ways in which this can be done so as to always include neither the manager nor the officer:

As all the four members are to be selected from 10 clerks and neither the manager nor the officer is to be included in the committee,

Number of possible ways = ${}^{10}C_{4} = \frac{\mathbf{10x9x8x7}}{\mathbf{4x3x2x1}} = 210$

[5] In a party, every person shakes hands with every other person present. If the total number of handshakes was 105, find the total number of persons present at the party.

Solution:

Assume there are n persons in the party, then the number of handshakes is ${}^{n}C_2$.

As ${}^{n}C_2 = 105$ i.e. $\frac{\mathbf{n(n-1)}}{\mathbf{2}} = 105$

Therefore, n (n-1) = 210

Thus, factors of 210 such that they are successive integers are 15 x 14 = 210

So, the number of persons present at the party are 15.

[6] A person has 12 friends of whom 8 are relatives. In how many way can he invite7 guests such that 5 of them are relatives?

Solution:

Of the 12 friends, 8 are relatives and the remaining 4 are not relatives. He has to invite 5 relatives and 2 friends as his guests. 5 relative can be chosen out of 8 in ${}^{8}C_5$ ways; 2 friends can be chosen out of 4 in ${}^{4}C_2$ ways.

Hence, by the fundamental principle, the number of ways in which he can invite 7 guests such that 5 of them are relatives and two of them are friends = ${}^{8}C_5 x {}^{4}C_2$

$$= \frac{8!}{5!\,(8-5)!} x \frac{4!}{2!\,(4-2)!}$$

$$= \frac{(\mathbf{8x7x6x5!})}{\mathbf{5!x3!}} x \frac{\mathbf{4x3x2x1}}{\mathbf{2!2!}}$$

$$= 8x7x6$$

$$=336$$

[7] A box contains 7 red, 6 white and 4 blue balls. How many selections of three balls can be made so that (a) all three are red, (b) none is red, (c) one is each colour?

Solution:

(a) all three are red:

All three balls will be red colour if taken out from 7 red balls and this can be done in ${}^{7}C_3$ ways i.e. $\frac{7!}{3!\,(7-3)!} = \frac{(\mathbf{7x6x5x4!})}{\mathbf{4!x3!}} = 35$

(b) **none is red** :

None of three balls is red if these are chosen from 6 white and 4 blue balls and this can be done in ${}^{10}C_3$ ways i.e. $\frac{10!}{3!\,(10-3)!} = \frac{(\mathbf{10x9x8x7!})}{\mathbf{7!x3!}} = 120$

(c) one is each colour:

The number of groups of three balls such that one is of each colour

$= {}^{7}C_1 x {}^{6}C_1 x {}^{4}C_1 = 168$

4.1.2 (ii) EXERCISES

(A) Multiple Choices

Choose the correct alternative from the following:

a) Permutation is used when……

(A) Order is not important (B) order is important
(C) either (A) OR (B) (D) neither (A) OR (B)

b) Combination is used when……

(A) order is not important (B) order is important
(C) either (A) OR (B) (D) neither (A) OR (B)

c) If ${}^{n}p_3 = 6\ {}^{n}c_4$, then value of n is
(A) 10 (B) 8 (C) 7 (D) 5

d) Number of way four different books can be arranged

(A) 4 (B) 24 (C) 1 (D) 0

e) In how many different ways can 6 people be photographed, if only 2 can be seated at a time?

(A) 6 (B) 2 (C) 15 (D) 30

f) The value of r if $5^4P_r = 6^5P_{r-1:}$

(A) 5 (B) 3 (C) 2 (D) 4

g) In how many four digit numbers can be formed from the digits 0, 1, 2… to 9, if no digit is repeated?

(A) 2688 (B) 4536 (C) 5040 (D) 3024

h) How many words can be formed of the letter ARTICLE, so that the vowels occupy only the even position?

(A) 576 (B) 840 (C) 600 (D) None

j) There are 8 questions in a question paper. In how many different ways can a student attempt 7 questions from the paper?

(A) 1 (B) 8 (C) 7 (D) 15

k) There are 10 people in a room. Each person shakes hand with every other person in the room. The total number of hand shake is:

(A) 10 (B) 100 (C) 20 (D) 45

l) A Supreme court bench consist of 5 judges. The number of ways in which the bench can give majority decision is
(A) 10 (B) 18 (C) 17 (D) 16

m) Relationship between ${}^{n}p_{r}$ and ${}^{n}c_{r}$ is
(A) ${}^{n}p_{r} = {}^{n}c_{r+r}!$ (B) ${}^{n}p_{r} = {}^{n}c_{r} \times r!$ (C) ${}^{n}p_{r} = {}^{n}c_{r-r}!$ (D) ${}^{n}p_{r} = {}^{n}c_{r}/r!$

(B) <u>Problems</u>

[1] In how many ways can the letters of the word COMMITTEE be arranged?

[2] A group of 6 students comprised of 3 boys and 3 girls. In how many ways could they be arranged in a straight line such that (1) the girls and boys occupy alternate position? (2) All three boys were sitting together?

[3] There are 3 different books of economics, 4 different books of commerce and 5 different books of statistics. In how many ways these can be arranged on a shelf when

(a) all the books are arranged at random**,**
(b) books of each subject are arranged together,
(c) books of only statistics are arranged together, and
(d) books of statistics and books of other subjects are arranged together?

[4] How many 4 lettered different words can be formed by using the letters a, b, c and d? When (i) repetition is allowed and (ii) repetition is not allowed

[5] In how many different ways can 6 people be photographed, if only 4 can be seated at a time?

[6] A cricket team of 11 members is to be selected from 18 players of whom 10 are batsmen, 6 are bowlers and 2 are wicket-keepers. In how many ways can the team be formed if at least one wicket keeper and at least 5 bowlers are to be included?

[7] In how many ways can a pack of 52 cards be divided equally among 4 players in order?

[8] There are 8 professors and 12 students out of whom a committee of 2 professors and 3 students is to be formed. Find the number of ways in which the committee can be formed such that (i) a particular professor is included. (ii) a particular student is to be excluded.

❖❖❖❖

5

PROBABILITY

UNIT STRUCTURE

5.(i) Introduction to Probability
5.(ii) Concept of Probability
5.(iii) Basic rules of Probability
5.(iv) Solved Examples
5.(v) Exercises

5. (i) INTRODUCTION TO PROBABILITY

UNCERTAINTY AND MYSTERY ARE ENERGIES OF LIFE.DON'T LET THEM SCARE YOU UNDULY, FOR THEY KEEP BOREDOM AT BAY AND SPARK CREATIVITY.

R.I.FITZHENRY

Uncertainty is an important part of everyday life. Some events are certain like the Sun rises every day, but many events like winning or losing the game, stock market etc. involved uncertainty or chance. When there is no guarantee of an event it belongs to the domain of uncertainty. Though such uncertainty is involved in most of the situations, such happenings have some "chance" or "probability" of their occurrence. To find a measure for probability, it is necessary to perform certain experiments, visualize the possible outcomes of the experiment. The following terms are used to describe the result of an experiment:

Statistical Experiment: A statistical experiment is a random i.e. nondeterministic experiment. An experiment is a process of measurement or observation which has number of possible results (outcomes). The experiments like tossing of an unbiased coin, throwing a fair dice, Selecting a committee of three persons from a group of ten persons, , Measuring copper content of brass, Number of customers enters a mall are examples statistical experiment. Measuring no. of hands of undergraduate students, measuring density of pure gold are examples non-statistical experiment.

Sample Space of an Experiment: A set of all possible outcomes of an experiment is called sample space of the experiment. It is denoted by S. e.g. sample space for the random experiment of tossing of an unbiased coin is S= {H, T}, where H denotes "head" and T denotes "tail". similarly sample space for the random experiment of throwing a fair dice is {1,2,3,4,5,6}

Sample points : The possible outcomes in a sample space of an experiment are called Sample points. The number of sample points in sample space is denoted by n(S).e.g. in case of tossing of an unbiased coin, n(S)=2 and in case of throwing a fair dice, n(S)=6

An Event: Any subset (say A) of sample space S of an experiment is called an Event A and the number of sample points in event A is denoted by n(A).e.g. If a fair dice is rolled and event A is that even number occurs on the uppermost face of the dice, then A={2,4,6}and n(A) =3.

Impossible Event: An event corresponding to null set is called an impossible event. e.g. If a fair dice is rolled and event A is that number 7 occurs on the uppermost face of the dice , then event A is impossible event.

Certain Event: An event that is certain to occur is called certain event. Certain event contains all the elements of the sample space.

Complementary Event: If A is any event from sample space of an experiment S, then non-occurrence of event A is an event called complementary event of event A and is denoted by A^c or A'.e.g. If S={1,2,3,4,5,6} and A={2,4,6} then A^c ={1,3,5}

Algebra of Events:

(i)A U B: Either event A or event B or both A and B occur.

(ii)A $\cap B$: Occurrence of both event A and B simultaneously.

Equally Likely Events: The events of a sample space which have same chance of occurring are called Equally Likely Events. Eg. Consider an experiment of throwing a die. Any one of the numbers cannot be expected to turn in preference to any other of them. These sample points are equally likely. The event corresponding to these sample points. i.e.{1},{2},{3},{4},{5},{6} are equally likely events.

Mutually Exclusive Events:

If two events A and B cannot occur together, then A and B are said to be mutually exclusive events. In such a case A $\cap B$ =φ. e.g. AnA^c = φ, hence A and A^c are mutually exclusive Events. Thus complementary events are mutually exclusive Events. Eg. In one throw of a fair die, A be the event obtaining even numbers and B be the event obtaining odd numbers on the uppermost face. Then S={1,2,3,4,5,6}; A={2,4,6} ; B={1,3,5}.

Here A $\cap B$ =φ. Hence events A and B are called mutually exclusive events.

Exhaustive events: If two events A and B are such that AUB =S, then A and B are called Exhaustive Events. e.g. AUA^c =S, hence A and A^c are exhaustive Events. Thus complementary events are exhaustive Events. . In one throw of a fair die, A be the event obtaining even numbers and B be the event obtaining odd numbers on the uppermost face. Then S={1,2,3,4,5,6}; A={2,4,6} ; B={1,3,5}.

Here AUB =S. Hence events A and B are called exhaustive events.

5. (ii) CONCEPT OF PRBABILITY

Mathematical or Classical definition of Probability of an event:

If sample space S of an experiment has n equally likely ,Mutually exclusive and Exhaustive outcomes of which mare favourable to the occurrence of event A ,then probability of event A ,denoted by P(A) is given by

$$\mathbf{P(A)} = \frac{Number\ of\ sample\ points\ in\ A}{Number\ of\ sample\ points\ in\ S} = \frac{n(A)}{n(S)} = \frac{m}{n}$$

It is obvious from the definition that, $0 \le P(A) \le 1$

Probability Assigning Techniques:

The assignment of probabilities to various elementary events of a sample space can be done in any one of the following three ways:

[1] **Subjective probability Assignment:** This is the technique of assigning probabilities on the basis of personal judgement. Under the subjective assignment, the probabilities to various elementary events are assigned on the basis of the expectations or the degree of belief of the statistician. Such assignment may differ from individual to individual and much depends upon the expertise of the statistician an assigning the probability. These probabilities, also known as personal probabilities, are very useful in the analysis of various business and economic problems.

[2] **Classical/Mathematical Probability Assignment:**

This is the technique under which the probability is assigned by calculating the ratio of the number of ways in which a given outcome can occur to the total number of possible outcomes. We know that various elementary events of a random experiment, under the classical definition, are equally likely and, therefore, can be assigned equal probabilities.

[3] **Empirical/Statistical probability assignment:** This is an objective method of assigning probabilities and is popular with the decision maker. Under this technique the probability is assigned by calculating the relative frequency of occurrence of of a given event over an infinite no. of occurrences. The probability assignment through this technique may well be unrealistic if future conditions do not happen to be a reflection of the past.

5. (iii) SOME BASIC RULES OF PROBABILITY

(i)Addition Theorem (General): If A and B are any two events associated with an experiment, then the probability of occurrence of events A or B or both A and B is given by

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If A, B and C are three events associated with an experiment, then the probability of occurrence of atleast one of the three events A, B and C is given by

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$$

(ii) Addition Theorem (Mutually exclusive event):

If A and B are mutually exclusive events then $A \cap B = \varphi$.

In this case Addition theorem becomes $P(A \cup B) = P(A) + P(B)$

If A, B and C are mutually exclusive events, then

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

(iii) If A and A^c are complementary events, then $P(A^c) = 1 - P(A)$

Conditional Probability: The probability of event A given that Event B has already occurred is called conditional probability of event A given that the event B has already occurred and is denoted by P(A/B).Similarly , conditional probability of event B given that A has already occurred is denoted by P(B/A).

Independent Event: If probability of event A is not affected by occurrence or non-occurrence of event B, then the events A and B are independent events. In such a case,

$$P(A/B) = P(A) \text{ and } P(B/A) = P(B).$$

Multiplication Theorem: If A and B are any two events associated with an experiment, then the probability of simultaneous occurrence of events A and B is given by

$$P(A \cap B) = P(A) \cdot P(B/A) = P(B) \cdot P(A/B)$$

Multiplication theorem (Independent Events): If events A and B are independent events then, P (A/B) = P (A) and P (B/A) =P (B). Hence Multiplication theorem is

P (AIB) =P (A) .P (B)

If A, B and C are three events associated with an experiment, then the probability of occurrence of atleast one of the three events A, B and C is given by

P (AIBIC) = P (A) .P (B).P(C)

Addition theorem (Independent Events):

P (AUB) = P(A) + P(B) –P(A).P(B)

Baye's Theorem On Inverse Probability:

Let A_1, A_2 , A_3, ……..A_n be n mutually exclusive and exhaustive events defined on a sample space S. Let H be an event defined on S, such that P(H)≠0.

Then

$$P(A_i/H) = \frac{P\left(\frac{H}{Ai}\right)P(Ai)}{P\left(\frac{H}{A1}\right)P(A1) + P\left(\frac{H}{A2}\right)P(A2) + \ldots\ldots\ldots P\left(\frac{H}{An}\right)P(An)}$$

Bayes' Theorem is based on the concept of conditional probability. The revised probabilities, thus obtained, are known as posterior or inverse probabilities. Using this theorem it is possible to revise various business decisions in the light of additional information.

5. (iv) SOLVED EXAMPLES:

[1] If P(A U B)= $\frac{5}{6}$, P ($\bar{A}$) = $\frac{1}{3}$, P(B) = $\frac{1}{2}$, Find P ($\bar{B}$), P(A I B) and P(A/B)

Solution:

P($\bar{B}$) =1- P(B) = $\frac{1}{2}$ P(A) =1- P($\bar{A}$) = 1- $\frac{1}{3}$ = $\frac{2}{3}$

P (A U B) = P(A) + P(B) – P(A I B)

Therefore, P (A I B) =P(A)+ P(B)- P (A U B)

$$= \frac{2}{3} + \frac{1}{2} - \frac{5}{6}$$

$$= \frac{1}{3}$$

[2] Given : P(A) =$\frac{1}{3}$, P(B) =$\frac{1}{4}$, P(C) =$\frac{1}{5}$,P(A I B) =$\frac{1}{12}$,P(A I C) =$\frac{1}{15}$, P(B I C) =$\frac{1}{20}$, P(A I B I C) =$\frac{1}{60}$. Find P (A U B U C)

Solution:

P(AUBUC)=P(A) + P(B)+P(C) –P(AIB)-P(BIC)-P(AIC)+P(AIBIC)

$$= \frac{1}{3} + \frac{1}{4} + \frac{1}{5} - \frac{1}{12} - \frac{1}{20} - \frac{1}{15} + \frac{1}{60}$$

$$= \frac{3}{5}$$

[3] Determine the probability of the following events in drawing a card from a standard desk of 52 cards. (a) a seven (b) A black card (c) An ace or a king (d) a black two or a black three (e) a red face card

Solution:

Let S be the sample space of the experiment. Therefore, n(S) = 5

(a)A: a card with number 7 is drawn

There can be 4 ways to get a seven from a deck of 52 cards. n(A) = 4

by definition, $P(A) = \frac{n(A)}{n(S)} = \frac{4}{52}$

(b) B: a black card is drawn

There are 26 black cards in a deck. n(B) = 26

by definition, $P(B) = \frac{n(B)}{n(S)} = \frac{26}{52} = \frac{1}{2}$

(c) C: an ace or a king is drawn

There are 4 ace cards and 4 kings. So number of ways in which ace or king is drawn

= n(C) = 4+4= 8

by definition, $P(C) = \frac{n(C)}{n(S)} = \frac{8}{52} = \frac{2}{13}$

(d) D : a black 2 or black 3 is drawn.

There are 2 black cards having number 2 (i.e. spade and club)
There are 2 black cards having number 3 (i.e. spade and club)

n(C)=2+2=4

by definition, $P(D) = \frac{n(D)}{n(S)} = \frac{4}{52} = \frac{1}{13}$

(e) E: a red face card is drawn

There are 12 face cards (4x Jack, Queen, King)
Therefore, there are 6 red face cards. n(E)=6

by definition, $P(E) = \frac{n(E)}{n(S)} = \frac{6}{52} = \frac{3}{26}$

[4] If 2 fair dice are rolled, find the probability that sum of the numbers appearing on the uppermost faces of the dice is sum is (i) greater than 8 (ii) an odd number (iii) a perfect square

Solution:

Let S be the sample space of the experiment

S= {(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),
(2,1),(2,2),2,3),(2,4),(2,5),(2,6),
(3,1),(3,2),3,3),(3,4),(3,5),(3,6),
(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),
(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),
(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)}

(i)A: sum greater than 8

A = {(3,6),(4,5),(5,4),(6,3),(4,6),(5,5),(6,4),(5,6),(6,5);(6,6)}

Therefore, n(A)= 10

by definition, P(A) = $\frac{n(A)}{n(S)} = \frac{10}{36} = \frac{5}{18}$

(ii) B: sum is an odd number (i.e. sum is 3,5,7 or 9)

Sum 3: {(1,2),(2,1)}
Sum 5: {(1,4),(2,3),(3,2),(4,1)}
Sum 7: {(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)}
Sum 9: {(3,6),(4,5),(5,4),(6,3)}
Sum 11:{(5,6),(6,5)}

B = {(1,2),(2,1),(1,4),(2,3),(3,2),(4,1), (1,6),(2,5),(3,4),(4,3),(5,2),(6,1),
(3,6),(4,5),(5,4),(6,3), (5,6)(6,5)}

Therefore, n(B)= 18

by definition, P(B) = $\frac{n(B)}{n(S)} = \frac{18}{36} = \frac{1}{2}$

(iii) C: sum is a perfect square (i.e. sum is 4 or 9)

C={(1,3),(2,2),(3,1),(3,6),(4,5),(5,4),(6,3)}

by definition, P(C) = $\frac{n(C)}{n(S)} = \frac{7}{36}$

[5] A coin is tossed thrice. What is the probability of getting 2 or more heads?

Solution:

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

2 or more heads imply 2 or 3 heads as only 3 coins are tossed

A : Occurrence of 2 heads

B : Occurrence of 3 heads

A ={HHT,HTH,THH}

B = {HHH}

P(A) = $\frac{n(A)}{n(S)} = \frac{3}{8}$ and P(B) = $\frac{n(B)}{n(S)} = \frac{1}{8}$

As A and B are mutually exclusive, probability of getting 2 or more heads is

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

$= \frac{3}{8} + \frac{1}{8}$

$= \frac{1}{2}$

[6] From the past experience it is known that A can solve 3 examples out of given 5 and B can solve 4 examples out of given 7. An example is given to both of them to solve independently. Find the probability that a) the example remains unsolved b) the example is solved c) only one of them solves the example. [Ans. 6/35, 29/35, 17/35]

Solution:

X : A solves an example

Y: B solves an example

P(X)= $\frac{3}{5}$ Therefore, P($\bar{X}$)=1-P(X)=1- $\frac{3}{5} = \frac{2}{5}$

And P(Y)= $\frac{4}{7}$ P($\bar{Y}$)=1-P(Y)=1- $\frac{4}{7} = \frac{3}{7}$

(a) P(Example remains unsolved):

=P(A cannot solve the problem and B cannot solve the problem)
=P($\bar{X}$ I $\bar{Y}$)
=P($\bar{X}$).P($\bar{Y}$)……[As X and Yare independent, $\bar{X}$ and $\bar{Y}$ are independent

$= \frac{2}{5} x \frac{3}{7}$

$= \frac{6}{35}$

(b) P(Example is solved):

=P(A solves the problem or B solves the problem or both of them solve)
=P(XUY)
=P(X)+P(Y)-P(XIY)
=P(X)+P(Y)-P(X).P(Y)

$= \frac{3}{5} + \frac{4}{7} - \frac{3}{5} x \frac{4}{7}$

$= \frac{29}{35}$

(c) P(Only one of them solves):

=P(A solves the problem and B cannot solve it)+
P(B solves the problem and A cannot solve it)
=P(X I $\bar{Y}$) + P($\bar{X}$ I Y)

$= \frac{3}{5} + \frac{3}{7} + \frac{2}{5} x \frac{4}{7}$

$= \frac{17}{35}$

[7] Shankar is known to hit a target in 5 out of 9 shots whereas Hari is known to hit the same target in 6 out of 11 shots. What is the probability that the target would be hit once they both try?

Solution:

Let A : Shankar hits the target
B : Hari hits the target

P(A)= $\frac{5}{9}$ P(A)= $\frac{6}{11}$

And P (A I B) = P(A) x P(B)

$= \frac{5}{9} x \frac{6}{11}$

$= \frac{10}{33}$

The probability that the target would be hit is given by
P(A U B) = P(A) + P(B) – P(A I B)

$$= \frac{5}{9} + \frac{6}{11} - \frac{10}{33}$$
$$= \frac{79}{99}$$

[8] the following data pertains to analysis of 200 engineers

Age(years)	Bachelor's degree only	Master's degree	Total
Under 30	90	10	100
30 to 40	20	30	50
Over 40	40	10	50
Total	150	50	200

If one engineer is selected at random from the company. Find the probability that:
(a) he has only a bachelor's degree.
(b) he has a master's degree given that he is over 40.
(c) he is under 30 given that he has only a bachelor's degree.

Solution:

Let A: Engineer has a bachelor's degree only
B: Engineer has a master's degree
C: Engineer is under 30 years of age
D: Engineer is over 40 years of age

(a) $P(A) = \frac{150}{200} = \frac{3}{4}$

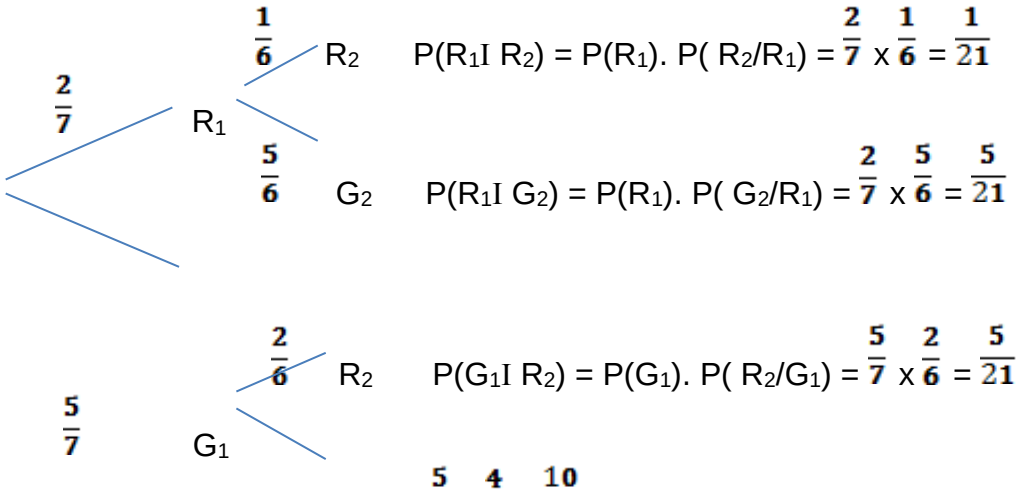
(b) $P(B/D) = \frac{P(B \cap D)}{P(D)} = \frac{\frac{10}{200}}{\frac{50}{200}} = \frac{1}{5}$

(c) $P(C/A) = \frac{P(C \cap A)}{P(A)} = \frac{\frac{90}{200}}{\frac{150}{200}} = \frac{3}{5}$

[9] A suitcase consists of 7 marbles, of which 2 are red and 5 are green. Two marbles are removed at random and their colour noted. The first marble is not replaced before the second is selected. Find the probability that the marbles will be (a) both red, (b) of different colours, (c) the same colour

Solution:
Let R_1: The first marble removed is red
R_2: The second marble removed is red
G_1: The first marble removed is green
G_2: The second marble removed is green

Therefore,



(a) P(Both marbles are red) = $P(R_1 I R_2) = P(R_1).\ P(R_2/R_1) = \frac{2}{7} \times \frac{1}{6} = \frac{1}{21}$

(b) P (Both marbles are of different colours) = $P(R_1 I G_2) + P(G_1 I R_2)$

$P(G_1 I R_2) = P(G_1).\ P(R_2/G_1) = \frac{5}{7} \times \frac{2}{6} = \frac{5}{21}$

$P(R_1 I G_2) = P(R_1).\ P(G_2/R_1) = \frac{2}{7} \times \frac{5}{6} = \frac{5}{21}$

P (Both marbles are of different colours) = $P(R_1 I G_2) + P(G_1 I R_2)$

$$= \frac{5}{21} + \frac{5}{21}$$

$$= \frac{10}{21}$$

(c) P(Both marbles are of same colour) = $P(R_1 I R_2) + P(G_1 I G_2)$

$$= \frac{1}{21} + \frac{10}{21}$$

$$= \frac{11}{21}$$

[10] Two sets of candidates are competing for the positions of board of directors in a company. The probabilities That the first and second set 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. The new product is introduced .What is the probability that the first set won?

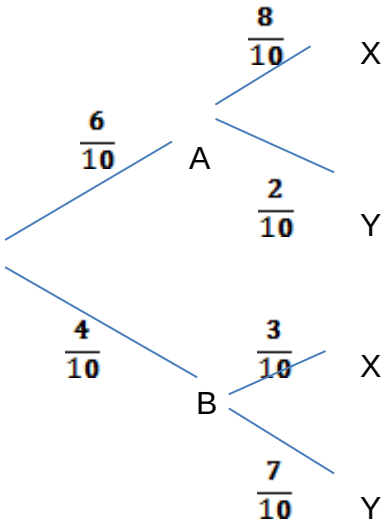
Solution:

Let A: First set wins P(A) = 0.6

B: Second set wins P(B) = 0.4

X: Product is introduced

Y: Product is not introduced.



By Baye's theorem,

P(First set wins given that the product is introduced) = P(A/X) = $\frac{P\left(\frac{X}{A}\right)P(A)}{P\left(\frac{X}{A}\right)P(A)+P\left(\frac{X}{B}\right)P(B)}$

$= \frac{(0.8)(0.6)}{(0.8)(0.6)+(0.3)(0.4)}$

$= \frac{8}{10}$

5 .(v) EXERCISES

- **Do as Directed:**

(1) Choose the correct alternative from the following

a) The value of probability is:

(A) Less than 0 (B) between 1and 10

(C) more then 0 (D) between 0 and1

b) All possible outcomes of a statistical experiment are called:

(A) Sample space (B) Hyper space

(C) Cyber space (D) Virtual space

c) Complimentary events are:

(A) Exhaustive (B) exclusive

(C) both A and B (D) none

d) If $A \bigcup B = S$ then A and B are said to be

(A) Exhaustive (B) exclusive

(C) both A and B (D) none

e) If $A \cap B = \phi$ then A and B are said to be

(A) Exhaustive (B) exclusive

(C) both A and B (D) none

(2) State whether the following are statistical experiment or not.
(a) Tossing of a coin.
(b) Throwing a fair die.
(c) Measuring no. of hands of undergraduate students.
(d) Selecting a committee of three persons from a group of eight persons.
(e) Measuring density of pure gold.
(f) Number of customers enters a mall.

(3) Write sample space for the following random experiment.
(a) Tossing of two coins.
(b) Throwing a fair dice.
(c) Selecting a card from pack of 52 cards.
(d) Number of customers enters a mall.

(4) In experiment of throwing a fair die:
A = {3,4} B = {4,5,6}
Give (a) A^c (b) $A^c \cap B^c$ (c) $(A \cup B)^c$ (d) $A^c \cap B$

(5) An insurance company insured 2,000 scooter drivers, 4,000 car drivers and 6,000 truck drivers. The probability of an accident is 0.01, 0.03 and 0.15 in the respective category. One of the insured driver meets an accident. What is the probability that he is a scooter driver?

(6) Consider a population of consumers consisting of two types. The upper class of consumers comprise 35% of the population and each member has a probability 0.8 of purchasing brand A of a product. Each member of the rest of the population has a probability 0.3 of purchasing brand A of the product. A consumer, chosen at random, is found to be buyer of brand A. What is the probability that the buyer belongs to the middle and lower class of consumers?

- **<u>Problems</u>:**

[1] A bag contains 7 white balls, 5 black balls and 4 red balls. If two balls are drawn at random from the bag, find the probability that a) both the balls are white b) One is black and other is red.
[Ans: 7/40, 1/6]

[2]A room has 3 lamps. From a collection of 10 light bulbs of which 6 are not good, a person select 3 at random and puts them in the sockets. What is the probability that he will have light from all the three lamps?
[Ans.: 1/30]

[3] P($\bar{A}$) = 2/3, P(B) = 1/4, P(AUB) = 5/12. Find P(A $\cap$ B), P(A/B) and P(B/A).
[Ans.: 1/6, 2/3, 1/2]

[4]A box contains 5 red and 4 blue balls and other box contains 4 red and 7 blue balls. A ball is selected at random from the first box and without noting the color put in the other. A ball is then drawn from the second box. What is the probability that it is blue?
[Ans.: 67/108]

[5] The probability that A will be alive 30 year hence is 0.3 and that B will be alive 30 years hence is 0.4 what is the probability that a) both will be alive b) only A will be alive c) both will not be alive d) at least one of them will be alive?
[Ans.: 0.12, 0.18, 0.42, 0.58]

[6] Students of a class were given a test in Economics & Accountancy. Of these students 20% failed in Economics, 15% failed in Accountancy and 5% failed in both. Find the chance that a student selected at random a) failed in at least one of the two subjects. b) failed in only Accountancy c) passed in both the subjects.
[Ans.: 0.3, 0.1, 0.7]

❖❖❖❖

6

RANDOM VARIABLE AND ITS PROBABILITY DISTRIBUTION

UNIT STRUCTURE

6. (i) Introduction to Probability Distribution
6. (ii) Various Probability Distribution
6. (iv) Exercises

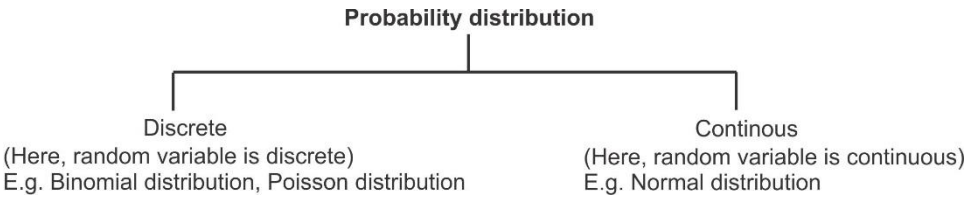
Aim and Objective :
1) Probability distribution are used both in theoretical aspects as well as practical aspects.
2) They are the fundamental concepts of statistics.

6. (i) INTRODUCTION :

Probability distribution can either be discrete or continuous. A discrete probability distribution is sometimes called a probability density function.

The study of a population can be done either by constructing an observed (or empirical) frequency distribution, often based on a sample from it, or by using a theoretical distribution. If a random variable satisfies the conditions of a theoretical probability distribution, then this distribution can be fitted to the observed data.

It finds applications in the understanding and analysis of a large number of business and economic situations. E.g. It is possible to test a hypothesis about a population, to take decision in face of uncertainty, to make forecasts, etc.



Probability distribution of a random variable X :

All possible values of a random variable 'X' along with their corresponding probabilities such that the sum of all these probabilities is unity, is called a probability distribution of the random variables 'X'.

In general,

X	X_1	X_2	X_3	.....	.....	.....	X_n
$P(X)$	P_1	P_2	P_3	.....	.....	.....	P_n

Where $P_i \geq O$ for all 'i'and $\sum_{i=1}^{n} P_i = 1$

Probability Mass Function :

Probability $P(X)$, such that the discrete random variable 'X' takes correspondence to every sample point in sample space S is called probability Mass Function (p, m, f,) of a random variable 'X'.

If $x_1, x_2, x_3, ...x_n$ are different values of a discrete random variable X and $P(x_1), P(x_2), P(x_3), ...P(x_n)$ are their respective probabilities such that

i) $P(x_i) \geq 0 \qquad i = 1, 2, 3....n$

ii) $\sum P(x_i) = 1 \qquad i = 1, 2, 3....n$

Then, $P(x)$ is known as the Probability Mass function of variable X.

Probability Density Function :

The probabilities associated with a continuous random variable X are determined by the probability density function $f(x)$ of a random variable X.

- **Properties of Probability density function '** $f(x)$**' :**

1) $f(x) \geq 0$ for all values of x

2) The probability that x will lie between two numbers 'a' and 'b' is equal to the area under the curve $y = f(x)$ between $x = a$ and $x = b$.

3) The total area under the entire curve $y = f(x)$ is equal to one.

Expectation and Variance of a random variable 'X' :

If a random variable X takes values $x_1, x_2, ...x_n$ with corresponding probabilities $P_1, P_2, ...P_n$ respectively, expectation of the random variable X is denoted by $E(X)$ and is given by

$$\sum(X) = x_1 p_1 + x_2 p_2 + ...x_n p_n$$

$$= \sum_{i=1}^{n} x_i p_i = \sum px$$

$$\sum(x^2) = \sum px^2$$

$E(X)$ is also known as expected value of X. If population mean is μ, then the expected value of x is given by,

$$E(X) = \mu$$

Variance of 'x'

$$V_{(x)} = \sigma^2 = E\left(x^2\right) - \left(E\left(x\right)\right)^2$$
$$= \sum px^2 - \left(\sum px\right)^2$$
$$= E\left(x^2\right) - \mu^2$$
$$= E\left(x - \mu\right)^2$$

Standard Deviation $S.D. = \sqrt{V_{(x)}} = \sqrt{\sum px^2 - \left(\sum px\right)^2}$

I) When x is discrete random variable with probability mass function $P\left(x\right)$, then its expected value is given by :

$\mu = \sum xP\left(x\right)$ and its variance is $\sigma^2 = E\left(x^2\right) - \mu^2$ where $E\left(x^2\right) = \sum x^2 P\left(x\right)$.

II) For a continuous random variable x defined in $\left[-\infty, \infty\right]$, its expected value (i.e. mean) and variance is given by,

$E\left(x\right) = \int_{-\infty}^{\infty} x\, f\left(x\right) dx$ and $\sigma^2 = E\left(x^2\right) - \mu^2$ where $E\left(x^2\right) = \int_{-\infty}^{\infty} x^2 f\left(x\right) dx$

- **Properties of Expected values**

1) Expectation of a constant K is K i.e. $E\left(x\right) = K$, for any constant K.

2) Expectation of sum of two random variables is sum of their expectation i.e. $E\left(x + y\right) = E\left(x\right) + E\left(y\right)$ for any two random variable 'x' & 'y'.

3) Expectation of the product of a constant and a random variable is the product of constant and the expectation of random variable. i.e. $E\left(K, x\right) = K.\ E\left(x\right)$ for any constant K.

4) Expectation of product of two random variables is the product of expectations of two random variables, provided the two variables are independent. i.e. $E\left(xy\right) = E\left(x\right) \times E\left(y\right)$ where x & y are independent.

E.g.

1) Two fair dice are rolled. If X denotes the sum of the numbers appearing on the uppermost faces on the dice, find

1) $P\left(X < 4\right)$
2) $P\left(X \geq 10\right)$
3) $P\left(3 < X < 7\right)$
4) $P\left(X > 3\right)$

If S is a sample space of experiment.

Solution :

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

$\therefore n(S)=36$

$\therefore P(X=2)=P\{(1,1)\}=\frac{1}{36}$

$P(X=3)=P\{(1,2),(1)\}=\frac{2}{36}$

The Probability distribution of X is given

X	2	3	4	5	6	7	8	9	10	11	12
$P(x)$	$1/36$	$2/36$	$3/36$	$4/36$	$5/36$	$6/36$	$5/36$	$4/36$	$3/36$	$2/36$	$1/36$

Here, $\sum_{x=2}^{12} P(X)=1$

1) $P(X<4)=P(X=2\ or\ X=3)$

$=P(X=2)+P(X=3)$

$=\frac{1}{36}+\frac{2}{36}$

$=\frac{1}{12}$

2) $P(X\geq 10)=P(X=10\ or\ X=11\ or\ X=12)$

$=P(X=10)+P(X=11)+P(X=12)$

$=\frac{3}{36}+\frac{2}{36}+\frac{1}{36}$

$=\frac{1}{6}$

3) $P(3<X<7)=P(X=4\ or\ X=5\ or\ X=6)$

$=P(X=4)+P(X=5)+P(X=6)$

$=\frac{3}{36}+\frac{4}{36}+\frac{5}{36}$

$=\frac{1}{3}$

4) $P(X<3)=P(X=4)+P(X=5)+...+P(X=12)$

But $\sum_{x=2}^{12} P(x)=1$

$P(X=2)+P(X=3)+....+P(X=12)=1$

$\therefore P(X=4)+P(X=5)+....+P(X=12)$

$=1-P(X=2)-P(X=3)$

$=1-\frac{1}{36}-\frac{2}{36}$

$=\frac{11}{12}$

Q.2 A random variable x has following probability distribution.

X	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
$P(x)$	0	2K	3K	K	2K	K^2	$7K^2$	$2K^2+K$

Find :

1) Value of K
2) $P(X<3)$
3) $P(X\geq 4)$
4) $P(2<x\leq 5)$

Solution :

1) $\sum P(x)=1$

$\therefore 0+2K+3K+K+2K+K^2+7K^2+2K^2+K=1$

$\therefore 10K^2+9K-1=0$

$(K+1)(10K-1)=0$

$K=1/10$

2) $P(X<3)=P(X=0)+P(X=1)+P(X=2)$

$=0+2K+3K$

$=5K$

$=0.5 \qquad [\because K=0.1]$

3) $P(X\geq 3)=P(X=4)+P(X=0)+P(X=7)$

$=2K+K^2+7K^2+2K^2+K$

$=10K^2+3K$

$=10(0.1)^2+3(0.1)$

$=0.4$

4) $P(2<x\leq 5)=P(X=3)+P(X=4)+P(X=5)$

$=K+2K+K^2$

$=K^2+3K$

$=0.31$

Q.3 If X is a random variable having probability mass function.

$$P(X=x)=\frac{x}{8}\ ;x=0,1$$
$$=\frac{K}{4}\ ,x=2$$
$$=\frac{Kx}{16};x=3$$

Find the value of K & $E(X)$

Solution :

$P(X=x)$ is p.m.f.

$$\therefore \sum_{x=0}^{3} P(X=x)=1$$
$$\therefore P(X=0)+P(X=1)+P(X=2)+P(X=3)=1$$
$$0+\frac{1}{8}+\frac{K}{4}+\frac{3K}{16}=1$$
$$\therefore 7K+2=16$$
$$\therefore K=2$$

$$P(X=x)=\frac{x}{8};x=0,1$$
$$=\frac{1}{2};x=2$$
$$=\frac{x}{8};x=3$$

Now, $E(x)=\sum_{x=0}^{3} x\ P(X=x)$

$$=O\times P(X=0)+1\times P(X=1)+2P(X=2)+3P(X=3)$$
$$=0+1\times\frac{1}{8}+2\times\frac{1}{2}+3\times\frac{3}{8}$$
$$=\frac{1}{8}+1+\frac{9}{8}$$
$$=\frac{9}{4}$$

Q.4 If X is a number appearing on the uppermost face of a fair dice. Find $E(X)$ and $V(X)$.

Solution :

Since X is a no. appearing on uppermost face of a fair dice, the probability distribution of X is given by

X	1	2	3	4	5	6

$P(X=x)$	1/6	1/6	1/6	1/6	1/6	1/6

$$E(X)=\sum x\ P(X=x)$$

$$=1\times\frac{1}{6}+2\times\frac{1}{6}+3\times\frac{1}{6}+4\times\frac{1}{6}+5\times\frac{1}{6}+6\times\frac{1}{6}$$

$$=\frac{1}{6}(1+2+3+4+5+6)$$

$$=\frac{21}{6}$$

$$=\frac{7}{2}$$

Now, $E(x^2)=\sum x^2P(X=x)$

$$=1^2\times\frac{1}{6}+2^2\times\frac{1}{6}+3^2\times\frac{1}{6}+4^2\times\frac{1}{6}+5^2\times\frac{1}{6}+6^2\times\frac{1}{6}$$

$$=\frac{1}{6}(1^2+2^2+3^2+4^2+5^2+6^2)$$

$$=\frac{91}{6}$$

$$V(X)=E(X^2)-[E(X)]^2$$

$$=\frac{91}{6}-\left(\frac{7}{2}\right)^2$$

$$=\frac{91}{6}-\frac{49}{4}$$

$$=\frac{35}{12}$$

Q.5 In a certain lottery, one price of `1000, 2 prizes of `500 each and 5 prizes of `100 each are to be awarded to 8 tickets drawn from the total number of 10,000 tickets sold at a price of `1 per ticket. Find the expected net gain to a person buying a particular ticket.

Solution :

Let the person get a prize of `X with probability $P(X)$

$P(x)$	**1000**	**500**	**100**	**0**
	$\frac{1}{10000}$	$\frac{2}{10000}$	$\frac{5}{10000}$	$\left[1-\frac{8}{10000}\right]$

$$E(x)=\sum x\,P(X=x)$$

$$=1000\times\frac{1}{10000}+500\times\frac{2}{10000}+100\times\frac{5}{10000}+0\times\left[1-\frac{8}{10000}\right]$$

$$=\frac{1}{4}$$

$$=0.25$$

∵ the person has to spend `1 for buying the ticket, his net expected gain = 0.25 - 1= ` - 0.75
i.e. the person will incur an expected loss of 0.75`.

6. (ii) SOME STANDARD DISTRIBUTIONS :

1) Binomial Distribution :

i) It is named after Swiss Mathematician Jacob James Bernoulli and is also known as “Bernoulli Trial”.

ii) Trial or Process, the literal meaning of word ‘Binomial is two groups’. Hence, in this distribution, frequencies are divided on basis of two aspects or two possible outcomes, called as “Success” and “Failure”.

iii) Each trial is associated with two mutually exclusive and exhaustive events i.e. $p+q=1$, one of them is called success (p) and other is called failure (q) E.g. When we toss a coin, there are only two possible outcomes - Head or tail and one of them must happen.

iv) Binomial distribution is symmetric when $p=q=0.5$.

v) The general form of binomial distribution is the expansion of $(p+q)^n$, in which the no. of success is written in a descending order. If the no. of success is written in ascending order, then $(q+p)^n$ will be expanded.

vi) The trials are independent.

vii) The random experiment is performed repeatedly for a fixed number of times. For example, a coin is tossed 8 times. In other words, the number of trials (n) are finite and fixed and it is a positive integer.

viii)If $x \square B(n,p)$

$$P(x\ success) = f(x) = {}^{n}C_{x} P^{x} q^{n-x} \quad (x = 0,1,2,...,n)$$

p : probability of success

q : probability of failure (q = 1 – p)

n : number of trials

x : no. of successes in ‘n’ trials

ix) It is a probability distribution in which each $p(r) \geq 0$

$$\therefore O \leq p(x) \leq 1$$

$$\therefore \sum p(x) = 1$$

x) Mean of $\mu = nP$ binomial distribution.

xi) Variance of Binomial Distribution

$$V_{(x)} = npq \quad \therefore V_{(x)} = \sigma_{(r)}^{\ 2}$$

$$\sigma_{(x)} = \sqrt{V_{(x)}} = \sqrt{npq}$$

Variance is always i.e. $npq < np$ less than mean

Eg (i) For a Binomial distribution, mean = S and standard deviation = 2. Find n and p.

Solution :

Mean = S;

S.d. = 2

$\therefore np = 5 \ldots\ldots\ldots (1)$

Variance = $2^2 = 4$

$\therefore npq = 4 \ldots\ldots\ldots (2)$

Dividing (2) by (1), we get

$$q = \frac{4}{5}$$

$$\therefore p = 1 - q = \frac{1}{5}$$

Mean $= np = 5 \ \therefore \boxed{n = 25}$ & $\boxed{p = \frac{1}{5}}$

ii) If X has a Binomial distributed with $F(X) = 2$ and $Var(X) = 4/3$, find the probability distribution of X.

Solution :

$$E(X) = np = 2$$

$$Var(X) = npq = \frac{4}{3}$$

$$\frac{npq}{np} = \frac{4}{6} \therefore q = \frac{2}{3}$$

$$\therefore p = 1 - q = \frac{1}{3}$$

$$np = 2 \therefore n = 6$$

Hence, the distribution is $P(X = x) = {}^{n}C_x p^x q^{n-x} = {}^{6}C_x \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{6-x}$

Putting $x = 0,1,2 \ldots 6$, we get the probability distribution of X.

X	**0**	**1**	**2**	**3**	**4**	**5**	**6**
$P(X = x)$	$\frac{64}{729}$	$\frac{192}{729}$	$\frac{240}{729}$	$\frac{160}{729}$	$\frac{60}{729}$	$\frac{12}{729}$	$\frac{1}{729}$

iii) Twelve cards, 4 Kings 4 Queens and 4 Jacks, are removed from a pack of cards and put in a box A card is picked up at random from this box, noted down and is replaced back in the box. Find the probability for each of the following when 6 cards are taken out, one by one.

a) P (at least one is a King)
b) P (S are Kings)
c) P (at most S are Kings)
d) P (none is a king)

Solution : Success is drawing a King out of 12 cards in a box.

$$p = P(success) = \frac{4}{12} = \frac{1}{3}$$

$$q = P(\ failure) = 1 - \frac{1}{3} = \frac{2}{3}$$

$$n = 6$$

a) P (at least one is a King)

$$= 1 - P(0)$$

$$= 1 - {}^{6}C_0 p^0 q^6$$

$$= 1 - \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^6$$

$$= 1 - \frac{64}{729}$$

$$= \frac{665}{729}$$

b) P(5 are Kings) = $P(5)$

$$= {}^{6}C_5 p^5 q^1$$

$$= {}^{6}C_5 \left(\frac{1}{3}\right)^5 \left(\frac{2}{3}\right)$$

$$= \frac{4}{243}$$

c) P(at most 5 are Kings) $= P(5) + P(6)$

$$= {}^{6}C_5 \left(\frac{1}{3}\right)^5 \left(\frac{2}{3}\right) + {}^{6}C_6 \left(\frac{1}{3}\right)^6 \left(\frac{2}{3}\right)$$

$$= \frac{13}{729}$$

d) P(none is a King) $= P(0)$

$$= {}^{6}C_5 \left(\frac{1}{3}\right)^5 \left(\frac{2}{3}\right)^6$$

$$= \frac{64}{729}$$

2) Poisson Distribution :

In events such as the accident in a factory, or the number of deaths in a city in one year by a rare disease, binomial distribution is not applicable because we don't know the value of 'n' in expression $(q + p)^n$. In order to deal such situations. Poisson distribution comes to our rescue. Here, one cannot count the number of times an event occurs or does not occur.

i) Poisson Distribution was originated by French Mathematician Simon Denis Poisson.

ii) The Poisson distribution is a discrete probability distribution in which the number of successes are given in whole numbers such as 0, 1,..,..,..,etc.

iii) Poisson distribution is used in those conditions where value of P is very small $(p \to o)$ and value of q is at most equal to 1 $(q \to 1)$ and value of n is large. E.g. behavior of rare events.

iv) A random variable x is defined to follow Poisson distribution with parameter 'm' and is given by $x \square P(m)$ then $P(x) = \frac{e^{-m} m^x}{x!}$.

v) 'm' is called uniparameter.

vi) e is transcendental value. It is an irrational number lying between 2 and 3 $(e \approx 2.718280)$

vii) Poisson distribution applies for discrete probability distribution i.e. $O \le P(r) \le 1 \quad \sum P(x) = 1$

viii) In practice, Poisson distribution may be used in place of binomial where $n \ge 100$ and $p \le 0.1$ and $0 < m < 1$.

ix) Mean of Poisson Distribution is given by,

$$\bar{x} = \mu m$$

x) Variance of Poisson Distribution

$V_{(x)} = m = n_p$

$\sigma_{(x)} = \sqrt{m} = \sqrt{n_p}$

m is always positive.

Mean = Expected value $\bar{x} = E(x)$

Coefficient of = $\frac{\sigma}{m} \times 100 = \frac{\sqrt{m}}{m} \times 100$ Variation.

xi) Applications are

a) In Biology, to count the number of bacteria.

b) In counting the number of defects per item in statistical quality control.

c) In determining the number of deaths due to a rare disease.

d) In insurance problems, to count the number of casualties.

E.g.

i) For a Poisson distribution with $m = 0.7$, find $P(X = 2), P(X \le 2)$, given $e^{-0.7} = 0.497$.

Solution :

$$P(X = x) = P(x) = \frac{e^{-m} m^x}{x!}$$

i.e. $P(X = 2) = P_{(2)} = \frac{e^{-0.7} (0.7)^2}{2!}$

$$P(X \le 2) = P(X = 0) + P(X = 1) + P(X = 2)$$
$$= P(0) + P(1) + P(2)$$
$$= \frac{e^{-0.7}(0.7)^0}{0!} + \frac{e^{-0.7}(0.7)^1}{1!} + \frac{e^{-0.7}(0.7)^2}{2!}$$
$$= 0.497 + 0.348 + 0.172$$
$$= 0.967$$

ii) If a random variable x follows Poisson distribution such that $P(1) = P(2)$, find its mean and variance.

Solution :

$$P(X = x) = P(x) = \frac{e^{-m}m^x}{x!}$$

$P(1) = P(2)$ ..... [given]

$$\therefore \frac{e^{-m}m^1}{1!} = \frac{e^{-m}m^2}{2!}$$

$m = 2$

Mean & variance are both 2.

iii) The average number of incoming telephone calls per minute in a stock firm is 2. Find the probability that during a given minute, 2 or more calls are received. $(e^{-2} = 0.135)$

Solution :

This is a Poisson distribution with mean m = 2.

$$P(x) = \frac{e^{-m}m^x}{x!}$$

$$P(X \ge 2) = P(X = 2) + P(X = 3) + ...$$
$$= P(2) + P(3) + ...$$
$$= 1 - [P(0) + P(1)]$$
$$= 1 - \left[\frac{e^{-2}2^0}{0!} + \frac{e^{-2}2^1}{1!}\right]$$
$$= 1 - [0.135 + 0.270]$$
$$= 1 - 0.405$$
$$= 0.595$$

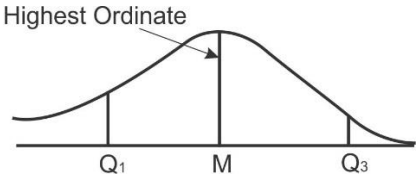
3) Normal Distribution :

i) It was originated by Carl Gauss and is also known as 'Gaussian distribution'.

ii) The normal distribution is symmetrical about x=0.

- **Properties of Normal Distribution :**

i) It is a bell shaped, symmetrical curve

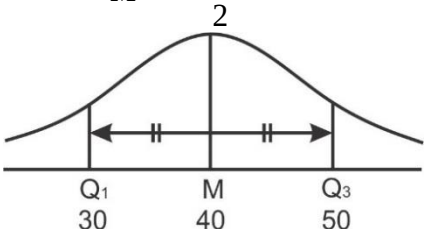


ii) The value of mean, median and mode are equal

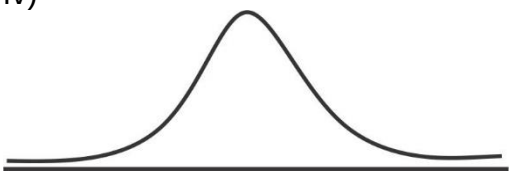
iii) Quartiles are situated at equal distances from median

$Q_3 - M = M - Q_1$

$$M = \frac{Q_3 + Q_1}{2}$$

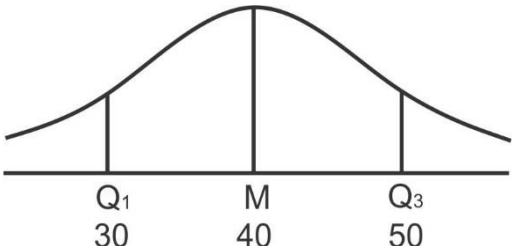


iv)



The curve is asymptotic (i.e. Curve comes closer & closer but never touches the base)

v) Semi Inter Quartile range (QD) is always 0.6745 times SD



$\therefore QD = 0.6745\, S.D.$

$$\therefore \frac{Q_3 - Q_1}{2} = 0.6745\sigma$$

i.e. $QD = \frac{2\sigma}{3}$

$Q_3 = \mu + 0.675\sigma$

$Q_1 = \mu - 0.675\sigma$

vi) Mean deviation M.D. = 0.7979 S.D.

$$M.D. = \frac{4}{5}\sigma$$

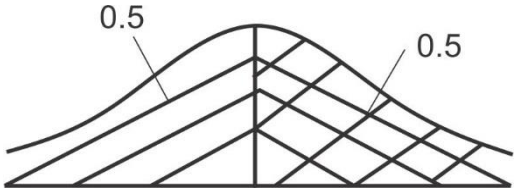
$$M.D. = 0.8\sigma$$

vii) QD:MD:SD = 10:12:15

viii)MD about mean, median and mode are equal.

ix) Normal Distribution is biparametric i.e. μ and S.D. σ are the parameters $x \square N(\mu\sigma)$ where x is a continuous variable E.g. Height, profit, rainfall, etc.

x) Area under standard 0.5 normal curve is 1
Area to left of the ordinate = Area to the right of the ordinate = 0.5



xi) Probability density function of x is given by PDF.

The equation of normal curve $y = \boxed{f(x) = \frac{1}{\sigma\sqrt{2\pi}} \times e^{-\frac{(x-\mu)^2}{2\sigma^2}}} \begin{pmatrix} -\infty < x < \infty \\ -\infty - \mu - \infty \\ \sigma > 0 \end{pmatrix}$

Where σ = Standard deviation
$\mu =$ the arithmetic mean
N= the number of observations

We can transform the variable x to $Z = \frac{x-\mu}{\sigma}$, then the equation takes the form

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}}; -\infty < Z < \infty$$

It is called P.D.F. for the standard normal variable Z. It is a normal variate with mean $\mu = 0$ and S.D. $\sigma = 1$.

xii) The standard normal distribution is also known as Unit Normal Distribution or Z- distribution.

Note :
1) x is any real number.
2) σ is always positive
3) Coefficient $= \frac{SD}{Mean} \times 100 = \frac{\sqrt{m}}{m} \times 100$ of variation Mean.
4) $\varnothing_{(1)} =$ mean area below Z=1 i.e. area between $-\infty$ to 1.
5) Probability = Area under the curve
6) % Area = Area $\times 100$
7) Standard Deviation $= \sqrt{\sigma_1^2 + \sigma_2^2}$

Methods of fitting a Normal Curve
1) Area method
2) Ordinate Method

- **Relation between Binomial & Poisson Distribution**

Conditions under which Binomial Distribution tends to become Poisson distribution are as follows :

i) When number of trials is very large $(n \to \infty)$

ii) When the occurrence of every trial is very small $(p \to 0, q \to 1)$

iii) When average number of success (m) is equal to positive finite quantity (np) $m = np$

- **Relationship between Binomial & Normal Distribution**

Conditions under which Binomial distribution tends to become Normal distribution.

i) n approaches to infinity $(n \to \infty)$

ii) Values of p and q are moderate

<u>Difference between Binomial, Poisson and Normal Distribution</u>

1) Number of Distributions
   Binomial, Poisson $\to$ Discrete Probability Distribution
   Normal $\to$ Continuous Probability Distribution

2) Definitions of Probability functions
   Binomial Distributions = $P_x = {}^n C_x p^x q^{n-x}$
   Poisson Distributions = $P_{(x)} = \dfrac{e^{-m} m^x}{x!}$
   Normal Distribution = $f(x) = \dfrac{1}{\sigma\sqrt{2r}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

3) Value of no. of trials 'n'
   Binomial : finite
   Poisson, Normal : Infinite / Very large $P(x)$

4) Parameters
   2 in Binomial : n, p
   2 in Normal : m, σ
   1 in Poisson : m

***Basic Comparison*: <u>Binomial</u>, <u>Poisson,Normal Distribution</u>**

1) Types of Distribution	Discrete	Discrete	Continuous
2) Parameters :	Two n, p	One m	Two μ, σ

3) Restriction $O < P < 1\ m > 0$ $\mu = 0$, $\sigma = 1$

on parameter

4) Mean μ : $\mu = np$ $\bar{X} = m = \boxed{n \not{p}} = 1$

5) Variance : $\sigma^2 = npq\ \sigma^2 = m = \boxed{n \not{p}}\ \sigma^2$

$\left(\sigma^2\right)$

6) Probability Function: $P_{(x)} = {}^{\eta} C_x P^x q^{n-x}$ $P_{(x)} = \dfrac{e^{-m} m^x}{x!}$ $P_{(x)} = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-x)^2}{2\sigma}}$

$x = 0,1,2....\infty$ *where* $\sigma = SD$

$m = Mean$ $\mu = mean$

E.g.

1) If X follows normal distribution with mean 120 and variance 1600, find,

i) $P\left(X \le 140\right)$

ii) $P\left(X \ge 110\right)$

iii) $P\left(100 \le X \le 130\right)$

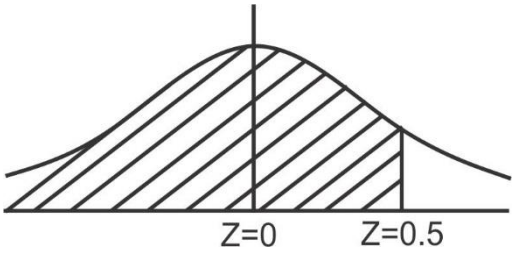
Area between Z = 0 and Z = 0.25 = 0.0987

Area between Z = 0 and Z = 0.5 = 0.1916

Solution :

$\because$ X follows normal distribution with mean 120 and variance 1600.

$\therefore \mu = 120\ \&\ \sigma = \sqrt{1600} = 40$



i) $P\left(X \le 140\right)$

$= P\left(\dfrac{X - 120}{40} \le \dfrac{140 - 120}{40}\right)$

$= P\left(Z \le 0.5\right)$

= Area to left of Z=0.5

= (Area to left of Z = 0) + (Area between Z = 0 & Z = 0.5)

= 0.5 + 0.1961

= 0.6916

ii) $P\left(X \ge 110\right)$

$= P\left(\frac{X-120}{40} \geq \frac{110-120}{40}\right)$

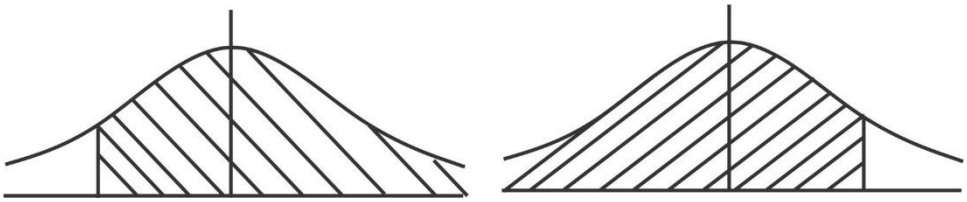
$= P(Z \geq -0.25)$

$= P(Z \leq 0.25)$

= (Area to left of Z = 0) + (Area between Z = 0 & Z = 0.25)

= 0.5 + 0.0987

= 0.598



iii) $P(100 \leq X \leq 130)$

$= P\left(\frac{100-120}{40} \leq \frac{X-120}{40} \leq \frac{130-120}{40}\right)$

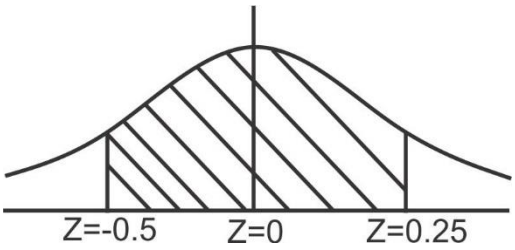
$= P(-0.5 \leq Z \leq 0.25)$

= Area between Z = -0.5 and Z = 0.25

= (Area between Z = -0.5 and Z = 0) + (Area between Z = 0 and Z = 0.25)

= 0.1916 + 0.0987

= 0.2903



2) If the weights of 10,000 soldiers in a regiment are normally distributed with a mean of 72kgs and a standard deviation of 5kgs. How many soldiers have weights above 80 kgs? Also find percentage of soldiers with weights between 70 and 77kgs. Given that

Area (Z = 0 to Z = 1.6) = 0.4452

Area (Z = 0 to Z = 0.4) = 0.1554

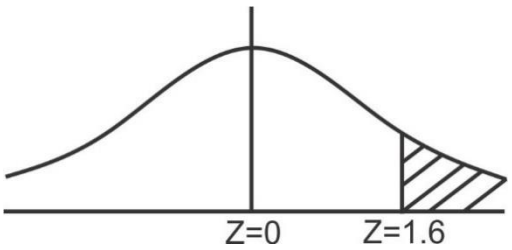
Area (Z = 0 to Z = 1) = 0.3413

Solution :

Let X be the weights of soldiers in the regiment. Let N be no. of soldiers. Let $\mu \& \sigma$ be the mean weight and SD of weights of the solders.

$\therefore N = 10,000, \mu = 72, \sigma = 5$

i) To find no of soldiers with weights above 80kgs, we find the probability that a soldier has weight above 80kgs.



$$P(X > 80)$$

$$= P\left(\frac{X-72}{5} > \frac{80-72}{5}\right)$$

$$= P(Z > 1.6)$$

= (Area of right to Z = 0) - (Area between Z = 0 and Z = 1.6)
= 0.5 - 0.4452
= 0.0548

∴ No. of soldiers with weight above 80kgs

$$= N \times P(X > 80)$$

$$= 10,000 \times 0.0548$$

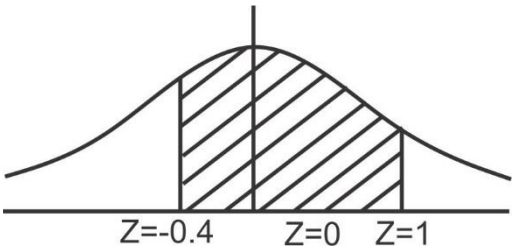
$$= 548$$

ii) To find the percentage of soldiers with weights between 70 and 77kgs.
i.e. $P(70 < x < 77)$

$$= P\left(\frac{70-72}{5} < \frac{X-72}{5} < \frac{77-72}{5}\right)$$

$$= P(-0.4 < Z < 1)$$

= (Area between Z = 0.4 and Z = 0) + (Area between Z = 0 and Z = 1)
= 0.1554 + 0.3413
= 0.4967



∴ % of soldiers with weight between 70 & 77kgs

$$= 100 \times P(70 < X < 77)$$

$$= 100 \times 0.4967$$

$$= 49.67$$

3) The incomes of a group of 10,000 persons were found to be normally distributed with mean Rs. 520 and standard deviation Rs. 60. Find the number of persons having incomes between Rs. 400 and 550 given that Area (Z = 0 to Z = 2) = 0.4772 & Area (Z = 0 to Z = 0.5) = 0.1915

Solution :

Standard Normal Variate $Z = \frac{X-m}{\sigma} = \frac{X-520}{60}$

When $x = 400$, $Z = \frac{400 - 520}{60} = -2$

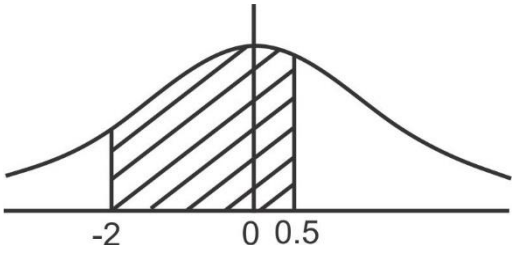
$x = 550$, $Z = \frac{550 - 520}{60} = 0.5$

$\therefore P(400 \le X \le 550)$

= Area $(Z = -2 \text{ to } o)$ + Area (Z = 0 to Z = 0.5)

= 0.4772 + 0.1913

= 0.6687



$\therefore$ The number of persons whose incomes are between Rs. 400 and Rs. 550.

$= Np$

$= 10000 \times 0.6687$

$= 66.87$

6. (iv) EXERCISES

(A) Multiple Choice Questions

[1] For a Binomial distribution, mean is 10 and n=30 then the probability of failure q is
(a) 1/3 (b) 3 (c) 2/3 (d) -1/3

[2] For a poission distribution with P (2) =P (3), then its Mean and Variance are
(a) 2 and 3 (b) 3 and 3 (c) 3 and 2 (d) 2 and 2

[3] The average number of incoming telephone calls at a switch board is 2 per minute. Find the probability that, during a given minute, less than 2 calls are received. (e^{-2}=0.135)
(a) 0.5 (b) 0.135 (c) 0.27 (d) 0.405

[4] When X is continuous function, then f(x) is called
(a) Probability mass function (b) probability density function
(c) both (d) None of these

[5] If x and y are independent, then
(a) E(xy)=E(x) x E(y) (b) E(xy)=E(x)+E(y)
(c) E(x +y)=E(x)+E(y) (d) E(x-y)=E(x) -E(y)

[6] Number of misprints per page of a thick book follows
(a) Binomial distribution (b) Normal distribution
(c) Poission distribution (d) Standard Normal distribution

[7] The total area of normal curve is
(a) one (b) 50 per cent
(c) 0.50 (d) any value between 0 and1

[8] The mean and mode of a normal distribution

(a) may be equal (b) may be different
(c) are always equal (d) (a) or (b)

[9] For a poission distribution,
(a) mean and standard deviation are equal
(b) mean and variance are equal
(c) standard deviation and variance are equal
(d) both (a) and (b)

[10] Which one is uniparametric distribution?
(a) Binomial distribution (b) Normal distribution
(c) Poission distribution (d) Hyper geometric distribution

[11] Probability Distribution may be
(a) discrete (b) continuous (c) infinite (d) (a) or (b)

[12] An unbiased coin is tossed 6 times. Find the probability of getting at least 4 tails.
(a) $\frac{11}{32}$ (b) $\frac{3}{8}$ (c) $\frac{21}{32}$ (d) $\frac{1}{16}$

(B) Problems

[1] Write down probability distribution of random variables **X** defined as number of heads appeared when four coins are tossed simultaneously.

[2] Two fair dice are rolled. X denotes the sum of the numbers appearing on the uppermost faces of the dices. From the probability distribution of X find a) P(X is a multiple of 3) b) P(X<5) c) P (5<X<10) d) P(X>4).

[**3]** A distributer makes a profit of Rs. 30 on each item that is received in perfect condition and suffers a loss of Rs. 6 on each item that is received in less-than-perfect condition. If each item received is in perfect condition with probability 0.4, what is the distributor's expected profit per item?

[4] In a game of throwing a fair dice, A wins Rs.60/- if a 6 is thrown. He gains Rs.30/- if the dice show 2 or 4 and he loses Rs.30/- if odd numbers occurs on the uppermost face of the dice. Find the expected gain of A?

[5] If the mean and variance of a Binomial distribution are 4 and 2.4 respectively, find the probability of (1) 5 successes (2) at least nine successes.

[6]An unbiased cubical dice is thrown 5times and the number appearing on its uppermost face is noted. Find the probability that the number of times an even appear is (1) 3 times (2) all 5 times.

[7] A student attempts an online test of 20 multiple-choice independent questions. Each question has 4 possible answers of which only one is correct. Find probability that (i) he has exactly 2 answers correct (ii) 3 or 4 answers correct(iii) none of the answer is correct.

[8] A Poisson distribution has standard deviation 3.Find P(o) and P (1).(Given: e^{-9} =0.000123).

[9] It is observed that 1% of mangoes in a box are bad. Find the probability that in a box of 100 mangoes, number of bad mangoes is (i) nil (ii) only 1 (iii) less than 2 (iv) more than 2. (Given: e^{-1} =0.3679)

[10] In an intelligence test administered to 1000 persons, the average I.Q was 100 with a standard deviation of 15. (a) How many people had their I.Q. between 70 and 110? (b) What is the percentage of person with I.Q. above 110?

Given : Area between z=0 and z=2 is 0.4772

: Area between z=0 and z=0.67 is 0.2486

[11] 1500 candidates appeared for a certain examination. The mean marks were 58 with a standard deviation of 5 marks. Assuming that the distribution of marks to be Normal. Find (i) the proportion of students securing more than 63 marks. (ii) the number of students securing marks between 60 and 68. (iii) the percentage of students with marks below 53.

Given : Area between z=0 and z=1 is 0.3413

: Area between z=0 and z=2 is 0.4772

: Area between z=0 and z=0.4 is 0.1554

❖❖❖❖

MODULE – V

7

INTRODUCTION TO SAMPLING AND REASONS FOR SAMPLING

Unit Structure :

Sampling & Sampling Distribution
Introduction to Sampling, Reason for Sampling, Sampling Design Process, Random Sampling vs Non Random Sampling, Sampling Types & Methods, Sampling Distribution, Central Limit Theorem

7.1.1 OBJECTIVES

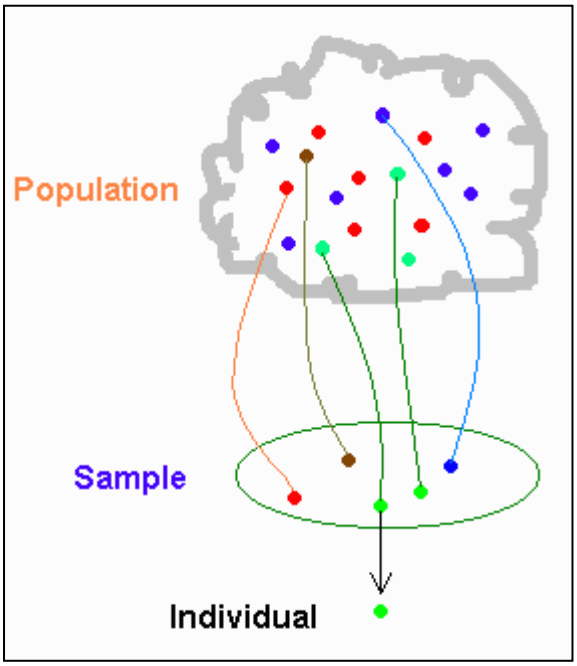
At the end of this unit the learners will be able to

- Understand the meaning of Sampling
- Understand the Reasons for Sampling.

7.1.2 INTRODUCTION TO SAMPLING

In statistics population refers to the total universe of objects being studied. Examples include: all votes in India or all possible outcomes when a dice is thrown. In reality it is not always possible to study the whole population and practically such study is not feasible. Sampling is widely used by researchers in business as a means of gathering useful information about a population. Data are gathered from samples and conclusions are drawn about the population as a part of the inferential statistics process.

A sample can be cheaper to obtain than a census for a given magnitude of questions. For example, if an eight-minute telephone interview is being undertaken, conducting the interviews with a sample of 100 customers rather than with a population of 100,000 customers obviously is less expensive. In addition to the cost savings, the significantly smaller number of interviews usually requires less total time. Thus, if obtaining the results is a matter of urgency, sampling can provide them more quickly. With the volatility of some markets and the constant barrage of new competition and new ideas, sampling has a strong advantage over a census in terms of research turnaround time.



7.1.3 REASONS FOR SAMPLING

Researchers usually cannot make direct observations of every individual in the population they are studying. Instead, they collect data from a subset of individuals – a sample – and use those observations to make inferences about the entire population.

Our knowledge, our attitudes, and our actions are based to a very large extension samples. This is equally true in everyday life and in scientific research. A person's opinion of an institution that conducts thousands of transactions everyday is often determined by the one or two encounters he has had with the institution in the course of several years; Travelers who spend 10 days in a foreign country and then proceed to write a book telling the inhabitants how to revive their industries, reform their political system, balance their budget, and improve the food in their hotels are a familiar figure of fun. But in a real sense they differ from the political scientist who devotes 20 years to living and studying in the country only in that they base their conclusions on a much smaller sample of experience and are less likely to be aware of the extent of their ignorance. In science and human affairs alike we lack the resources to study more than a fragment of the phenomena that might advance our knowledge.

Ideally, the sample corresponds to the larger population on the characteristic(s) of interest. In that case, the researcher's conclusions from the sample are probably applicable to the entire population. This type of correspondence between the sample and the larger population is most important when a researcher wants to know what proportion of the population has a certain characteristic – like a particular opinion or a demographic feature. Public opinion polls that try to describe the percentage of the population that plans to vote for a particular candidate, for example, require a sample that is highly representative of the population.

Taking a sample instead of conducting a census offers several advantages.

1. The sample can save money.

2. The sample can save time.

3. For given resources, the sample can broaden the scope of the study.

4. Because the research process is sometimes destructive, the sample can save product.

5. If accessing the population is impossible, the sample is the only option.

A sample can be cheaper to obtain than a census for a given magnitude of questions.

For example, if an eight-minute telephone interview is being undertaken, conducting the interviews with a sample of 100 customers rather than with a population of 100,000 customers obviously is less expensive. In addition to the cost savings, the significantly smaller number of interviews usually requires less total time. Thus, if obtaining the results is a matter of urgency, sampling can provide them more quickly. With the volatility of some markets and the constant barrage of new competition and new ideas, sampling has a strong advantage over a census in terms of research turnaround time.

If the resources allocated to a research project are fixed, more detailed information can be gathered by taking a sample than by conducting a census. With resources concentrated on fewer individuals or items, the study can be broadened in scope to allow for more specialized questions. One organization budgeted Rs.100, 000 for a study and opted to take a census instead of a sample by using a mail survey. The researchers mass-mailed thousands of copies of a computer card that looked like a Major League Baseball all-star ballot. The card contained 20questions to which the respondent could answer Yes or No by punching out a perforated hole.

The information retrieved amounted to the percentages of respondents who answered Yes and No on the 20 questions. For the same amount of money, the company could have taken a random sample from the population, held interactive one-on-one sessions with highly trained interviewers, and gathered detailed information about the process being studied. By using the money for a sample, the researchers could have spent significantly more time with each respondent and thus increased the potential for gathering useful information.

Some research processes are destructive to the product or item being studied. For example, if light bulbs are being tested to determine how long they burn or if candy bars are being taste tested to determine whether the taste is acceptable, the product is destroyed.

If a census were conducted for this type of research, no product would be left to sell. Hence, taking a sample is the only realistic option for testing such products.

Sometimes a population is virtually impossible to access for research. For example, some people refuse to answer sensitive questions, and some telephone numbers are unlisted. Some items of interest are so scattered that locating all of them would be extremely difficult. When the population is inaccessible for these or other reasons, sampling is the only option.

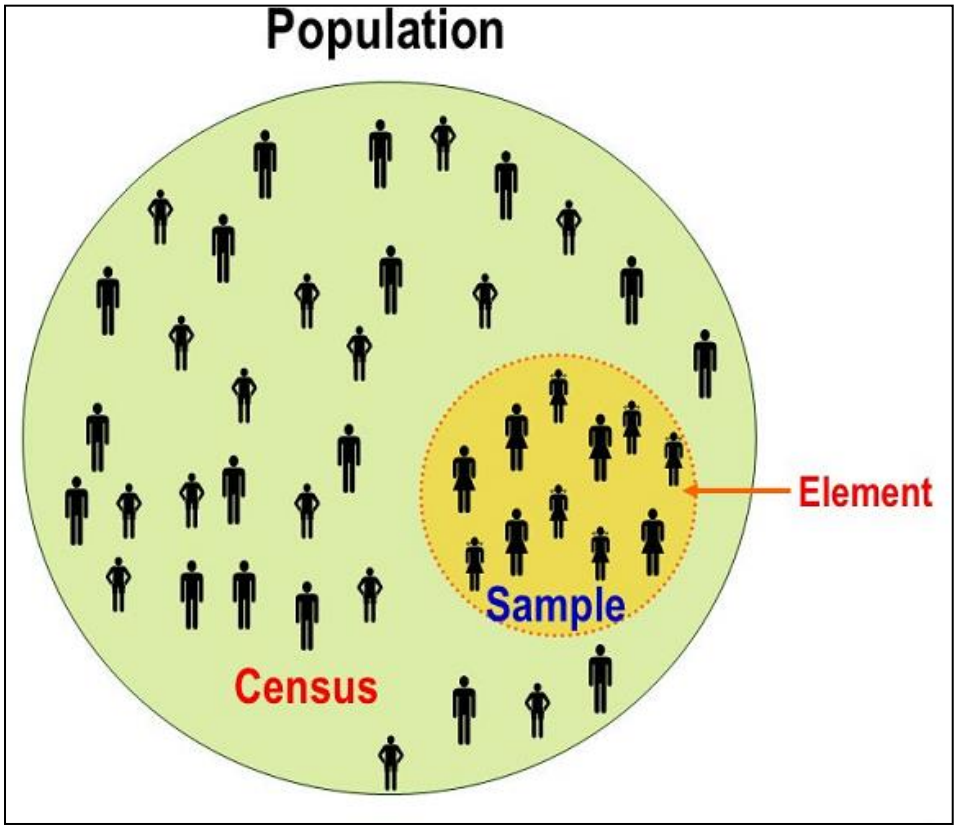
7.1.4 SAMPLING DESIGN PROCESS

- **Distinguishing Between a Sample and a Population**

Before describing sampling procedures, we need to define a few key terms. The term **population** means *all* members that meet a set of specifications or a specified criterion. For example, the population of India is defined as all people residing in India. The population of Tamil Nadu means all people living within the Tamil Nadu state boundary. A population of inanimate objects can also

exist, such as all automobiles manufactured in India in the year 2017. A single member of any given population is referred to as an **element.** When only some elements are selected from a population, we refer to that as a **sample;** when all elements are included, we call it a **census**.

Let's look at a few examples that will clarify these terms.



Two research psychologists were concerned about the different kinds of training that graduate students in clinical psychology were receiving. They knew that different programs emphasized different things, but they did not know which clinical orientations were most popular. Therefore, they prepared a list of all doctoral programs in clinical psychology (in India) and sent each of them a questionnaire regarding aspects of their program. The response to the survey was excellent; nearly 95% of the directors of these programs returned the completed questionnaire. The researchers then began analyzing their data and also classifying schools into different clinical orientations: psychoanalytic, behaviorist, humanistic, rogueries, and so on. When the task was complete, they reported the percentage of schools having these different orientations and described the orientations that were most popular, which were next, and so on. They also described other aspects of their data. The study was written up and submitted for publication to one of the professional journals dealing with matters of clinical psychology. The editor of the journal read the report and then returned it with a letter rejecting the manuscript for publication. In part, the letter noted that the manuscript was not publishable at this time because the proper statistical analyses had not been performed. The editor wanted to know whether the differences in orientation found among the different schools were significant or if they were due to chance

The researchers were unhappy, and rightly so. They wrote back to the editor, pointing out that their findings were not estimates based on a sample. They had surveyed all training programs (that is, the population). In other words, they had obtained a census rather than a sample. Therefore, their data were exhaustive; they included all programs and described what existed in the real world. The editor would be correct only if they had sampled some schools and then wanted to generalize to all schools. There searchers were not asking whether a sample represented the population; they were dealing with the population.

A comparable example would be to count all students (the population) enrolled in a particular university and then report the number of male and female students. If we found that 60% of the students were female, and 40% male, it would be improper and irrelevant to ask whether this difference in percentage is significantly different from chance. The fact is that the percentages that exist in the school population are parameters. They are not estimates derived from a sample. Had we taken a small sample of students and found this 60/40 split, it would then be appropriate to ask whether differences this large could have occurred by chance alone.

Data derived from a sample are treated statistically. Using sample data, we calculate various statistics, such as the mean and standard deviation. These sample statistics summarize (describe) aspects of the sample data. These data, when treated with other statistical procedures, allow us to make certain inferences. From the sample statistics, we make corresponding estimates of the population. Thus, from the sample mean, we estimate the population mean; from the sample standard deviation, we estimate the population standard deviation.

The above examples illustrate a problem that can occur when the terms population and sample are confused. The accuracy of our estimates depends on the extent to which the sample is representative of the population to which we wish to generalize.

- **Definitions**

Population: It is the totality of the objects or individuals regarding inferences are made in a sampling study.

Sample: It is the smaller representation of a large whole.

Sampling: It is a process of selecting a subset of randomized number of the members of the population of a study.
Sampling Frame / Source List: It is a complete list of all the members/ units of the population from which each sampling unit is selected.

Sample Design / Sample Plan: It is a definite plan for obtaining a sample from a given population.

Sampling Unit: It is a geographical one (state, district).

Sample Size: It is the number of items selected for the study.

Sampling Error: It is the difference between population value and sample value.

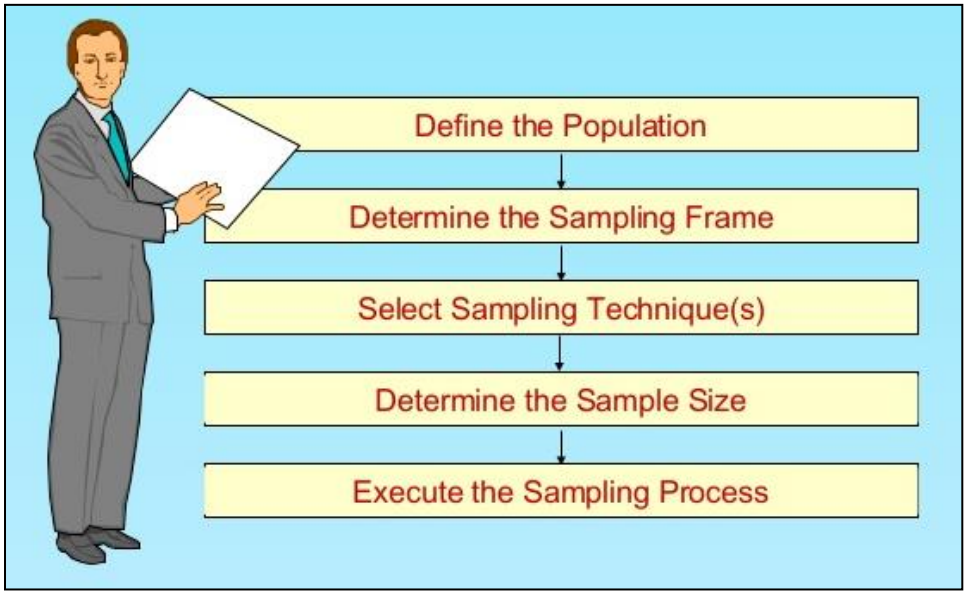
Sampling Distribution: It is the relative frequency distribution of samples.

Census: It is the collection of data from whole population.

Sampling: It is taking any portion of a population or universe as representative of that population.

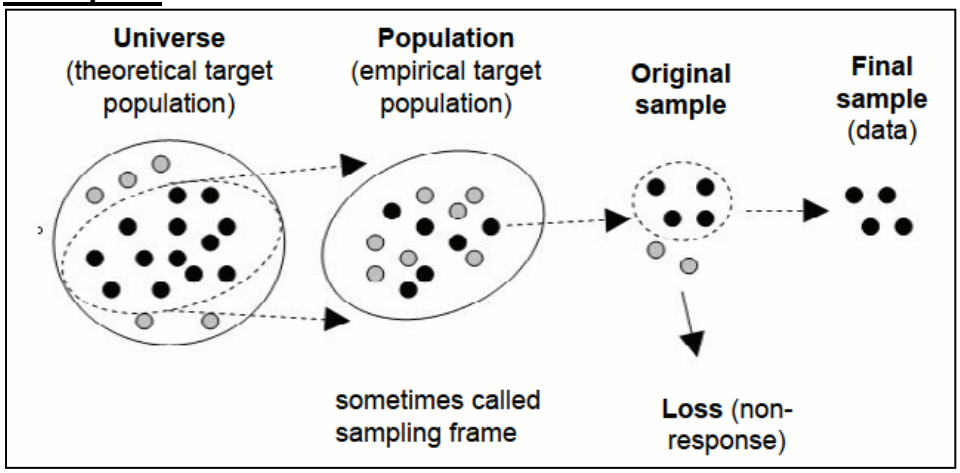
Sampling method has been using in social science research since1754 by A.L. Bowley.

The **sampling design process** includes five steps which are closely related and are important to all aspect of any research project. The five steps are: defining the **target population**; determining the **sample** frame; selecting a **sampling** technique; determining the **sample** size; and executing the **sampling process**.

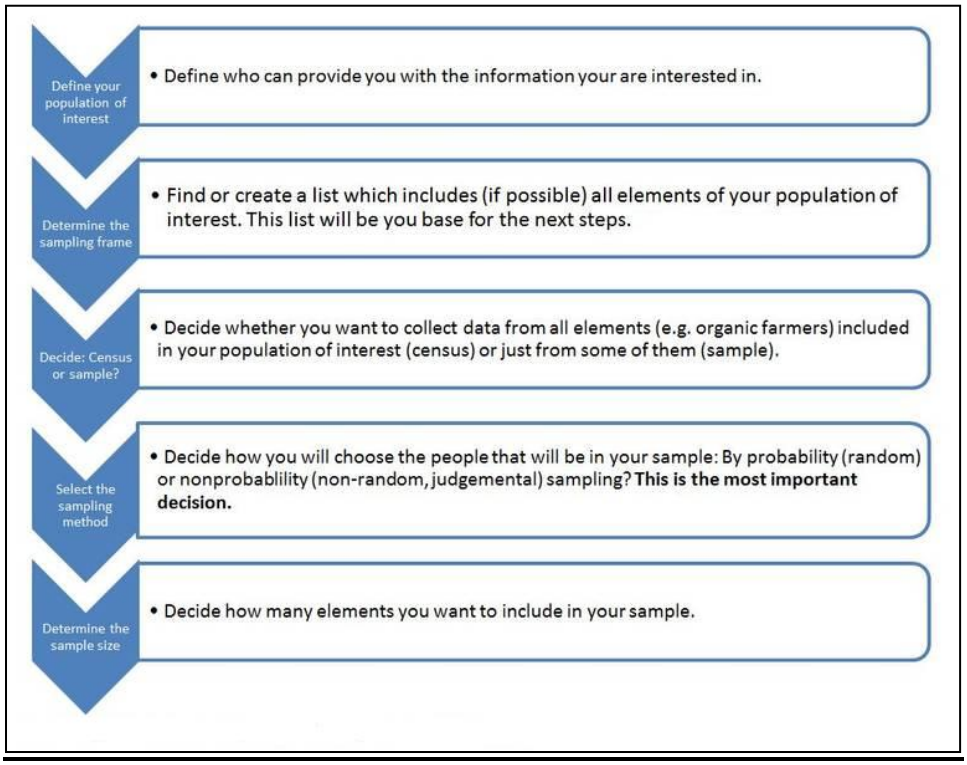


- **Examples of Sampling Designing Process**

Example 1



Example 2



Example 3

Target Population: All women who delivered live infants in Cameron County and Matamoros during the study period (August 21st -November 9th, 2005)

↓

Study Population: Women who delivered live infants during the study period in Cameron County and Matamoros in hospitals with 100 or more deliveries in 2004
- 4 hospitals in Cameron County
- 6 hospitals in Matamoros

↓

Sampling: 84 blocks of 2 consecutive days selected during the study period
- Cameron County: 36 blocks (9 in each study hospital)
- Matamoros: 48 blocks (8 in each study hospital)

↓

Women Chosen for the Study: All women who delivered live infants during any of the selected blocks of 2 consecutive days (n = 999)
- Sample size for Cameron County = 525
- Sample size for Matamoros = 474

Example 5

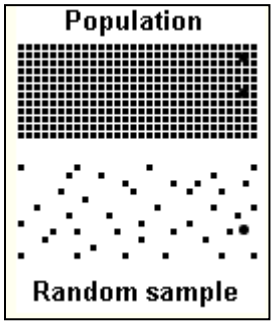
Sampling Design

Sample	• A subset, or some part, of a larger population. • A finite subset of a population selected from if which the objective of investigating its population is called sample of that population.
Population / Universe	• A complete group of entities . • All items in any field of inquiry.
Census	• An investigation of all the individual elements making up a population.
Sampling / Sampling Frame	• Sampling may be defined as the process of obtaining information about an entire population by examining only a part of it. In any investigation, if the data are collected from a representative part of the universe, the data is collected by sampling

From the definition of aims of the study and from secondary research the researcher would already have a group of people, organizations or maybe companies (so-called 'elements') in mind from whom the data can be collected for statistical investigation. This group is called as the '**population of interest**' (in some textbooks it is called the '**target population**').If the group population of interest is rather small, it might be possible to collect data from all elements in this group – this is called a **census**. However, in most cases the researcher will collect data from only a few elements – these then make up the **sample**. During sampling, the researcher will also have to decide about the **sampling unit**: For example, the elements in the sample are individual people, but the researcher might want to know about household spending on organic products. In this case, the sampling unit would be households. Sampling is easiest when the researcher already have an existing **sampling frame**, which is a list of all the elements included in the population of interest (i.e. a list of all organic farmers in the country). If such a list does not yet exist, the phrase 'sampling frame' refers to the procedure of creating this list.

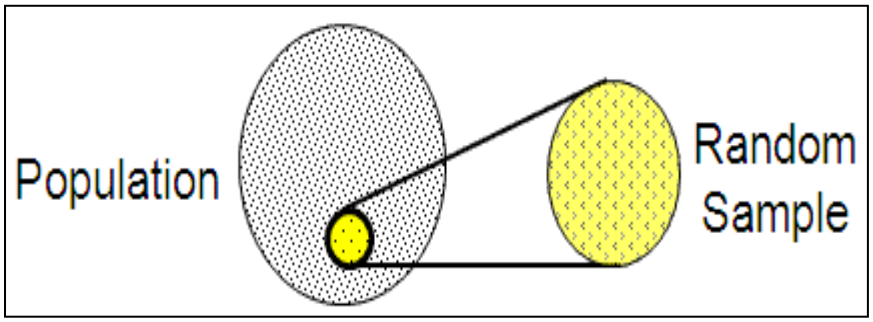
7.1.5 RANDOM SAMPLING

Random sampling is a part of the sampling technique in which each sample has an equal probability of being chosen. A sample chosen randomly is meant to be an unbiased representation of the total population. If for some reasons, the sample does not represent the population, the variation is called a sampling error.



Random sampling is one of the simplest forms of collecting data from the total population. Under random sampling, each member of the subset carries an equal opportunity of being chosen as a part of the sampling process. For example, the total workforce in organizations is 300 and to conduct a survey, a sample group of 30 employees is selected to do the survey. In this case, the population is the total number of employees in the company and the sample group of 30 employees is the sample. Each member of the workforce has an equal opportunity of being chosen because all the employees which were chosen to be part of the survey were selected randomly. But, there is always a possibility that the group or the sample does not represent the population as a whole, in that case, any random variation is termed as a sampling error.

An unbiased random sample is important for drawing conclusions. For example when we took out the sample of 30 employees from the total population of 300 employees, there is always a possibility that a researcher might end up picking over 25 men even if the population consists of 200 men and 100 women. Hence, some variations when drawing results can come up, which is known as a sampling error. One of the disadvantages of random sampling is the fact that it requires a complete list of population. For example, if a company wants to carry out a survey and intends to deploy random sampling, in that case, there should be total number of employees and there is a possibility that all the employees are spread across different regions which make the process of survey little difficult.



7.1.6 Non-Random Sampling

In random sampling, every item in a population has a known chance of being included in a sample.

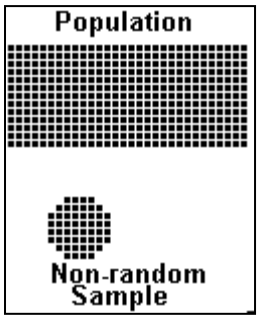
In non-random sampling this is not the case. Indeed, one of the main criticisms of non-random sampling is: because it's non-random, bias is almost certainly introduced.

A sample is said to be biased if:

NOT ALL OUTCOMES HAVE A KNOWN CHANCE OF OCCURRING OR IF SOME OUTCOMES HAVE A ZERO CHANCE OF OCCURRING.

Non-random sampling is useful when descriptive comments about the sample itself are desired.

However, it can be difficult to draw conclusions about the population based on information derived from a sample, as samples are often unrepresentative of the population.



7.1.7 EXERCISE

Q.1. Explain the concept and the need of sampling.

Q.2. Explain with examples sampling design process.

Q.3. What do you understand by random sampling?

Q.4. What do you understand by non-random sampling?

❖❖❖❖

8

SAMPLING TYPES AND METHODS – RANDOM SAMPLING & NON- RADOM SAMPLING

Unit Structure :

8.2.1 Sampling types and Methods - Random Sampling
- a. Simple Random Sampling
- b. Stratified Random Sampling
- c. Systematic Sampling
- d. Cluster (or Area) Sampling
- e. Multi-Stage Sampling

8.2.2 Sampling Types and Methods – Non-Random Sampling
- a. Convenience sampling
- b. Judgment sampling
- c. Quota Sampling
- d. Snowball Sampling
- e. Purposive Sampling
- f. Haphazard Sampling

8.2.3 a. Sampling Distribution
- b. Central Limit Theorem

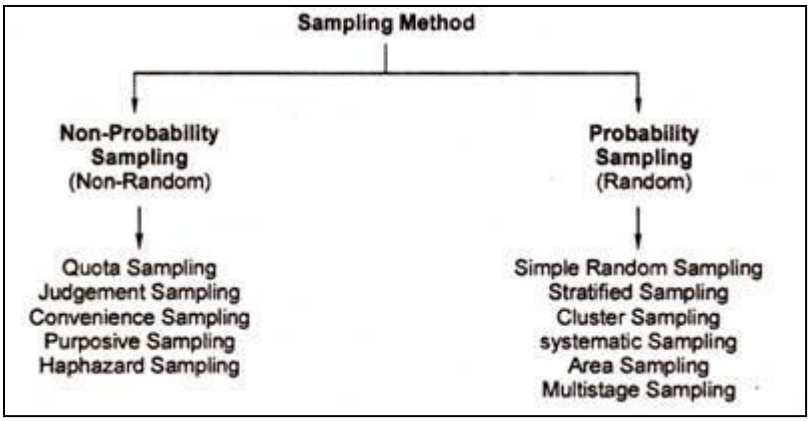
8.2.4 Exercise

8.2.1 SAMPLING TYPES AND METHODS – RANDOM SAMPLING

The two main types of sampling are random and non-random. In **random sampling,** *every unit of the population has the same probability of being selected into the sample.* Random sampling implies that chance enters into the process of selection. For example, most Indians would like to believe that winners of nationwide magazine sweepstakes or numbers selected as state lottery winners are selected by some random draw of numbers.

Sometimes random sampling is called *probability sampling* and non-random sampling is called *non probability sampling.* Because every unit of the population is not equally likely to be selected, assigning a probability of occurrence in non-random sampling is impossible.

The statistical methods presented and discussed are based on the assumption that the data come from random samples.



The four basic random sampling techniques are simple random sampling, stratified random sampling, systematic random sampling, and cluster (or area) random sampling. Each technique offers advantages and disadvantages. Some techniques are simpler to use, some are less costly, and others show greater potential for reducing sampling error.

8.2.1.a Simple Random Sampling

The most elementary random sampling technique is **simple random sampling**. Simple random sampling can be viewed as the basis for the other random sampling techniques.

With simple random sampling, each unit of the frame is numbered from 1 to *N* (where *N* is the size of the population). Next, a table of random numbers or a random number generator is used to select *n* items into the sample. A random number generator is usually a computer program that allows computer-calculated output to yield random numbers.

A brief table of random numbers is shown in the following lines in this section. The spaces in the table are there only for ease of reading the values. For each number, any of the 10 digits(0–9) is equally likely, so getting the same digit twice or more in a row is possible.

A Brief Table of Random Numbers

91567 42595 27958 30134 04024 86385 29880 99730
46503 18584 18845 49618 02304 51038 20655 58727
34914 63974 88720 82765 34476 17032 87589 40836
57491 16703 23167 49323 45021 33132 12544 41035
30405 83946 23792 14422 15059 45799 22716 19792
09983 74353 68668 30429 70735 25499 16631 35006
85900 07119 97336 71048 08178 77233 13916 47564

As an example, from the population frame of companies listed in following table, we will use simple random sampling to select a sample of six companies. First, we number every member of the population. We select as many digits for each unit sampled as there are in the largest number in the population. For example, if a population has 2,000 members, we select four-digit numbers. Because the population in this table contains 30 members, only two digits need be selected for each number. The population is numbered from 01 to 30, as shown in the Numbered Population of 30 Companies table.

A Population Frame of 30 Companies

Alaska Airlines	DuPont	Lubrizol
Alcoa	ExxonMobil	Mattel
Ashland	General Dynamics	Merck
Bank of America	General Electric	Microsoft
BellSouth	General Mills	Occidental Petroleum
Chevron	Halliburton	JCPenney
Citigroup	IBM	Procter & Gamble
Clorox	Kellogg	Ryder
Delta Air Lines	Kmart	Sears
Disney	Lowe's	Time Warner

The population is numbered from 01 to 30, as shown in the Numbered Population of 30 Companies table.

Numbered Population of 30 Companies

01 Alaska Airlines	11 DuPont	21 Lubrizol
02 Alcoa	12 ExxonMobil	22 Mattel
03 Ashland	13 General Dynamics	23 Merck
04 Bank of America	14 General Electric	24 Microsoft
05 BellSouth	15 General Mills	25Occidental Petroleum
06 Chevron	16 Halliburton	26 JCPenney
07 Citigroup	17 IBM	27Procter & Gamble
08 Clorox	18 Kellogg	28 Ryder
09 Delta Air Lines	19 Kmart	29 Sears
10 Disney	20 Lowe's	30 Time Warner

The object is to sample six companies, so six different two-digit numbers must be selected from the table of random numbers. Because this population contains only 30 companies, all numbers greater than 30 (31–99) must be ignored. If, for example, the number 67 is selected, the process is continued until a value between 1 and 30 is obtained. If the same number occurs more than once, we proceed to another number. For ease of understanding, we start with the first pair of digits in the above table and proceed across the first row until n = 6different values between 01 and 30 are selected. If additional numbers are needed, we proceed across the second row, and so on. Often a researcher will start at some randomly selected location in the table and proceed in a predetermined direction to select numbers. In the first row of digits in the above table of random numbers, the first number is 91. This number is out of range so it is cast out. The next two digits are 56. Next is 74, followed by 25, which is the first usable number. From the table of numbered population of 30 companies, we see that 25 is the number associated with Occidental Petroleum, so Occidental Petroleum is the first company selected into the sample. The next number is 95, unusable, followed by 27, which is usable. Twenty-seven is the number for Procter & Gamble, so this company is selected. Continuing the process, we pass over the numbers 95 and 83. The next usable number is 01, which is the value for Alaska Airlines. Thirty-four is next, followed by 04 and 02, both of which are usable. These numbers are associated with Bank of America and Alcoa, respectively. Continuing along the first row, the next usable number is 29, which is associated with Sears. Because this selection is the sixth, the sample is complete. The following companies constitute the final sample:

Alaska Airlines
Alcoa
Bank of America
Occidental Petroleum
Procter & Gamble
Sears

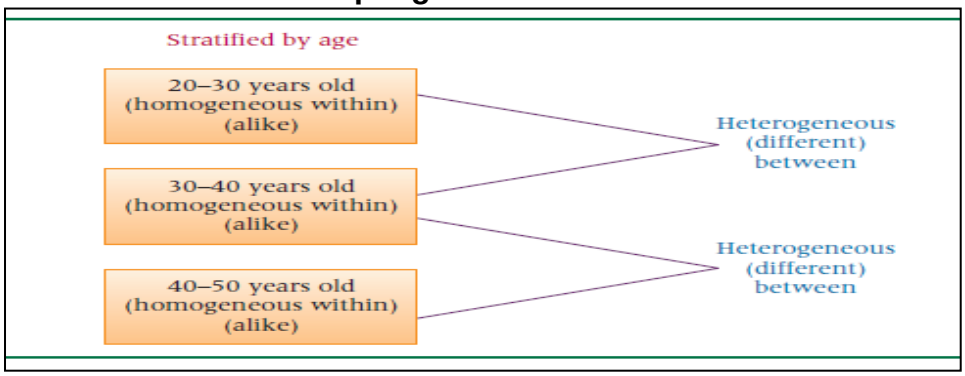
Simple random sampling is easier to perform on small than on large populations. The process of numbering all the members of the population and selecting items is cumbersome for large populations.

8.2.1.b Stratified Random Sampling

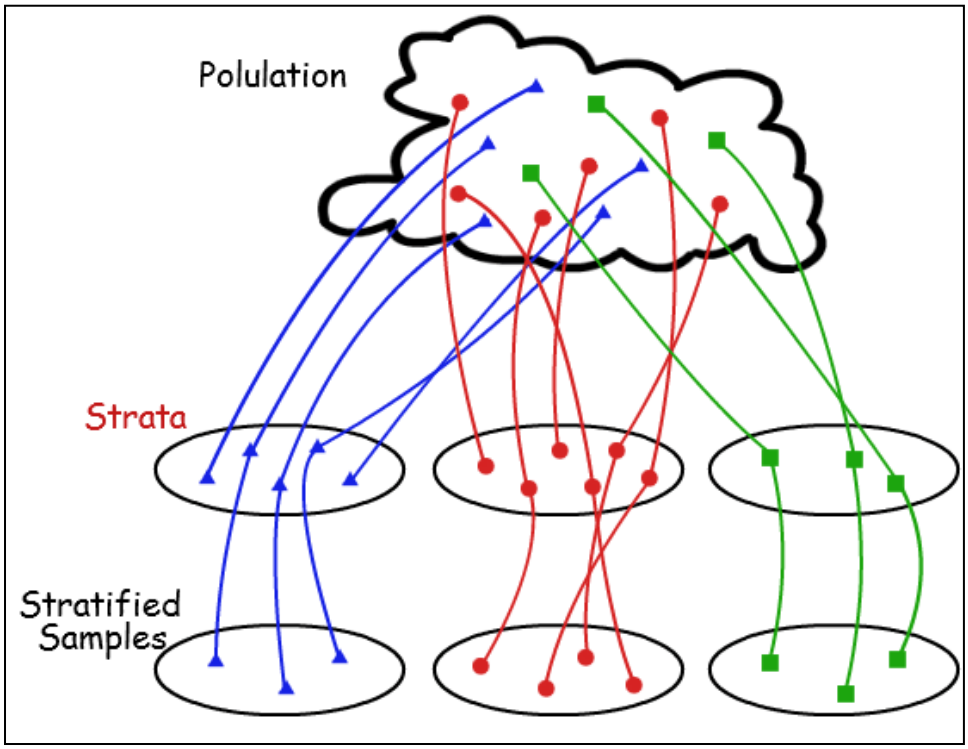
A second type of random sampling is **stratified random sampling**, in which the population is divided into non overlapping subpopulations called strata. The investigator then extracts a random sample from each of the subpopulations. The main reason for using stratified random sampling is that it has the potential for reducing sampling error. Sampling error occurs when, by chance, the sample does not represent the population. With stratified random sampling, the potential to match the sample closely to the population is greater than it is with simple random sampling because portions of the total sample are taken from different population subgroups. However, stratified random sampling is generally more costly than simple random sampling because each unit of the population must be assigned to a stratum before the random selection process begins.

Strata selection is usually based on available information. Such information may have been gleaned from previous censuses or surveys. Stratification benefits increase as the strata differ more. Internally, a stratum should be relatively homogeneous; externally, strata should contrast with each other. Stratification is often done by using demographic variables, such as sex, socioeconomic class, geographic region, religion, and ethnicity.

Stratified Random Sampling of FM Radio Listeners



In FM radio markets, age of listener is an important determinant of the type of programming used by a station. Figure contains stratification by age with three strata, based on the assumption that age makes a difference in preference of programming. This stratification implies that listeners 20 to 30 years of age tend to prefer the same type of programming, which is different from that preferred by listeners 30 to 40 and 40 to 50 years of age. Within each age subgroup (stratum), homogeneity or alikeness is present; between each pair of subgroups a difference, or heterogeneity, is present.



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

Proportionate & Disproportionate Stratified Random Sampling

Job level	Number of elements	Number of subjects in the sample: Proportionate sampling (20% of the elements)	Number of subjects in the sample: Disproportionate sampling
Top management	10	2	7
Middle-level management	30	6	15
Lower-level management	50	10	20
Supervisors	100	20	30
Clerks	500	100	60
Assistants	20	4	10
Total	710	142	142

8.2.1.c Systematic Sampling

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

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

DETERMINING THE VALUE OF *k*

$$k = \frac{N}{n}$$

where

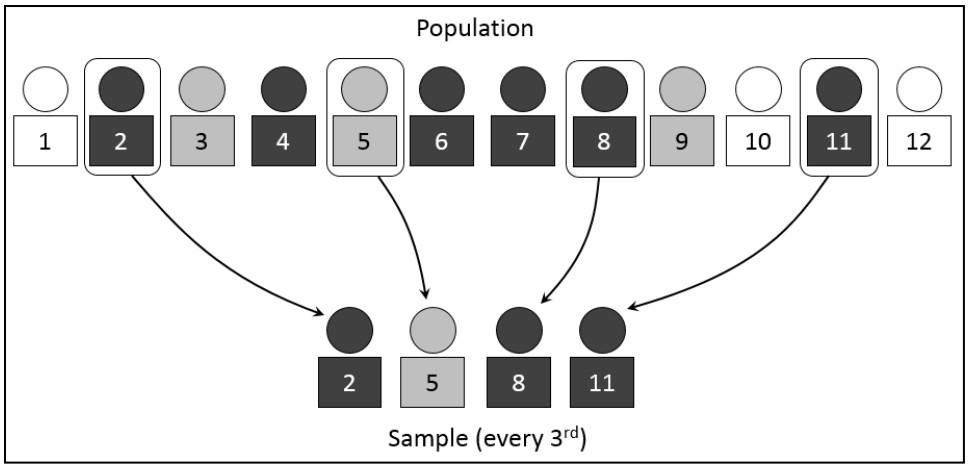
n = sample size
N = population size
k = size of interval for selection

As an example of systematic sampling, a management information systems researcher wanted to sample the manufacturers in Texas. He had enough financial support to sample 1,000 companies (*n*). The *Directory of Texas Manufacturers* listed approximately 17,000 total manufacturers in Texas (*N*) in alphabetical order. The value of *k* was 17 (17,000/1,000) and the researcher selected every 17th company in the directory for his sample.

Did the researcher begin with the first company listed or the 17th or one somewhere between? In selecting every *k* th value, a simple random number table should be used to select a value between 1 and *k* inclusive as a starting point. The second element for the sample is the starting point plus *k*. In the example, *k* = 17, so the researcher would have gone to a table of random numbers to determine a starting point between 1 and 17. Suppose he selected the number 5. He would have started with the 5th company, then selected the 22nd(5 + 17), and then the 39th, and so on.

Besides convenience, systematic sampling has other advantages. Because systematic sampling is evenly distributed across the frame, a knowledgeable person can easily determine whether a sampling plan has been followed in a study. However, a problem with systematic sampling

can occur if the data are subject to any periodicity, and the sampling interval is in syncopation with it. In such a case, the sampling would be non-random. For example, if a list of 150 college students is actually a merged list of five classes with 30 students in each class and if each of the lists of the five classes has been ordered with the names of top students first and bottom students last, then systematic sampling of every 30th student could cause selection of all top students, all bottom students, or all mediocre students; that is, the original list is subject to a cyclical or periodic organization. Systematic sampling methodology is based on the assumption that the source of population elements is random.



8.2.1.d Cluster (or Area) Sampling

Cluster (or area) sampling is a fourth type of random sampling. **Cluster (or area) sampling** involves dividing the population into non-overlapping areas, or clusters. However, in contrast to stratified random sampling where strata are homogeneous within, cluster sampling identifies clusters that tend to be internally heterogeneous. In theory, each cluster contains a wide variety of elements, and the cluster is a miniature, or microcosm, of the population.

Examples of clusters are towns, companies, homes, colleges, areas of a city, and geographic regions. Often clusters are naturally occurring groups of the population and are already identified, such as states or Standard Metropolitan Statistical Areas. Although area sampling usually refers to clusters that are areas of the population, such as geographic regions and cities, the terms *cluster sampling* and *area sampling* are used interchangeably in this chapter.

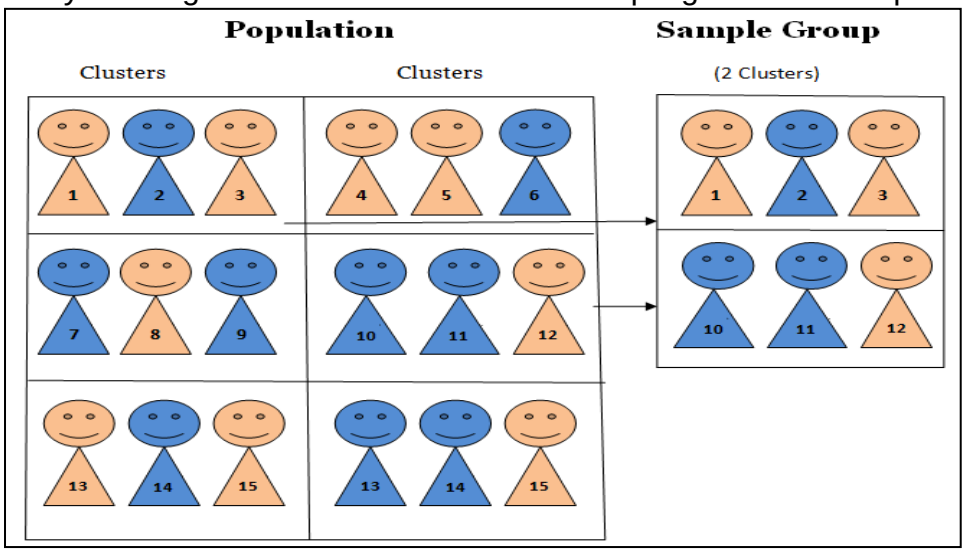
After randomly selecting clusters from the population, the business researcher either selects all elements of the chosen clusters or randomly selects individual elements into the sample from the clusters. One example of business research that makes use of clustering is test marketing of new products. Often in test marketing, the United States is divided into clusters of test market cities, and individual consumers within the test market cities are surveyed.

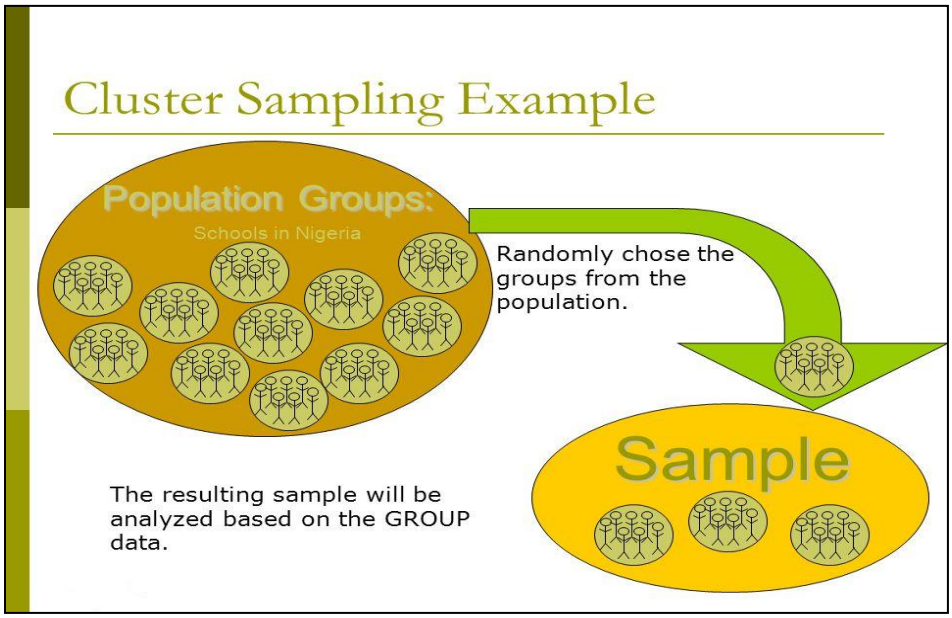
Sometimes the clusters are too large, and a second set of clusters is taken from each original cluster. This technique is called **two-stage sampling**. For example, a researcher could divide the United States into clusters of cities. She could then divide the cities into clusters of blocks and randomly select individual houses from the block clusters. The first stage is selecting the test cities and the second stage is selecting the blocks.

Cluster or area sampling offers several advantages. Two of the foremost advantages are convenience and cost. Clusters are usually convenient to obtain, and the cost of sampling from the entire population is reduced because the scope of the study is reduced to the clusters. The cost per element is usually lower in cluster or area sampling than in stratified sampling because of lower element listing or locating costs. The time and cost of contacting elements of the population can be

reduced, especially if travel is involved, because clustering reduces the distance to the sampled elements. In addition, administration of the sample survey can be simplified. Sometimes cluster or area sampling is the only feasible approach because the sampling frames of the individual elements of the population are unavailable and therefore other random sampling techniques cannot be used.

Cluster or area sampling also has several disadvantages. If the elements of a cluster are similar, cluster sampling may be statistically less efficient than simple random sampling. In an extreme case—when the elements of a cluster are the same—sampling from the cluster maybe no better than sampling a single unit from the cluster. Moreover, the costs and problems of statistical analysis are greater with cluster or area sampling than with simple random sampling.





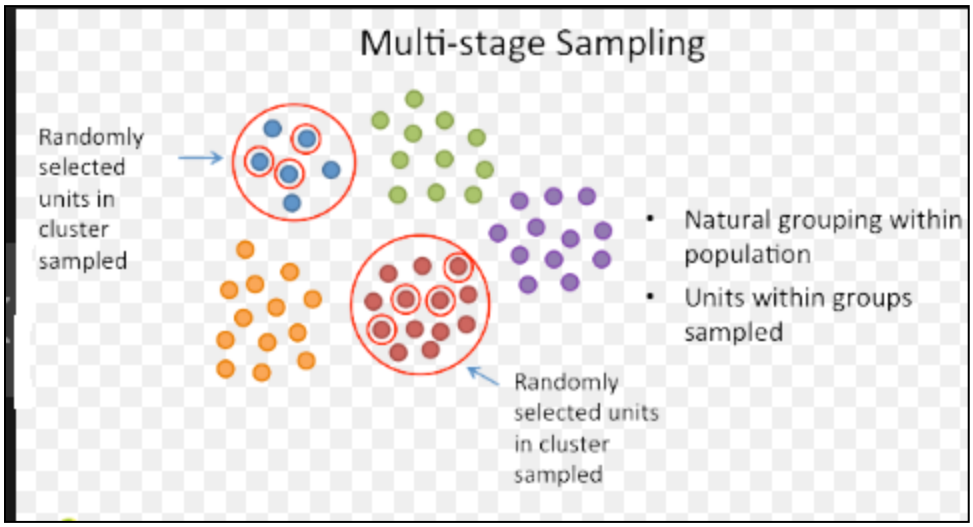
8.2.1.e Multi-Stage Sampling

The **Multistage Sampling** is the probability sampling technique wherein the sampling is carried out in several stages such that the sample size gets reduced at each stage.

The multistage sampling is a complex form of **cluster sampling**. The cluster sampling is yet another random sampling technique wherein the population is divided into subgroups called as clusters; then few clusters are chosen randomly for the survey.

While in the multistage sampling technique, the first level is similar to that of the cluster sampling, where the clusters are formed out of the population, but further, these clusters are sub-divided into smaller targeting groups, i.e. sub-clusters and then the subject from each sub-clusters are chosen randomly. Further, the stages can be added depending on the nature of research and the size of the population under study.

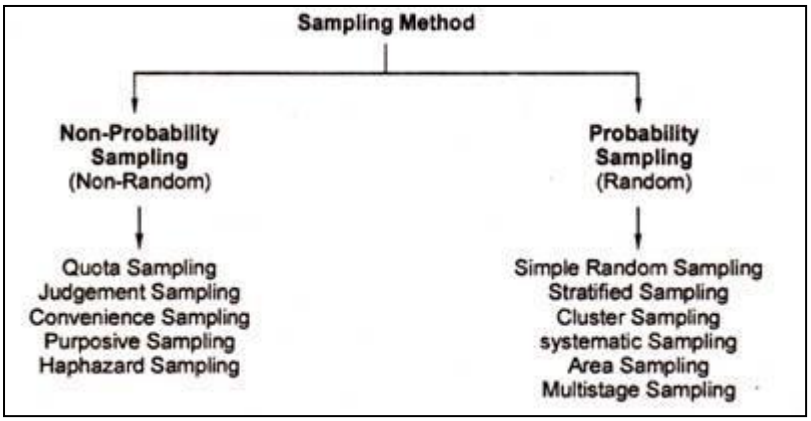
For example, if the government wants to take a sample of 10,000 households residing in Gujarat state. At the first stage, the state can be divided into the number districts, and then few districts can be selected randomly. At the second-stage, the chosen districts can be further sub-divided into the number of villages and then the sample of few villages can be taken at random. Now at the third-stage, the desired number of households can be selected from the villages chosen at the second stage. Thus, at each stage the size of the sample has become smaller and the research study has become more precise.



8.2.2 SAMPLING TYPES AND METHODS – NON-RANDOM SAMPLING

In **nonrandom sampling** *not every unit of the population has the same probability of being selected into the sample.* Members of nonrandom samples are not selected by chance. For example, they might be selected because they are at the right place at the right time or because they know the people conducting the research.

Non-random sampling methods are not appropriate techniques for gathering data to be analyzed by most of the statistical methods presented here.



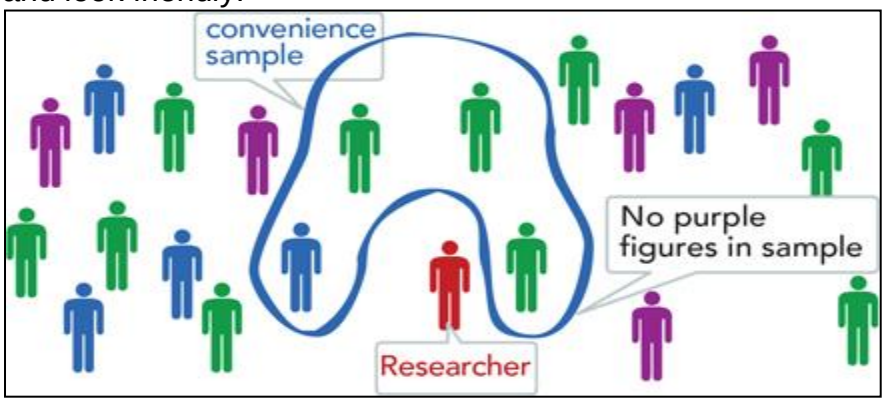
Sampling techniques used to select elements from the population by any mechanism that does not involve a random selection process are called nonrandom sampling techniques. Because chance is not used to select items from the samples, these techniques are non-probability techniques and are not desirable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

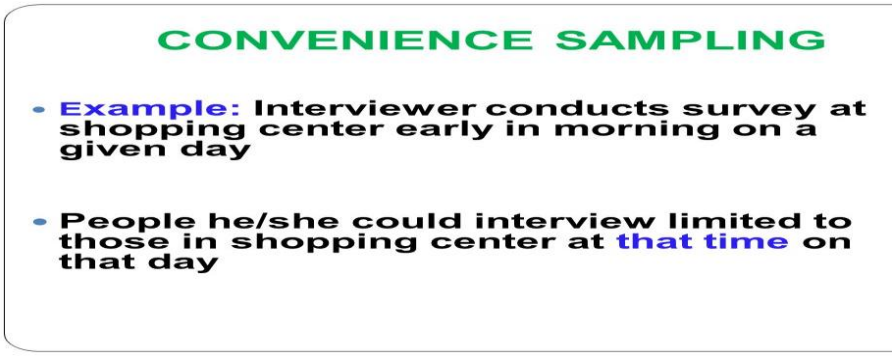
Four nonrandom sampling techniques are presented here:

- Convenience sampling
- Judgment sampling
- Quota sampling, and
- Snowball sampling
- Purposive sampling
- Haphazard sampling.

8.2.2.a Convenience sampling

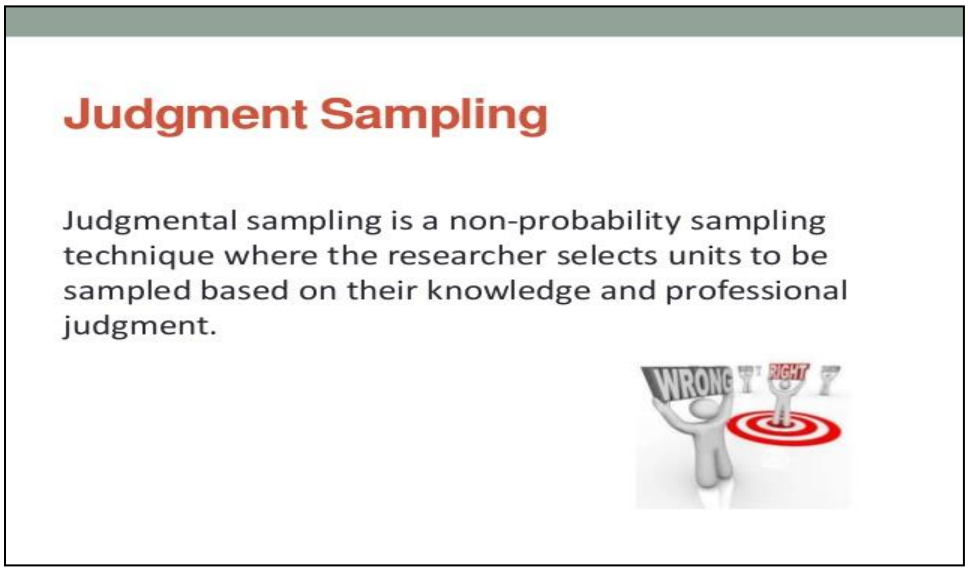
In **convenience sampling**, *elements for the sample are selected for the convenience of the searcher.* The researcher typically chooses elements that are readily available, nearby, or willing to participate. The sample tends to be less variable than the population because in many environments the extreme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

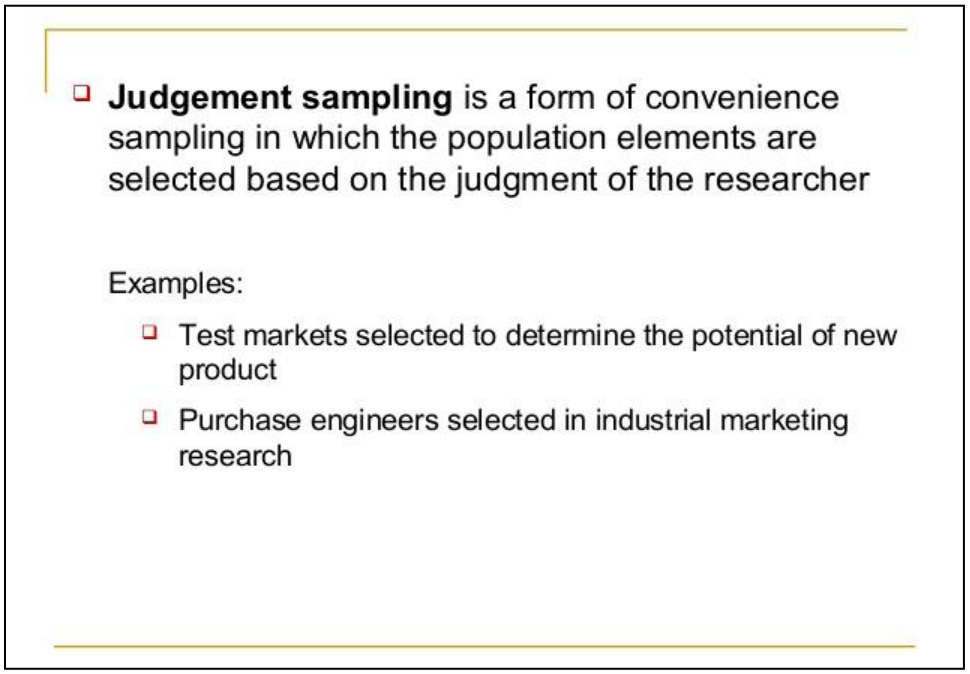




8.2.2.b Judgment sampling

Judgment sampling occurs when *elements selected for the sample are chosen by the judgment of the researcher.* Researchers often believe they can obtain a representative sample by using sound judgment, which will result in saving time and money. Sometimes ethical, professional researchers might believe they can select a more representative sample than the random process will provide. They might be right! However, some studies show that random sampling methods outperform judgment sampling in estimating the population mean even when the researcher who is administering the judgment sampling is trying to put together a representative sample. When sampling is done by judgment, calculating the probability that an element is going to be selected into the sample is not possible. The sampling error cannot be determined objectively because probabilities are based on *nonrandom* selection. Other problems are associated with judgment sampling. The researcher tends to make errors of judgment in one direction. These systematic errors lead to what are called *biases*. The researcher also is unlikely to include extreme elements. Judgment sampling provides no objective method for determining whether one person's judgment is better than another's.





8.2.2.c Quota Sampling

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

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

Quota sampling can be useful if no frame is available for the population. For example, suppose a researcher wants to stratify the population into owners of different types of cars but fails to find any lists of Toyota van owners. Through quota sampling, the researcher would proceed by interviewing all car owners and casting out non–Toyota van owners until the quota of Toyota van owners is filled.

Quota sampling is less expensive than most random sampling techniques because it essentially is a technique of convenience. However, cost may not be meaningful because the quality of nonrandom and random sampling techniques cannot be compared.

Another advantage of quota sampling is the speed of data gathering. The researcher does not have to call back or send out a second questionnaire if he does not receive a response; he just moves on to the next element. Also, preparatory work for quota sampling is minimal.

The main problem with quota sampling is that, when all is said and done, it still is only a *nonrandom* sampling technique. Some researchers believe that if the quota is filled by *randomly* selecting elements and discarding those not from a stratum, quota sampling is essentially a version

of stratified random sampling. However, most quota sampling is carried out by the researcher going where the quota can be filled quickly. The object is to gain the benefits of stratification without the high field costs of stratification. Ultimately, it remains a non probability sampling method.

Quota Sampling

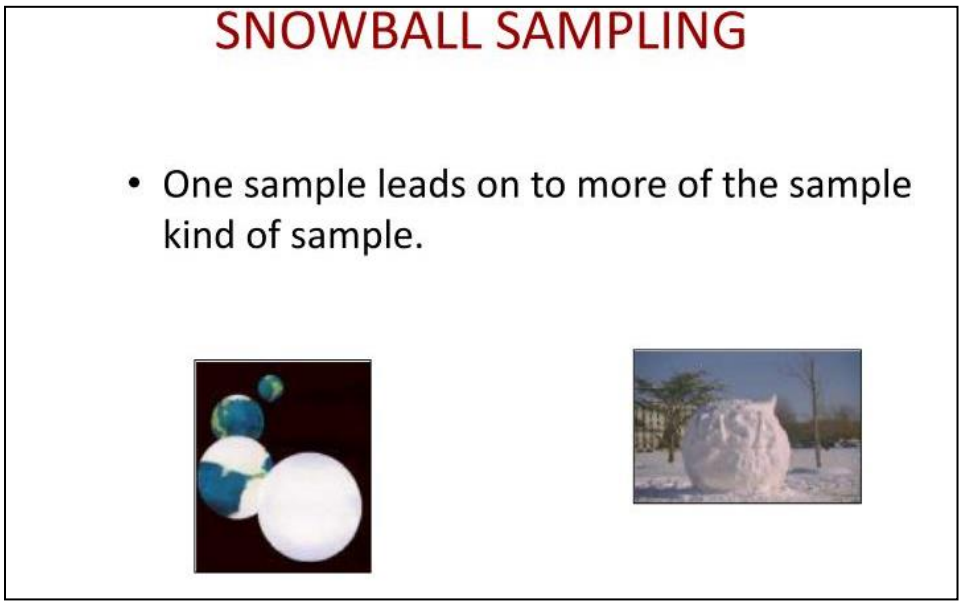
- It is nonprobability sampling technique wherein the researcher ensures equal or proportionate representation of subjects, depending on which trait is considered as the basis of the quota.
- The bases of the quota are usually age, gender, education, race, religion, & socio-economic status.
- For example, if the basis of the quota is college level & the research needs equal representation, with a sample size of 100, he must select 25 first-year students, another 25 second-year students, 25 third-year, & 25 fourth-year students.

What is Quota ?

8.2.2.d Snowball Sampling

Another nonrandom sampling technique is **snowball sampling**, in which *survey subjects are selected based on referral from other survey respondents.* The researcher identifies a person who fits the profile of subjects wanted for the study. The researcher then asks this person for the names and

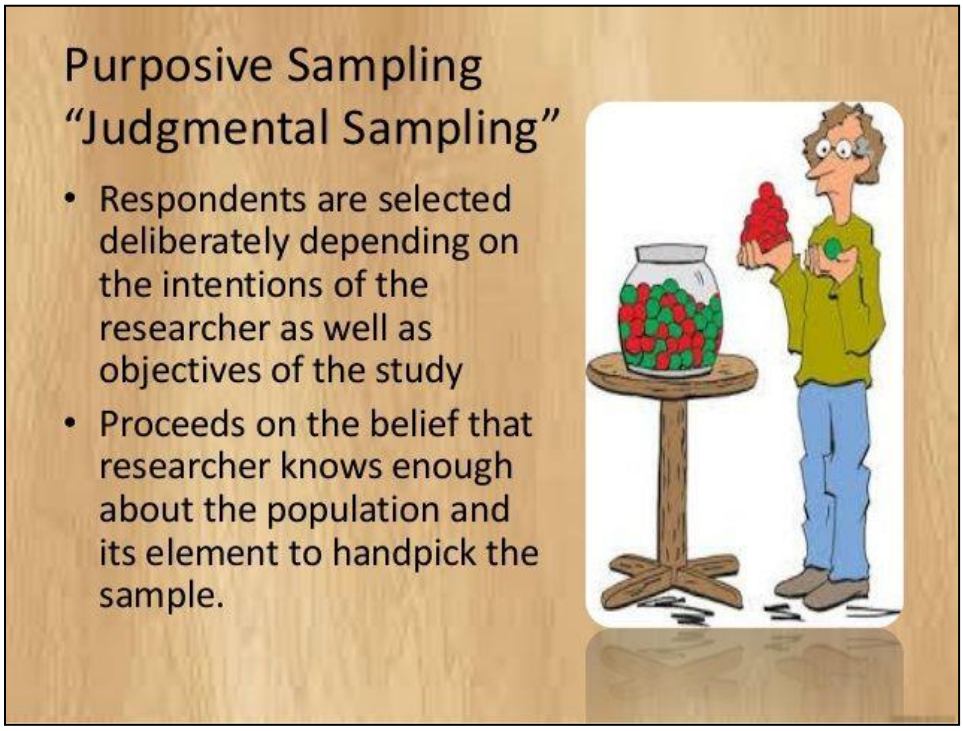
locations of others who would also fit the profile of subjects wanted for the study. Through these referrals, survey subjects can be identified cheaply and efficiently, which is particularly useful when survey subjects are difficult to locate. It is the main advantage of snowball sampling; its main disadvantage is that it is nonrandom.



8.2.2.e Purposive Sampling

A purposive sample is a non-probability sample that is selected based on characteristics of a population and the objective of the study. Purposive sampling is also known as judgmental, selective, or subjective sampling.

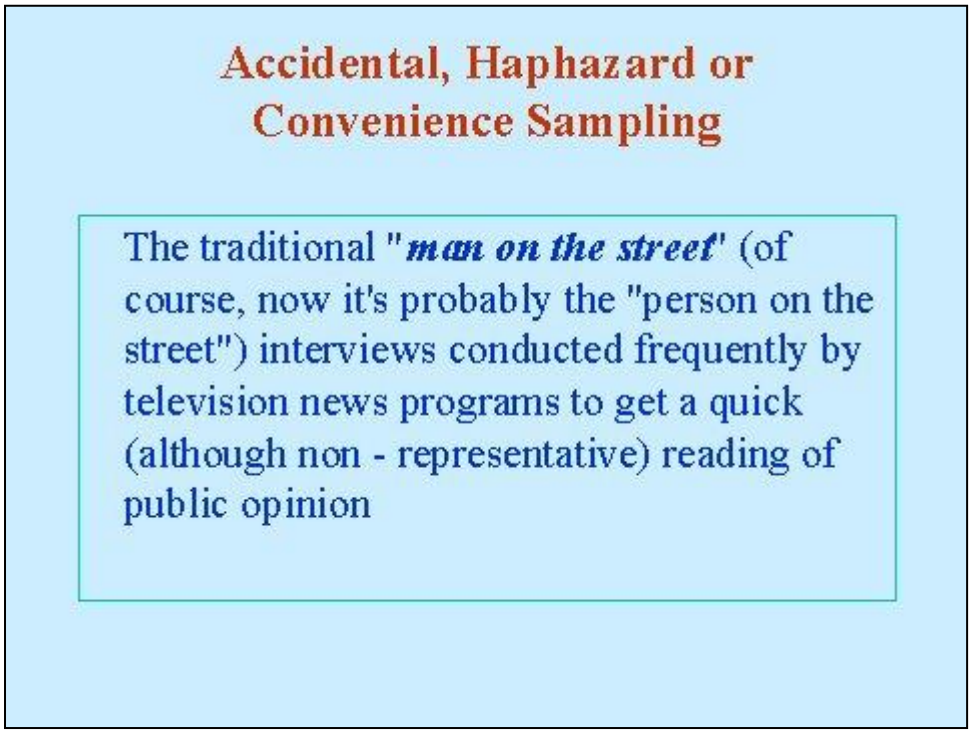
This type of sampling can be very useful in situations when you need to reach a targeted sample quickly, and where sampling for proportionality is not the main concern. There are seven types of purposive samples, each appropriate to a different research objective.



8.2.2.f Haphazard Sampling

It is a non-probability sample selection method in which the interviewer arbitrarily selects respondents for the survey without using systematic or random selection methods. There is no way to ensure that the estimates derived from a haphazard sample will be unbiased.

An auditor may choose from several methodologies for determining what to look at when auditing a company. One of the frequently employed techniques is called haphazard sampling used by auditors to simulate a variety of random sampling techniques when testing for potential errors in various accounting populations such as inventory and accounts receivable.



8.2.3.a Sampling Distribution

Sampling Error

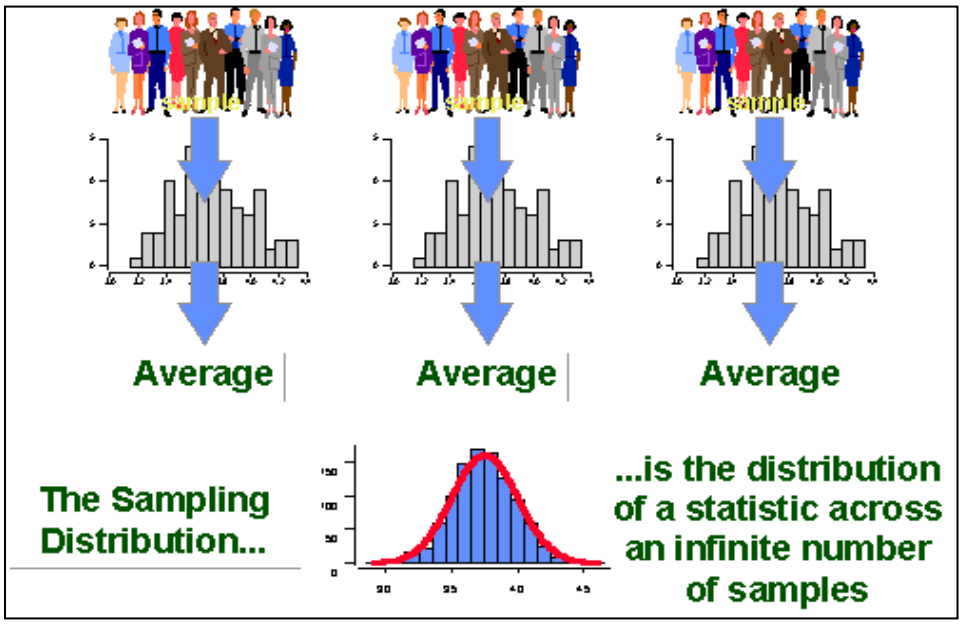
Sampling error occurs *when the sample is not representative of the population.* When random sampling techniques are used to select elements for the sample, sampling error occurs by chance. Many times the statistic computed on the sample is not an accurate estimate of the population parameter because the sample was not representative of the population. This result is caused by sampling error. With random samples, sampling error can be computed and analyzed.

Non-sampling Errors

All errors other than sampling errors are **non-sampling errors**. The many possible non-sampling errors include missing data, recording errors, input processing errors, and analysis errors. Other non-sampling errors result from the measurement instrument, such as errors of unclear definitions, defective questionnaires, and poorly conceived concepts. Improper definition of the frame is a non-sampling error. In many cases, finding a frame that perfectly fits the population is impossible. Insofar as it does not fit, a non-sampling error has been committed.

Sampling Distribution of *sample- mean*(*μ*)

In the inferential statistics process, a researcher selects a random sample from population, computes a statistic on the sample, and reaches conclusions about the population parameter from the statistic. In attempting to analyze the sample statistic, it is essential to know the distribution of the statistic.

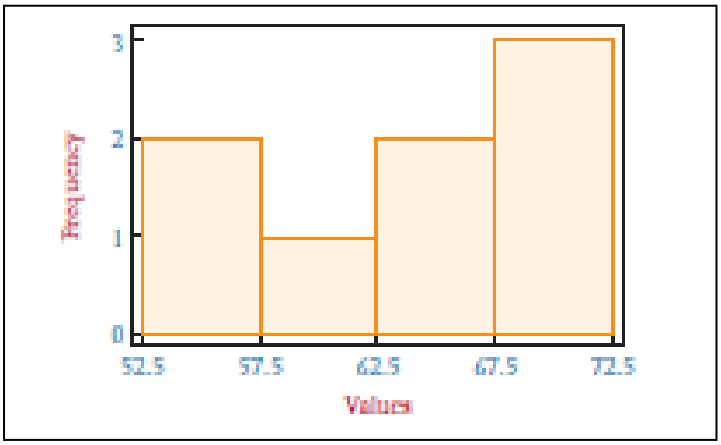


The sample-mean is one of the more common statistics used in the inferential process. To compute and assign the probability of occurrence of a particular value of a sample mean, the researcher must know the distribution of the sample means. One way to examine the distribution possibilities is to take a population with a particular distribution, randomly select samples of a given size, compute the sample means, and attempt to determine how the means are distributed.

Suppose a small finite population consists of only N = 8 numbers:

54 55 59 63 64 68 69 70

Using an Excel-produced histogram, we can see the shape of the distribution of this population of data.



Suppose we take all possible samples of size $n = 2$ from this population with replacement

The result is the following pairs of data:

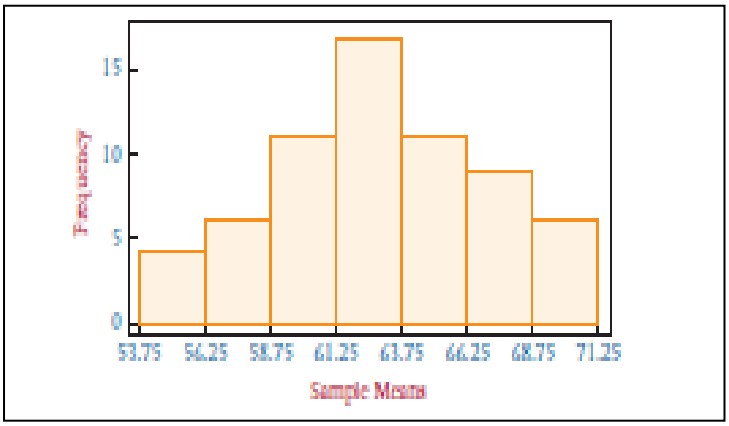
(54,54)	(55,54)	(59,54)	(63,54)
(54,55)	(55,55)	(59,55)	(63,55)
(54,59)	(55,59)	(59,59)	(63,59)
(54,63)	(55,63)	(59,63)	(63,63)
(54,64)	(55,64)	(59,64)	(63,64)
(54,68)	(55,68)	(59,68)	(63,68)
(54,69)	(55,69)	(59,69)	(63,69)
(54,70)	(55,70)	(59,70)	(63,70)
(64,54)	(68,54)	(69,54)	(70,54)
(64,55)	(68,55)	(69,55)	(70,55)
(64,59)	(68,59)	(69,59)	(70,59)
(64,63)	(68,63)	(69,63)	(70,63)
(64,64)	(68,64)	(69,64)	(70,64)
(64,68)	(68,68)	(69,68)	(70,68)
(64,69)	(68,69)	(69,69)	(70,69)
(64,70)	(68,70)	(69,70)	(70,70).

The means of each of these samples are:

54	54.5	56.5	58.5	59	61	61.5	62
54.5	55	57	59	59.5	61.5	62	62.5
56.5	57	59	61	61.5	63.5	64	64.5
58.5	59	61	63	63.5	65.5	66	66.5
59	59.5	61.5	63.5	64	66	66.5	67
60	61.5	63.5	65.5	66	68	68.5	69
61.5	62	64	66	66.5	68.5	69	69.5
62	62.5	64.5	66.5	67	69	69.5	70

Again using an Excel-produced histogram, we can see the shape of the distribution of these sample means.

8.2.3.b Central Limit Theorem



Notice that the shape of the histogram for sample means is quite unlike the shape of the histogram for the population. The sample means appear to “pile up” toward the middle of the distribution and “tail off” toward the extremes.

The sample means form a distribution that approaches a symmetrical, nearly normal-curve-type distribution.

Observe the shape of the distributions. Notice that even for small sample sizes, the distributions of sample means for samples taken from the uniformly distributed population begin to “pile up” in the middle. As sample sizes become much larger, the samples mean distributions begin to approach a normal distribution and the variation among the means decreases.

However, the sample means for samples taken from these populations appear to be approximately normally distributed, especially as the sample sizes become larger. What would happen to the distribution of sample means if we studied populations with differently shaped distributions?

The answer to that question is given in the **central limit theorem**.

Central Limit Theorem

If samples of size n are drawn randomly from a population that has a mean of μ and a standard deviation of σ, the sample means, $\bar{x}$, are approximately normally distributed for sufficiently large sample sizes ($n \geq 30$) regardless of the shape of the population distribution. If the population is normally distributed, the sample means are normally distributed for any size sample.

From mathematical expectation,* it can be shown that the mean of the sample means is the population mean.

$$\mu_{\bar{x}} = \mu$$

and the standard deviation of the sample means (called the standard error of the mean) is the standard deviation of the population divided by the square root of the sample size.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The central limit theorem creates the potential for applying the normal distribution to many problems when sample size is sufficiently large. Sample means that have been computed for random samples drawn from normally distributed populations are normally distributed.

However, the real advantage of the central limit theorem comes when sample data drawn from populations not normally distributed or from populations of unknown shape also can be analyzed by

using the normal distribution because the sample means are normally distributed for sufficiently large sample sizes.

The central limit theorem states that sample means are normally distributed regardless of the shape of the population for large samples and for any sample size with normally distributed populations. Thus, sample means can be analyzed by using z scores.

The formula to determine z scores for individual values from a normal distribution:

$$z = \frac{x - \mu}{\sigma}$$

If sample means are normally distributed, the z score formula applied to sample means would be

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

This result follows the general pattern of *z* scores: the difference between the statistic and its mean divided by the statistic's standard deviation. In this formula, the mean of the statistic of interest is $\mu_{\bar{x}}$, and *the standard deviation of the statistic of interest* is $\sigma_{\bar{x}}$, sometimes referred to as **the standard error of the mean**. To determine $\mu_{\bar{x}}$, the researcher would randomly draw out all possible samples of the given size from the population, compute the sample means, and average them. This task is virtually impossible to accomplish in any realistic period of time. Fortunately, $\mu_{\bar{x}}$ equals the population mean, μ, which is easier to access. Likewise, to determine directly the value of $\sigma_{\bar{x}}$, the researcher would take all possible samples of a given size from a population, compute the sample means, and determine the standard deviation of sample means. This task also is practically impossible. Fortunately, $\sigma_{\bar{x}}$ can be computed by using the population standard deviation divided by the square root of the sample size.

As sample size increases, the standard deviation of the sample means becomes smaller and smaller because the population standard deviation is being divided by larger and larger values of the square root of *n*. The ultimate benefit of the central limit theorem is a practical, useful version of the *z* formula for sample means.

z formula for sample mean

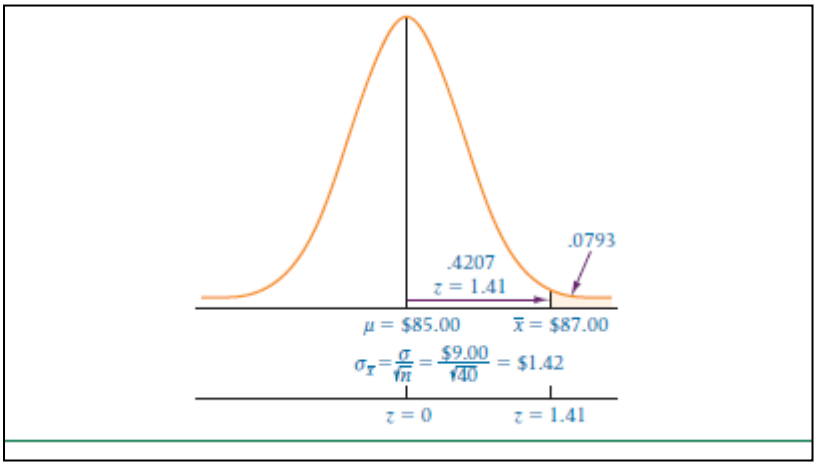
$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Suppose, for example, that the mean expenditure per customer at a tire store is 85.00, with a standard deviation of 9.00. If a random sample of 40 customers is taken, what is the probability that the sample average expenditure per customer for this sample will be 87.00 or more? Because the sample size is greater than 30, the central limit theorem can be used, and the sample means are normally distributed. With mean = 85.00, standard deviation = 9.00, and the z formula for sample means, z is computed as

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{87.00 - 85.00}{\frac{9.00}{\sqrt{40}}} = \frac{2.00}{1.42} = 1.41$$

From the *z* distribution table *z* = 1.41 produces a probability of 0.4207. This number is the probability of getting a sample mean between 87.00 and 85.00 (the population mean). Solving for the tail of the distribution yields
0.5000 - 0.4207 = 0.0793

which is the probability of $\bar{x} \geq$ 87.00. That is, 7.93% of the time, a random sample of 40 customers from this population will yield a sample mean expenditure of 87.00 or more.



8.2.4 EXERCISES

Q.1 Develop a frame for the population of each of the following research projects:

a. Measuring the job satisfaction of all union employees in a company

b. Conducting a telephone survey in a city, to determine the level of interest in opening a new hunting and fishing specialty store in the mall

c. Interviewing passengers of a major airline about its food service

d. Studying the quality control programs of boat manufacturers

e. Attempting to measure the corporate culture of cable television companies.

Q.2 Make a list of 20 people you know. Include men and women, various ages, various educational levels, and so on. Number the list and then use the random number table to select six people randomly from your list. How representative of the population is the sample? Find the proportion of men in your population and in your sample. How do the proportions compare? Find the proportion of20-year-olds in your sample and the proportion in the population. How do they compare?

Q.3 For each of the following research projects, list three variables for stratification of the sample:

a. A nationwide study of motels and hotels is being conducted. An attempt will be made to determine the extent of the availability of online links for customers. A sample of motels and hotels will be taken.

b. A consumer panel is to be formed by sampling people in a town. Members of the panel will be interviewed periodically in an effort to understand current consumer attitudes and behaviors.

c. A large soft drink company wants to study the characteristics of the bottlers of its products, but the company does not want to conduct a census.

d. The business research bureau of a large university is conducting a project in which the bureau will sample paper-manufacturing companies.

Q.4 In each of the following cases, the variable represents one way that a sample can be stratified in a study. For each variable, list some strata into which the variable can be divided.

a. Age of respondent (person)
b. Size of company (sales volume)
c. Size of retail outlet (square feet)
d. Geographic location
e. Occupation of respondent (person)
f. Type of business (company).

Q.5 A city's telephone book lists 100,000 people. If the telephone book is the frame for a study, how large would the sample size be if systematic sampling were done on every200th person?

Q.6 Give an example of how judgment sampling could be used in a study to determine how district attorneys feel about attorneys advertising on television.

Q.7 Give an example of how convenience sampling could be used in a study of Fortune 500 executives to measure corporate attitude toward paternity leave for employees.

Q.8 Give an example of how quota sampling could be used to conduct sampling by a company test marketing a new personal computer.

Q.9 Give an example of snowball sampling.

Q.10 A population has a mean of 50 and a standard deviation of 10. If a random sample of 64 is taken, what is the probability that the sample mean is each of the following?
a. Greater than 52
b. Less than 51
c. Less than 47
d. Between 48.5 and 52.4
e. Between 50.6 and 51.3

Q.10 -- Answer:
a. 0.0548
b. 0.7881

c. 0.0082
d. 0.8575
e. 0.1664

Q.11 Suppose a random sample of size 36 is drawn from a population with a mean of 278. If 86% of the time the sample mean is less than 280, what is the population standard deviation?

Q.11 -- Answer:

11.11

Q.12 Suppose a subdivision on the southwest side of Denver, Colorado, contains 1,500houses. The subdivision was built in 1983. A sample of 100 houses is selected randomly and evaluated by an appraiser. If the mean appraised value of a house in this subdivision for all houses is 177,000, with a standard deviation of 8,500, what is the probability that the sample average is greater than 185,000?

Q.12 -- Answer:

0.0000

Q.13 According to Nielsen Media Research, the average number of hours of TV viewing per household per week in the United States is 50.4 hours. Suppose the standard deviation is 11.8 hours and a random sample of 42 U.S. households is taken.
a. What is the probability that the sample average is more than 52 hours?
b. What is the probability that the sample average is less than 47.5 hours?
c. What is the probability that the sample average is less than 40 hours? If the sample average actually is less than 40 hours, what would it mean in terms of the Nielsen Media Research figures?
d. Suppose the population standard deviation is unknown. If 71% of all sample means are greater than 49 hours and the population mean is still 50.4 hours, what is the value of the population standard deviation?

Q.13 -- Answers:

a. 0.1894
b. 0.0559
c. 0.0000
d. 16.4964

❖❖❖❖

9

TESTING OF HYPOTHESIS - ONE SAMPLE

Unit Structure

9.1 INTRODUCTION

Hypothesis testing begins with an assumption, called a Hypothesis, that we make about a population parameter. A hypothesis is a supposition made as a basis for reasoning. Accordingto Prof. Morris Hamburg, "A Hypothesis in statistics is simply a quantitative statement abouta population." Palmer O. Johnson has beautifully described hypothesis as "islands in the uncharted seas of thought to be used as bases for consolidation and recuperation as we advance into the unknown."

In order to test a hypothesis, we collect sample data, produce sample statistics, and use this information to decide how likely it is that our hypothesized population parameter is correct. Say that we assume a certain value for a population mean. To test the validity of our assumption, we gather sample data and determine the difference between the hypothesized value and the actual value of the sample mean. Then we judge whether the difference is significant. The smaller the difference, the greater the likelihood that our hypothesized value for the mean is correct. The larger the difference, the smaller the likelihood.

Unfortunately, the difference between the hypothesized population parameter and the actual sample statistic is more often neither so large that we automatically reject our hypothesis nor so small that we just as quickly accept it. So in hypothesis testing as in most significant rela- life decisions, clear-cut solutions are the exception, not the rule.

There can be several types of hypotheses. For example, a coin may be tossed 200 times and we may get heads 80 times and tails 120 times. We may now be interested in testing the hypothesis that the coin is unbiased. To take another example we may study the average weight of the 100 students of a particular college and may get the result as 110 lb. We may now be interested in testing the hypothesis that the sample has been drawn from a population with average weight 115 lb. Similarly, we may be interested in testing the hypothesis that the variables in the population are uncorrelated.

Suppose a manager of a large shopping mall tells us that the average work efficiency of her employees is 90%. How can we test the validity of her hypothesis? using the sampling methods we learnt earlier, we could calculate the efficiency of a sample of her employees. If we did this and the sample statistic came out to be 93%, we would readily accept the manager's statement. However, if the sample statistic were 46 percent, we would reject her assumption as untrue. We can interpret both these outcomes, 93 percent and 46 percent, usingour common sense.

Now suppose that our sample statistic reveals an efficiency of 81 percent. This value is relatively close to 90%. But is it close enough for us to accept the manager's hypothesis? Whether we accept or reject the manager's hypothesis, we cannot be absolutely certain that our decision is correct; therefore, we will have to learn to deal with uncertainty in our decision making. We cannot accept or reject a hypothesis about a population parameter simply by intuition. Instead, we need to learn how to decide objectively, on the basis of sample information, whether to accept or reject a hunch.

9.2 HYPOTHESIS TESTING

Use a statistic calculated from the sample to test an assertion about the value of a population parameter.

STEP 1: Determine the sample statistic to be calculated and formulate the hypothesis.

1. The decision about which sample statistic to calculate depends upon the scale used to measure the variable.
    - a proportion (π) is calculated for nominal scaled variables.
    - a median (med) is calculated for ordinal scaled variables.
    - a mean (μ) is calculated for interval or ratio scaled variables.
2. The hypotheses are:

Null Hypothesis (H_0): H_0 specifies a value for the population parameter against which the sample statistic is tested. H_0 always includes an equality.

Alternative Hypothesis (H_a): H_a specifies a competing value for the population parameter. H_a

- is formulated to reflect the proposition the researcher wants to verify.
- includes a non-equality that is mutually exclusive of H_0.
- is set up for either a one tailed test or a two tailed test.

The decision about using a one tailed vs. two tailed test depends upon the proposition the researcher wants to verify. For example, if the mean age of the students in this class is tested against the value 21, the hypotheses could be:

ONE TAILED TEST	**TWO TAILED TEST**
H_0: $\mu = 21$ or H_0: $\mu = 21$ H_a: $\mu > 21$ or H_a: $\mu > 21$	H_0: $\mu = 21$ H_a: $\mu \neq 21$

STEP: 2 Conduct the test.

1. All hypothesis tests take action on H_0. H_0 is either rejected or not rejected. When H_0 is rejected (not rejected), the proposition in H_a is verified (not verified).
2. Conducting the test involves deciding if H_0 should be rejected or not to be rejected.
3. There is always a chance a mistake will be made when H_0 is rejected or not rejected. This is because the decision is based on information obtained from a sample rather than the entire target population, i.e., sampling error. Hypothesis tests are designed to control for Type I error: rejecting a true null hypothesis.
4. One approach to deciding if H_0 should be rejected or not rejected is the critical value approach. The researcher controls the chance of Type I error by setting the test's level of significance (α). Traditionally, α is set at either .01, .05, or .10.

With the critical value approach:

- Rejecting H_0 when the researcher sets $\alpha = .01$ means the researcher is willing to accept no more than a 1% chance that a true null hypothesis is being rejected. The results of a test at the 1% level of significance are highly significant.
- Rejecting H_0 when the researcher sets $\alpha = .05$ means the researcher is willing to accept no more than a 5% chance that a true null hypothesis is being rejected. The results of a test at the 5% level of significance are significant.

- Rejecting H_0 when the researcher sets $\alpha = .10$ means the researcher is willing to accept no more than a 10% chance that a true null hypothesis is being rejected. The results of a test at the 10% level of significance are marginally significant.

5. An alternative approach to deciding if H_0 should be rejected or not reject is the p-value approach. The researcher knows precisely the chance of Type I error because the statistical package calculates the exact probability that a true null hypothesis is being rejected. This exact probability is called the "p-value."

With the p-value approach:

- The researcher sets the test's α level based on how much risk of Type I error the researcher is willing to tolerate. The α level can be set at any value as long as it is less than or equal to 0.10.
- The researcher rejects H_0 if the p-value $< \alpha$.
- The Methods section of a research report that uses the p-value approach should include a statement about the level that has been set for α.
- Most Statistical packages calculate the p-value for a 2-tailed test. If you're conducting a 1-tailed test you must divide p-value by 2 before deciding if it is acceptable.
- In SPSS output, the p-value is labelled "Sig(2-tailed)".

An Interesting Note

Because the p-value precisely measures the test's chances of Type I error, it measures the exact α. level the test obtains. Consequently:

- The p-value is also called the "obtained α. level".
- The smaller (larger) the obtained α. level, the more (less) statistically significant the results.

STEP 3: State the results of the test as they relate to the problem under study. When H_0 is rejected, there is sufficient "evidence" in the data to support the assertion made in H_a. When H_0 is not rejected, the data do not contain sufficient "evidence" to support the assertion made in H_a.

EXAMPLE RESEARCH PROBLEM

An ongoing concern of University of Wisconsin System administrators is one frequently expressed by students and their parents: earning a degree from a System University takes longer than the advertised four years. As aspiring UW-L Bus 230 team decides to look into the problem. Their research is guided by the hypothesis that the problem, at least in part, is due UW-L students' lack of commitment. The team reasons that for students to be committed to graduating "on time" they must average 15 credit hours a semester (the minimum number needed to graduate in four years), and study hard enough so they won't have to repeat classes.The team hypothesises that UW-L students are averaging fewer than 15 credit hours per semester, and are studying less than most faculty recommend: two hours per week for each credit hour attempted. The team interviews 200 randomly selected CBA undergraduates. Their questionnaire asks:

1. How many credits are you taking this semester?
2. In a typical week, how many hours do you study?

The results of the analysis of these data appear below. Do these data confirm the research team's hypothesis?

Step 1: Determine the sample statistic to calculate and formulate the hypotheses.

- The sample statistic is a mean (μ) because the variable is measured with a ratio scale.
- The test is set up as a one-tail test to evaluate the researchers' position that students are averaging fewer than 15 credits per semester.

Null Hypothesis H_0: μ credits = 15

Alternative Hypothesis H_a: μ credits < 15 ⇨ 1 tailed test ⇨ divide Sig (2-tailed) by 2.

Step 2: Conduct the test.

One-Sample Test

Test Value = 15						
	t	Df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Credits	-4.096	199	.000	- .8850	- 1.3111	- .4589

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Credits	200	14.1150	3.0559	.2161

SPSS OUTPUT: Analyse>Compare Means>One Sample t-test:

- p-value/2≤ .0005/2 = .000⇨the chance a true null hypothesis is being rejected is less than .025%.
- .005 < .05⇨ reject H_0 because the p-value is less than the α level.

Note: These results are highly significant because the test's obtained α level is almost zero.

Step 3: State the Results

The data contain sufficient evidence to conclude that UW-L students are averaging fewerthan 15 credit hours per semester.

Step 1: Determine the sample statistic to calculate and formulate the hypotheses.

- The sample statistic is a mean (μ) because the variable is measured with a ratio scale.
- The test is set up as a one-tail test to evaluate the researchers' position that students are averaging fewer than 28 hours of studying per week.

Null Hypothesis H_0: μ study = 28

Alternative Hypothesis H_a: μ study < 28 ⇨1 tailed test ⇨divide Sig (2-tailed) by 2.

Step 2: Conduct the test

Set α = .05

One Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
STUDY	200	20.7000	11.8619	.8388

SPSS OUTPUT: Analyse>Compare Means>One Sample t-test:

One-Sample Test

Test Value = 15					
	t	Df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference

					Lower	Upper
Credits	-8.703	199	.000	- 7.3000	- 8.9540	- 5.6460

- p-value/2≤ .0005/2 = .000⇨the chance a true null hypothesis is being rejected is less than .025%.
- .005 < .05⇨ reject H_0 because the p-value is less than the α level.

Note: These results are highly significant because the test's obtained α level is almost zero.

Step 3: State the results.

The data contain sufficient evidence to conclude that on average UW-L students study fewer than 28 hours per week.

PRESENTING STATISTICAL RESULTS

The sample estimate for the average number of credit hours UW-L students take per semesteris 14.2 (Figure 1). This value is statistically less than 15 (p-value/2 ≤ .00025, Appendix 2, p.1), the minimum number of credit hours needed per semester to graduate in four years. Students study an average of 20.7 hours per week (Figure 1). This value is statistically less than 28 (p-value/2≤ .00025, Appendix 2, p.2), the number of study hours per week faculty would recommend for a 14 hour credit load.

DISCUSSING RESEARCH RESULTS

The results indicate that UW-L student behaviour contributes to terms to graduation that exceed four years. Students average only 14.2 credit hours per semester. This value is statistically less than 15(p-value/2≤ .00025), the minimum number of credit hours per semester needed to graduate on time. Also, students study less than the amount most faculty recommend. Given a 14 credit hour load, faculty recommend that students study 28 hours per week. The 20.7 hours UW-L students study is statistically less than 28 (p-value/2≤ .00025). While UW-L, students may be brighter than most thereby needing to study less, it is more likely that the lack of study effort leads to poor classroom performance and a need to retake some classes. This would extend the number of semester needed to graduate.

EXAMPLE RESEARCH PROBLEM

One objective of the authors of "Alcohol Consumption and College Life" was to evaluate the UW-L Spring Core Alcohol and Drug Survey finding that "Most UW-L students have 0-5 drink a week." To do so their questionnaire asked:

During a typical week, how many days per week do you consume alcoholic beverages?On average, how many drinks do you consume each time you drink?

To do the analysis, the authors multiplied the responses to Q2 and Q3, and used SPSS to generate a frequency table of the product, which they labelled Weekly Consumption:

SPSS Output: Analyse > Descriptive Statistics > Frequencies:

Weekly Consumption

Valid	Frequency	Percent	Valid Percent	Cumulative Percent
0	24	16.2	16.2	16.2
1	2	1.4	1.4	17.6
2	7	4.7	4.7	22.3

3	11	7.4	7.4	29.7
4	10	6.8	6.8	36.5
5	7	4.7	4.7	41.2
6	7	4.7	4.7	45.9
7	1	0.7	0.7	46.6
8	10	6.8	6.8	60.8
9	1	0.7	0.7	54.1
10	10	6.8	68	60.8
12	8	5.4	5.4	66.2
14	4	2.7	2.7	68.9
15	4	2.7	2.7	71.6
16	7	4.7	4.7	76.4
18	6	4.1	4.1	80.4
20	3	2.0	2.0	82.4
21	1	.07	0.7	83.1
24	3	.20	2.0	85.1
27	2	1.4	1.4	86.5
30	6	4.1	4.1	90.5
33	1	0.7	0.7	91.2
36	2	1.4	1.4	92.6
39	2	1.4	1.4	93.9
40	3	2.0	2.0	95.9
45	1	0.7	0.7	96.6
54	1	0.7	0.7	97.3
60	1	0.7	0.7	98.0
72	1	0.7	0.7	98.6
75	1	0.7	0.7	99.3
120	1	0.7	0.7	100.0
Total	**148**	**100.0**	**100.0**	

Using the same approach as the Core Study, the authors concluded that most UW-L studentshave 0-8 drinks per week.

EXAMPLE RESEARCH PROBLEM CONTINUED

The authors of "Alcohol Consumption and College Life" wanted to test the hypothesis thatthe average number of drinks UW-L student's consume was greater than 8.6, the value that was found in the Core Study.

Step 1: Determine the sample statistic to calculate and formulate the hypotheses.

- The sample statistic is a mean (μ) because the variable is measured with a ratio scale.
- The test is set up as a one-tail test to evaluate the researchers' position that students drink more than 8.6 drinks per week.

Null Hypothesis H_0: μ Weekly Consumption = 8.6

Alternative Hypothesis H_a: μ Weekly Consumption > 8.6 1 tailed test⇨Divide Sig (2-⇨ tailed) by 2.

- **Step 2: Conduct the test.**
- **Set α = .05**

SPSS OUTPUT: Analyse >Compare Means > One Sample t-test:

One-Sample Test

Test Value = 8.6						
	T	Df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Weekly Consumption	3.179	147	.002	4.31	1.63	6.98

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Weekly Consumption	148	12.91	16.48	1.35

- p-value/2 = .002/2 = .001 ⇨ the chance a true null hypothesis is being rejected is less than -1%.
- .001 < .05 ⇨ reject H_0 because the p-value is less than the α level.

Note: These results are highly significant because the test's obtained α level is almost .001.

Step 3: State the results.

The data contain sufficient evidence to conclude that on average UW-L students are consuming on average more than 8.6 drinks per week.

PRESENTING STATISTICAL RESULTS

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	24	16.2	16.2	16.2
	1	2	1.4	1.4	17.6
	2	7	4.7	4.7	22.3
	3	11	7.4	7.4	29.7
	4	10	6.8	6.8	36.5
	5	7	4.7	4.7	41.2
	6	7	4.7	4.7	45.9
	7	1	.7	.7	46.6
	8	10	6.8	6.8	53.4
	9	1	.7	.7	54.1

Weekly ConsumptionFigure 6

Another hypothesis tested was that most UW-L students consume five or less drinks per week. According to the cumulative frequency observed, most (53.4%) UW-L students drink zero to eight alcoholic beverages per week (Figure 6). Furthermore, the sample estimate for the average number of drinks consumed per week is 12.91. A one sample t-test found this figure to be statistically larger than 8.6, the mean figure reported in the Core Study (p-value/2 = .001, Appendix B, page 30).

DISCUSSING RESEARCH RESULTS

there are some stark differences in the findings of this study and those of the Core Study. In contrast to the Core Study, which concluded that most UW-L students have 0-5 drinks a week, this study found that most students have 0-8 drinks a week. Using the same

methodology as the Core Study, the cumulative frequency for drinks per week exceeded 50% (53.4%) at eight drinks. Furthermore, statistical evidence exists to estimate of 12.91 is clearly statistically larger than the value reported in the Core Study. These differences may be the consequence of how the samples were chosen. This study's sample was randomly chosen from a list of all UW-L students. The Core's sample was a "modified stratified random sampling of classes" in which General Education, junior and senior level classes were randomly selected in an effort to reflect the proportion of freshmen, sophomore etc., in the population. All students in attendance at 24 of the 60 classes selected were surveyed. While this procedure may result in a sample that is a fair representation of the academic classes, the time of day the surveyed class met may have influenced the results. For example, 7:45 a.m. classes may be those that students skip most, especially if the preceding night involved drinking. A sampling procedure that might miss drinkers would bias the consumption numbers downward, and lead to the differences in the findings of the two studies.

EXAMPLE RESEARCH PROBLEM

Ted Skemp, a La Crosse area attorney, was startled when he read the April, 1996 edition of the ABA Journal. It reported that lawyers' "[clients complained most often about being ignored.....more than 20% believed their lawyers failed to return phone calls promptly, [and] more than 20% believed their lawyers did not pay adequate attention to their cases." To make sure he was treating his clients right, Mr. Skemp commissioned a Bus 230 team to survey his clients. The research team prepared a questionnaire that included the questions:

1. When you call Mr. Skemp and have to leave a message, does he return your callspromptly?

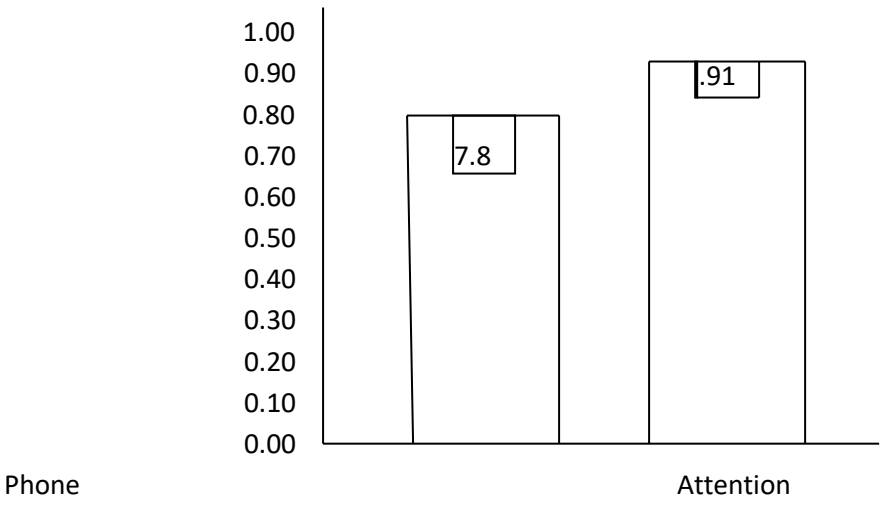
0. No 1. Yes

2. Does Mr. Skemp pay adequate attention to your case?

0. No 1. Yes

The team named the variable measured by Q1 "PHONE," and the variable measured by Q2"ATTENTION." Output from statistical analysis of these variables appears below. Present the statistical resultsand discuss them in terms of whether or not they are favourable for Mr. Skemp.

Statistician with Mr. Skemp



9.3 TEST OF HYPOTHESIS EXAMPLES ERROR TYPE I & TYPE II

Example 1

The Alpha-Fetoprotein (AFP) Test has both Type I and Type II error possibilities. This test screens the mother's blood during pregnancy for AFP and determine risk. Abnormally high or low levels may indicate Down Syndrome.

H_0: patient is healthy

H_a: patient is unhealthy

Error Type I (False Positive) is: Test wrongly indicates that patient has a Down Syndrome, which means that pregnancy must be aborted for no reason.

Error Type II (False Negative) is: Test is negative and the child will be born with multiple anomalies.

Example 2

The Head of the Cartel is trying to uncover the mole from within his crew.

H_0: The henchman was not an undercover Miami Dade Police Officer**H_a:**

The henchman was an undercover Miami Dade Police Officer **Error Type**

1: (False Positive)

The head of the Cartel ended up murdering the henchman that was not an undercover Miami Dade Police Officer. Although the henchman was innocent, he was killed preventing him from ever flipping and giving the government information.

Error Type 2: (False Negative)

The head of the Cartel interviews a henchman that wan an undercover Miami Dade Police Officer, but fails to unveil his true identity. Consequently, he continues to allow exposure of his operation to the undercover Miami Dade Police officer, and further reveals the ins and outs of his operation, that will eventually bring him to his demise.

Example 3

Airplane mechanic inspects plane for any irregularities or malfunction.

H_0: Plane seems to meet all standards of FAA and is ok-ed to fly.

H_a: Plane seems to NOT meet all standards of FAA and is AOG (airplane on the ground). **Error Type 1: (False Positive):** Airplane Reverse Thruster is visually fine and operable butwhile check testing light indicator states it is not, it is replaced even though thruster was fineand operable, thus avoiding any accident or problem.

Error Type 2: (False Negative): Airplane Reverse Thruster seems visually to be malfunctioning but check testing light indicator states it is Fine & Operable, it is NOT replaced. At landing a pilot reports a malfunction with the thruster and cannot reduce speed atlanding, plane is involved in accident and many innocent lives are lost.

Example 4

The mechanic inspects the brake pads for the minimum allowable thickness.H_0:

Vehicles breaks meet the standard for the minimum allowable thickness.

H_a: Vehicles brakes do not meet the standard for the minimum allowable thickness.

Error Type 1: (False Positive)

The brakes are fine, but the check indicates you need to replace the brake pads; therefore any possible problems with brakes are avoided even though the brakes were not worn.

Error Type 2: (False Negative)

The brake pads are worn to beyond the minimum allowable thickness, but the mechanic does not find anything wrong with them and does not replace them. Consequently, the driver of thevehicle gets into an accident because she was unable to break effectively and gets into a fatal accident.

Example 5

During a boxing match, two contenders bump heads. The referee checks the concussion on one of the boxers.

H_0: The boxer is find and able to continue boxing.

H_a: The boxer is injured and must call the bout.

Error Type 1

The boxer is fine and not seriously injured but the referee finds the concussion too severe andstops the fight.

Error Type 2

The boxer is seriously injured and the concussion is detrimental to his health, but the referee does not find the concussion severe, and allows the fight to continue. Due to the severity of the cut, the boxer faints in mid fight and goes into a coma.

PROCEDURE OF TESTING A HYPOTHESIS

Following are the steps required for testing a hypothesis:

1. Setting up of the hypothesis.
2. Test Statistic
3. Type I & Type II Error
4. Level of Significance
5. Critical Region and Rejection Region
6. Tailed Test Observation
7. Taking a Decision

1. Setting up of the hypothesis: A statistical hypothesis or simply a hypothesis is a tentative solution logically drawn concerning any parameter or the population.

Generally two hypothesis are set up. They are referred to as,

a) Null Hypothesis (H_0): A statistical hypothesis which is stated for the purpose of possible acceptance is called null hypothesis. It is usually referred to by the symbol (H_0). In the words of FISHER, **"Null Hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true."**

b) Alternative Hypothesis (H_a): Any hypothesis which is set up as a complementary to the null hypothesis is called as alternate hypothesis and is denoted by **(H_a).**

For example, Null Hypothesis and Alternative Hypothesis in the above examples would be as follows:

i) $H_0 : \mu = \mu_0$ and $H_a : \mu > \mu_0$ or $\mu < \mu_0$.

ii) H_0 : There is no difference between the two Drugs A and B.

Or H_a : Drug A is better than Drug B.

Or H_a : Drug A is inferior to Drug B.

Then from the above, it is clear that the null hypothesis indicates no preferential attitude. Hence a null hypothesis is a hypothesis of no difference. The main problem of the testing of hypothesis is to accept or reject the null hypothesis. As against the null hypothesis, the

alternative hypothesis specifies a range of the other values that the statistician believes to be true. Only one alternative hypothesis is tested against the null hypothesis.

2. **Test Static:** The next step is to compute an appropriate test static which is based upon an appropriate probability distribution. It is used to test whether the null hypothesis set up should be accepted or rejected.

3. **Type I and Type II Errors:** Acceptance or rejection of a hypothesis is based on the result of the sample information which may not always be consistent with the population. The decision may be correct in two ways:

- Accepting the null hypothesis when it is true.
- Rejecting the null hypothesis when it is false.

The decision may be wrong in two ways:
1. Rejecting the null hypothesis when it is true.
2. Accepting the null hypothesis when it is false.

Actual	**Decision**	
	Accept	**Reject**
H_0 is true	Correct Decision (No error)	Wrong (Type I Error)
H_0 is false	Wrong Decision (Type II Error)	Correct Decision (No Error)

4. Level of Significance: The next step is the fixation of the level of significance. Level of significance is the maximum probability of making Type I error. These types of risks should be kept low as far as possible say at 5% or 1%.

5. Critical region or Rejection Region: Critical region is the region of rejection of the null hypothesis. It is a region corresponding the value of the sample observations in the sample space which leads to rejection of the null hypothesis. A single function of the sample observations can be fixed and we can determine a region or range of values which lead to rejection of H_0 whenever the value of the function fails in this region.

If the observed set of results has the probability of more than 5% then the difference between the sample result and hypothetical parameter is not significant at 5% level i.e. the difference is due to fluctuations of sampling and H_0 is accepted. It implies that the sample result supports the hypothesis. Similarly, if the observed set of results has the probability less than 5% then the difference is significant at 5% level i.e. the difference is not wholly due to fluctuations of sampling and H_0 is rejected.

6. Tailed test observation: The critical region is represented by the portion of the area under the normal curve. The test of hypothesis is confirmed after looking into this table of hypothesis.

7. Taking the decision: Lastly the decision should be arrived at as to the accepting or rejecting the null hypothesis. If the computed value of the test static is less than the critical value as per the table, the hypothesis should be accepted or vice versa.

STANDARD ERROR

The standard deviation of the sampling distribution of a statistic such as mean, median etc. is known as standard error.

USES OF STANDARD ERROR

1. S.E. plays a vital role in the large sample theory and is used in testing of hypothesis.

If the difference between the observed and theoretical value of a statistic is greater than 1.96 times the S.E the hypothesis is rejected at 5% level of significance and say that the differenceis significant at 5% level.

2. The confidence or probable limits within which the population parameter is expected to lie, can be determined with the help of S.E.

3. It serves as a measure of reliability: As the S.E. increases the deviation of actual values from the expected one increase. This indicates that the sample is more unreliable.

9.4 TESTING OF HYPOTHESIS USING VARIOUS DISTRIBUTION TESTS

1. T-Distribution

W. S. Gosset under the nom de plume (pen name) of 'student' first found the distribution t = $\frac{\bar{x}-\mu}{s}$ of R.A. Fisher later on defined t = $\frac{\bar{x}-\mu}{s}\sqrt{n}$ correctly by the equation and found its distribution in 1926.

Using the notation of the previous article, we define a new statistic t by the equation

t = $\frac{\bar{x}-\mu}{s}\sqrt{n}$ = $\frac{\bar{x}-\mu}{\sigma s}\sqrt{(n-1)}$ or t = $\frac{\bar{x}-\mu}{s}\sqrt{(v+1)}$

where v = (n - 1) denote the number of degrees of freedom of t.

Then it may be shown that, for samples of size n from a normal population, the distribution oft is given by

$_{t2}$ y = $\frac{y0}{1+\frac{1}{v}v+1/2}$

If we choose y_0 so that the total area under the curve is unity; we shall get

y_0 = $\frac{1}{\sqrt{nB}\frac{n}{2},1/2}$

We can easily study the form of the t-distribution. Since only even powers of t appear in its equation it is symmetrical about t = 0 like the normal distribution, but unlike the normal distribution, it has $g_2 > 0$ so that it is more peaked than the normal distribution with the same standard deviation. Also y attain its maximum value at t = 0 so that the mode coincides withthe mean at t = 0. Again the limiting form of the distribution when y ® ¥ is given by

$y = y_0^{e - 1/2t2}$

It follows that t is normally distributed for large samples.

9.5 USES OF T-DISTRIBUTION

We have seen that if the sample is large, the use is made of the tables of the normal probability integral in interpreting the results of an experiment and on the basis of that to reject or accept the null hypothesis.

If, however, the sample size n is small, the normal probability tables will no longer be useful. Following are the uses of t-distribution:

a. To test the significance of the mean of a small random sample from a normal population.

b. To test the significance of the difference between the means of two samples taken from a normal population.

c. To test the significance of an observed coefficient of correlation including partial and rank correlations.

d. To test the significance of an observed regression coefficient.

2. z-TABLES OF POINTS AND THE SIGNIFICANCE TEST

We take y_0 so that the total area under the curve given by unity. The probability that we get a given value z_0 or greater on random sampling will be given by the area to the right of the ordinate at z_0. Tables for this probability for various values of z are not available, since this probability is difficult to evaluate, since it depends upon two numbers v_1 and v_2.

Fisher has prepared tables showing 5% and 1% points of significance for z. Colcord and Deming have prepared a table of 0.1 % points of significance. Generally, these tables are sufficient to enable us to gause the significance of an observed value of z.

It should be noted that the z-tables given only critical values corresponding to right-tail areas.Thus 5% points of z imply that the area to the right of the ordinate at the variable z is 0.05. A similar remark applies to 1% points of z. In other words, 5% and 1% points of z correspondto 10% and 2% levels of significance respectively.

USES OF z-DISTRIBUTION

1. To test the significance of mean of various samples having two or more than two values.

2. To test the significance of difference between two samples from given population.

3. To test the significance of an observed coefficients based upon the table prepared by "FISHER" since, the probability is difficult to evaluate based upon two numbers.

4. To test the significance on any observed set of values deriving its critical values corresponding to 5% and 1% of z (since it uses only "Right Tailed Test" for valuing the significance testing).

9.6 EXERCISES

Q1. Write Explanatory Notes on the following:

a. Type I Error

b. Type II Error

c. Procedure for hypothesis testing.

d. t-distribution test

e. z-distribution test

f. Uses of t-test and z-test

10

TESTING OF HYPOTHESIS - TWO SAMPLES (Related and Independent)

Unit Structure:

10.1. INTRODUCTION

Having discussed the problems relating to sampling of attributes in the previous section, we now come to the problems of sampling of variables such as height, weight etc. which may take any value. It shall not, therefore, be possible for us to classify each member of a sample under one of two heads, success or failure. The values of the variables given by different trials will spread over a range, which will be unlimited - limited by practical considerations, as in the case of weight of people or limited by theoretical considerations as in the case of correlation coefficient which cannot lie outside the range +1 to - 1.

There are three main objects in studying problems relating to sampling of variables:

i. To compare observation with expectation and to see how far the deviation of one from the other can be attributed to fluctuations of sampling;

ii. To estimate from samples some characteristic of the parent population, such as the mean of a variable; and

iii. To gauge the reliability of our estimates.

DIFFERENCES BETWEEN SMALL AND LARGE SAMPLES

In this section, we shall be studying problems relating to large samples only. Though it is difficult to draw a clear-cut line of demarcation between large and small samples, it is normally agreed amongst statisticians that a sample is to be recorded as large only if its size exceeds 30. The tests of significance used for dealing with problems relating to large samplesare different from the ones used for small samples for the reasons that the assumptions that we make in case of large samples do not hold good for small samples. The assumptions madewhile dealing with problems relating to large samples are:

i. The random sampling distribution of a statistic is approximately normal; and

ii. Values given by the samples are sufficiently close to the population value and can be sued in its place for calculating the standard error of the estimate.

While testing the significance of a statistic in case of large samples, the concept of standard error discussed earlier is used. The following is a list of the formulae for obtaining standard error for different statistics:

1. Standard Error of Mean

i. When standard deviation of the population is known

$$\text{S. E.}_{\bar{X}} = \frac{\sigma p}{\sqrt{n}}$$

where $\text{S.E.}_{\bar{X}}$ refers to the standard error of the mean

σp = Standard deviation of the population n = number of observations in the sample.

ii. When standard deviation of population is not known, we have to use standard deviation of the sample in calculating standard error of mean. Consequently, the formula for calculating standard error is

$$\text{S. E.}_{\bar{X}} = \frac{(sample)}{\sqrt{n}}$$

where σ denotes standard deviation of the sample.

It should be noted that if standard deviation of both sample as well as population are availablethen we should prefer standard deviation of the population for calculating standard error of mean.

Fiducial limits of population mean:

95% fiducial limits of population mean are

$$\bar{X} \pm 1.96 \frac{}{\sqrt{n}}$$

99% fiducial limits of population mean are

$$\bar{X} \pm 2.58 \frac{}{\sqrt{n}}$$

2. S.E. of Median or S.E. Med = $1.25331 \frac{\sigma}{\sqrt{n}}$
3. S.E. of Quartiles or S.E. = $1.36263 \frac{\sigma}{\sqrt{n}}$
4. S.E. of Quartile Deviation or $S.E_{QD} = 0.78672 \frac{\sigma}{\sqrt{n}}$
5. S.E. of Mean Deviation or $S.E._{MD} = 0.6028 \frac{\sigma}{\sqrt{n}}$
6. S.E. of Standard Deviation or $S.E._{\sigma} = \frac{\sigma}{\sqrt{2n}}$
7. S.E of Regression Estimate of Y on X or $S.E_{xy}. = \sigma_x \sqrt{1 - r^2}$
8. S.E. of Regression Estimate of X on Y or $S.E._{yx} = \sigma_y \sqrt{1 - r}$

The following examples will illustrate how standard error of some of the statistics iscalculated:

Examples

1. Calculate standard error of mean from the following data showing the amount paid by 100 firms in Calcutta on the occasion of Durga Puja.

Mid Value (Rs.)	39	49	59	69	79	89	99
No. of firms	2	3	11	20	32	25	7

Solution:

S.E. $\bar{X} = \frac{}{\sqrt{n}}$

Calculation of Standard Deviation

Mid-valuem	F	(m-69)/10 d'	fd'	fd'^{2}
39	2	-3	-6	18
49	3	-2	-6	12
59	11	-1	-11	11
69	20	0	0	0
79	32	+1	32	32
89	25	+2	50	100
99	7	+3	21	63
	N = 100		Σfd' = 80	Σ fd'^{2} = 236

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times C = \sqrt{\frac{236}{100} - \left(\frac{80}{100}\right)^2} \times 100 = \sqrt{2.36 - 0.64} \times 10 = 1.311 \times 10 = 13.11$$

S.E. $\bar{X} = \frac{13.11}{\sqrt{100}} = \frac{13.11}{10} = 1.311.$

10.2. STANDARD ERROR OF THE DIFFERENCE BETWEEN THE MEANS OF TWOSAMPLES

i. If two independent random samples with n_1 and n_2 numbers respectively are drawn from the same population of standard deviation σ, the standard error of the difference between the sample means is given by the formula:

S.E. of the difference between sample means

$$= \sqrt{\sigma 2 \quad \frac{1}{n1+n2}}$$

If σ is unknown, sample standard deviation for combined samples must be substituted.

ii. If two random samples with $\bar{X}_1$, σ_1, n_1 and $\bar{X}_2$, σ_2, n_2 respectively are drawn from different populations, then S.E. of the different between the means is given by the formula:

$$= \sqrt{\frac{\sigma 12}{n1} + \frac{\sigma 22}{n2}}$$ and where σ_1 and σ_2 are unknown.

S.E. of difference between means

$$= \sqrt{\frac{S12}{n1} + \frac{S22}{n2}}$$

where S_1 and S_2 represent standard deviation of the two samples.

EXAMPLES

1. Intelligence test on two groups of boys and girls gave the following results:

	Mean	S.D.	N
Girls	75	15	150
Boys	70	20	250

Is there a significant difference in the mean scores obtained by boys and girls?

Solution:

Let us take the hypothesis that there is no significant difference in the mean scored obtainedby boys and girls.

$$S.E.(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma 12}{n1} + \frac{\sigma 22}{n2}}$$ where $\sigma_1 = 15$, $\sigma_2 = 20$, $n_1 = 150$ and $n_2 = 250$

Substituting the values

$$S.E.(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{(15)2}{150} + \frac{(20)2}{250}} = \sqrt{1.5 + 1.6} = 1.781$$

$$\frac{Difference}{S.E} = \frac{75-70}{1.781} = 2.84$$

Since the difference is more than 2.58 (1% level of significance) the hypothesis is rejected. There seems to be a significant difference in the mean score obtained by boys and girls.

STANDARD ERROR OF THE DIFFERENCE BETWEEN TWO STANDARD DEVIATIONS

In case of two large random samples, each drawn from a normally distributed population, the S.E. of the difference between the standard deviation is given by:

$$S.E\ (\sigma_1 - \sigma_2) = \sqrt{\frac{\sigma 12 + \sigma 22}{2n1+2n2}}$$

where population standard deviations are not known

$$S.E\ (S_1 - S_2) = \sqrt{\frac{S12 + S22}{2n1+2n2}}$$

EXAMPLE

1. Intelligence test of two groups of boys and girls gave the following results:

Girls: Mean = 84, S.D. = 10, n = 121

Boys: Mean = 81, S.D. = 12, n = 81

a. Is the difference is mean scores significant?

b. Is the difference between standard deviations significant?

SOLUTION:

a. Let us take the hypothesis that there is no difference in mean scores.

S.E.$(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ where $\sigma_1 = 10$, $\sigma_2 = 12$, $n_1 = 121$ and $n_2 = 81$

Substituting the values

S.E.$(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{(10)^2}{121} + \frac{(12)^2}{81}} = \sqrt{100/121 + 144/81} = \sqrt{2.604} = 1.61$

Difference of means (84 - 81) = 3

$\frac{Difference}{S.E} = \frac{3}{1.61} = 1.86$

Since the difference is less than 1.96 S.E. (5% level of significance) the given factors support hypothesis. Hence the difference in mean scores of boys and girls is not significant.

b. Let us take the hypothesis that there is no difference between the standard deviation of the two samples.

S.E $(\sigma_1 - \sigma_2) = \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2n_1 + 2n_2}}$ where $\sigma 1 = 10$, $\sigma 2 = 12$, $n_1 = 121$, $n_2 = 81$

S.E $(\sigma_1 - \sigma_2) = \sqrt{\frac{(10)^2}{2 X 121} + \frac{(12)^2}{2 X 81}} = \sqrt{\frac{100}{242} + \frac{144}{162}} = \sqrt{1.302} = 1.14$

Difference between the two standard deviations - (12 - 10) = 2

$\frac{Difference}{S.E} = \frac{2}{1.14} = 1.75$

Since the difference is less than 1.96 S.E. (5% level of significance) the given factors support hypothesis. Hence the difference in mean scores of boys and girls is not significant.

TWO-SAMPLE Z-TEST FOR COMPARING TWO MEANS

Requirements: Two normally distributed but independent populations, σ is known

Hypothesis test

Formula: $z = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of the two samples, Δ is the hypothesized difference between the population means (0 if testing for equal means), σ_1 and σ_2 are the standard deviations of the two populations, and n_1 and n_2 are the sizes of the two samples.

The amount of a certain trace element in blood is known to vary with a standard deviation of 14.1 ppm (parts per million) for male blood donors and 9.5 ppm for female donors. Random samples of 75 male and 50 female donors yield concentration means of 28 and 33 ppm, respectively. What is the likelihood that the population means of concentrations of the element are the same for men and women?

Null hypothesis: H_0: $\mu_1 = \mu_2$

or H_0: $\mu_1 - \mu_2 = 0$

alternative hypothesis: $H_a: \mu_1 \neq \mu_2$
or: $H_a: \mu_1 - \mu_2 \neq 0$ $z = \frac{28-33-0}{\sqrt{\frac{14.1^2}{75}+\frac{9.5^2}{50}}} = \frac{-5}{\sqrt{2.65+1.81}} = -2.37$

The computed *z*-value is negative because the (larger) mean for females was subtracted from the (smaller) mean for males. But because the hypothesized difference between the populations is 0, the order of the samples in this computation is arbitrary— could just as well have been the female sample mean and the male sample mean, in which case *z* would be 2.37 instead of –2.37. An extreme *z*-score in either tail of the distribution (plus or minus) will lead to rejection of the null hypothesis of no difference.

The area of the standard normal curve corresponding to a *z*-score of –2.37 is 0.0089. Because this test is two-tailed, that figure is doubled to yield a probability of 0.0178 that the population means are the same. If the test had been conducted at a pre-specified significance level of $\alpha < 0.05$, the null hypothesis of equal means could be rejected. If the specified significance level had been the more conservative (more stringent) $\alpha < 0.01$, however, the null hypothesis could not be rejected.

In practice, the two-sample *z*-test is not used often, because the two population standard deviations σ_1 and σ_2 are usually unknown. Instead, sample standard deviations andthe *t*-distribution are used.

Inferences About the Difference Between Two Population Means for Paired Data
Paired samples: The sample selected from the first population is related to thecorresponding sample from the second population.
It is important to distinguish independent samples and paired samples. Some examples are given as follows.

Compare the time that males and females spend watching TV.
Think about the following, then click on the icon to the left to compare your answers.

A. We randomly select 20 males and 20 females and compare the average time they spend watching TV. Is this an independent sample or paired sample?

B. We randomly select 20 couples and compare the time the husbands and wives spend watching TV. Is this an independent sample or paired sample?

The paired *t*-test will be used when handling hypothesis testing for paired data.

The Paired *t*-Procedure
Assumptions:

1. Paired samples
2. The differences of the pairs follow a normal distribution or the number of pairs is large (note here that if the number of pairs is < 30, we need to check whether the *differences* are normal, but we do not

need to check for the normality of *eachpopulation*)

Hypothesis:
H0:μd=0
Ha:μd≠0
OR

H0:μd=0
Ha:μd<0
OR

H0:μd=0
Ha:μd>0
***t*-statistic**:

Let *d* = differences between the pairs of data, then $\bar{d}$ = mean of these differences.
The test statistics is: $t*=\bar{d}-0sd/n\sqrt{}$
degrees of freedom = *n* - 1
where *n* denotes the number of pairs or the number of differences.
Paired *t*-interval:
$\bar{d}\pm t\alpha/2\cdot sdn--\sqrt{}$

Note: $sd^{-}=sdn\sqrt{}$ where sd^{-} is the standard deviation of the sample differences.

Example: Drinking Water
Trace metals in drinking water affect the flavor and an unusually high concentration can pose a health hazard. Ten pairs of data were taken measuring zinc concentration in bottom water and surface water (zinc_conc.txt).
Does the data suggest that the true average concentration in the bottom water exceeds that of surface water?

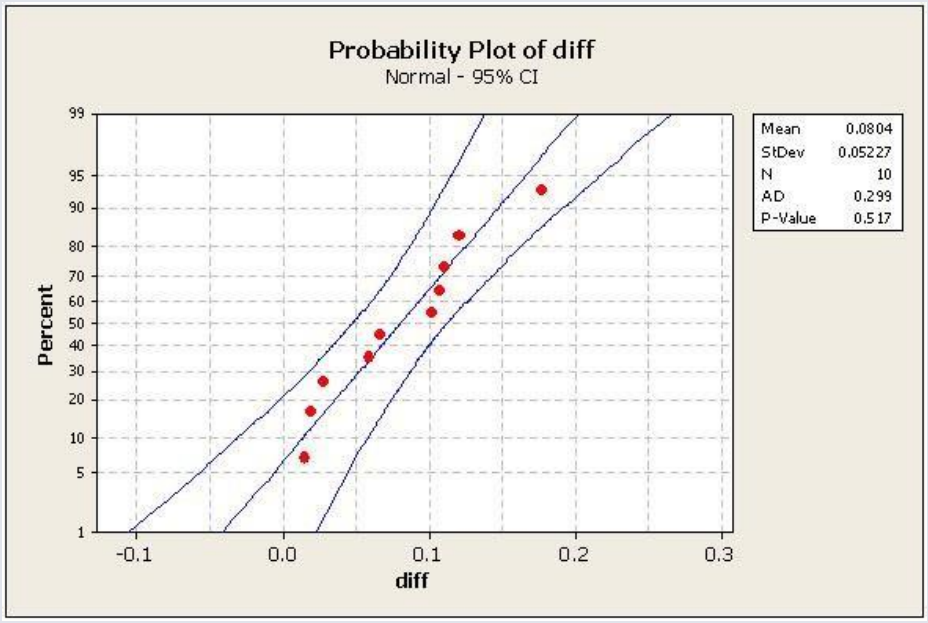
	Location									
	1	2	3	4	5	6	7	8	9	10
Zinc concentration in bottom water	.430	.266	.567	.531	.707	.716	.651	.589	.469	.723
Zinc concentration in surface water	.415	.238	.390	.410	.605	.609	.632	.523	.411	.612

To perform a paired *t*-test for the previous trace metal example:
Assumptions:
1. Is this a paired sample? - Yes.

2. Is this a large sample? - No.

3. Since the sample size is not large enough (less than 30), we need to check whether the differences follow a normal distribution.

In Minitab, we can use Calc > calculator to obtain *diff = bottom - surface* and then perform a probability plot on the differences.



Thus, we conclude that the difference may come from a normal distribution.

Step 1. Set up the hypotheses:
$H_0: \mu_d = 0$
$H_a: \mu_d > 0$
where 'd' is defined as the difference of bottom - surface.

Step 2. Write down the significance level $\alpha = 0.05$.
Step 3. What is the critical value and the rejection region?
$\alpha = 0.05$, $df = 9$
$t_{0.05} = 1.833$
rejection region: $t > 1.833$
Step 4. Compute the value of the test statistic:
$t_* = \bar{d} s_d \sqrt{n} = 0.0804 0.05231 0 \sqrt{} = 4.86$

Step 5. Check whether the test statistic falls in the rejection region and determine whether to reject H_o.
$t_* = 4.86 > 1.833$
reject H_0
Step 6. State the conclusion in words.
At $\alpha = 0.05$, we conclude that, on average, the bottom zinc concentration is higher than the surface zinc concentration.

Using Minitab to Perform a Paired *t*-Test

You can used a paired *t*-test in Minitab to perform the test. Alternatively, you can perform a 1-sample *t*-test on difference = bottom - surface.
1. Stat > Basic Statistics > Paired t

2. Click 'Options' to specify the confidence level for the interval and the alternative hypothesisyou want to test. The default null hypothesis is 0.

The Minitab output for paired T for bottom - surface is as follows:

```
Paired T for bottom - surface
```

	N	Mean	StDev	SE Mean
Bottom	10	0.5649	0.1468	0.0464
Surface	10	0.4845	0.1312	0.0415
Difference	10	0.0804	0.0523	0.0165

```
95% lower bound for mean difference: 0.0505
T-Test of mean difference = 0 (vs > 0): T-Value = 4.86 P-Value =
0.000
```

Note: In Minitab, if you choose a lower-tailed or an upper-tailed hypothesis test, an upper or lower confidence bound will be constructed, respectively, rather than a confidence interval.

Click on the 'Minitab Movie' icon to display a walk through of 'Conducting a Paired t-Test'.
Using the *p*-value to draw a conclusion about our example:
p-value = 0.000 < 0.05
Reject H_0 and conclude that bottom zinc concentration is higher than surface zinc concentration.

Note: For the zinc concentration problem, if you do not recognize the paired structure, but mistakenly use the 2-sample *t*-test treating them as independent samples, you will not be able to reject the null hypothesis. This demonstrates the importance of distinguishing the two types of samples. Also, it is wise to design an experiment efficiently whenever possible.
What if the assumption of normality is not satisfied? In this case we would use a nonparametric1-sample test on the difference.

10.3. HYPOTHESIS TESTING OF THE DIFFERENCE BETWEEN TWO POPULATIONMEANS

B) Hypothesis testing of the difference between two population means

This is a two sample z test which is used to determine if two population means are equal or unequal. There are three possibilities for formulating hypotheses.

l. $H_0 : \mu_1 = \mu_2$ $H_A : \mu_1 \neq \mu_2$

2. $H_0 : \mu_1 \geq \mu_2$ $H_A : \mu_1 < \mu_2$

3. $H_0 : \mu_1 \leq \mu_2$ $H_A : \mu_1 > \mu_2$

Procedure

The same procedure is used in three different situations

- Sampling is from normally distributed populations with known variances

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}}$$

- Sampling from normally distributed populations where population variances are unknown
  - population variances equal

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{s_p^2 / n_1 + s_p^2 / n_2}}$$

This is with *t* distributed as Student's *t* distribution with ($n_1 + n_2$ -2) degrees of freedom and a pooled variance.

  - population variances unequal

When population variances are unequal, a distribution of *t'* is used in a manner similar to calculations of confidence intervals in similar circumstances.

- Sampling from populations that are not normally distributed

If both sample sizes are 30 or larger the central limit theorem is in effect. The test statistic is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}}$$

If the population variances are unknown, the sample variances are used.

Sampling from normally distributed populations with population variances known

Example 7.3.1

Serum uric acid levels

Is there a difference between the means between individuals with Down's syndrome andnormal individuals?

(1) Data

= 4.5 $\bar{x}_1$ $n_1 = 12$ σ_1^2 1

= 3.4 $\bar{x}_2$ $n_2 = 15$ σ_2^2 1.5

= .05 α

(2) Assumptions

- two independent random samples

- each drawn from a normally distributed population

(3) Hypotheses

: $H_0 \quad \mu_1 = \mu_2$

: $H_A \quad \mu_1 \neq \mu_2$

(4) Test statistic

This is a two sample z test.

(a) Distribution of test statistic

If the assumptions are correct and H_0 is true, the test statistic is distributed as the normal distribution.

(b) Decision rule

With α = .05, the critical values of z are -1.96 and +1.96. We reject H_0 if z < -1.96 or z > +1.96.

(5) Calculation of test statistic

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}}$$

$$z = \frac{(4.5 - 3.4) - 0}{\sqrt{1/12 + 1.5/15}} = \frac{1.1}{.4282} = 2.57$$

(6) Statistical decision

Reject H_0 because 2.57 > 1.96.

(7) Conclusion

From these data, it can be concluded that the population means are not equal. A 95%confidence interval would give the same conclusion.

p = .0102.

Sampling from normally distributed populations with unknown variances

With equal population variances, we can obtain a pooled value from the sample variances.

Example 7.3.2

Lung destructive index

We wish to know if we may conclude, at the 95% confidence level, that smokers, in general,have greater lung damage than do non-smokers.

(1) Data

Smokers:	$\bar{x}_1 = 17.5$	$n_1 = 16$	$s_1^2 = 4.4752$
Non-Smokers:	$\bar{x}_2 = 12.4$	$n_2 = 9$	$s_2^2 = 4.8492$

$\alpha = .05$

Calculation of Pooled Variance:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$s_p^2 = \frac{(15)(4.4711) + (8)(4.8492)}{16 + 9 - 2}$$

$$s_p^2 = \frac{299.86 + 188.12}{23}$$

$$s_p^2 = 21.2165$$

(2) Assumptions

- independent random samples
- normal distribution of the populations
- population variances are equal

(3) Hypotheses

: $H_0 \quad \mu_1 \leq \mu_2$

: $H_A \quad \mu_1 > \mu_2$

(4) Test statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{s_p^2 / n_1 + s_p^2 / n_2}}$$

(a) Distribution of test statistic

If the assumptions are met and H_0 is true, the test statistic is distributed as Student's t distribution with 23 degrees of freedom.

(b) Decision rule

With α = .05 and df = 23, the critical value of *t* is 1.7139. We reject H_0 if t > 1.7139.

(5) Calculation of test statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{s_p^2/n_1 + s_p^2/n_2}}$$

$$t = \frac{(17.5 - 12.4) - 0}{\sqrt{21.2165/16 + 21.2165/9}} = \frac{5.1}{1.92} = 2.6563$$

(6) Statistical decision

Reject H_0 because 2.6563 > 1.7139.

(7) Conclusion

On the basis of the data, we conclude that $\mu_1 > \mu_2$.

Actual valuest =
2.6558
p = .014

Sampling from populations that are not normally distributed

Example 7.3.4

These data were obtained in a study comparing persons with disabilities with persons without disabilities. A scale known as the Barriers to Health Promotion Activities for Disabled Persons (BHADP) Scale gave the data. We wish to know if we may conclude, at the 99% confidence level, that persons with disabilities score higher than persons without disabilities.

(1) Data

Disabled: $\bar{x}_1 = 31.83$ $n_1 = 132$ $s_1 = 7.93$

Nondisabled: $\bar{x}_2 = 25.07$ $n_2 = 137$ $s_2 = 4.80$

α = .01

(2) Assumptions

- independent random samples

(3) Hypotheses

: H_0 $\mu_1 \leq \mu_2$

: H_A $\mu_1 > \mu_2$

(4) Test statistic

Because of the large samples, the central limit theorem permits calculation of the z score as opposed to using *t*. The z score is calculated using the given sample standard deviations.

(a) Distribution of test statistic

If the assumptions are correct and H_0 is true, the test statistic is approximately normally distributed

(b) Decision rule

With α = .01 and a one tail test, the critical value of z is 2.33. We reject H_0 z > 2.33.

(5) Calculation of test statistic

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{s_1^2 / n_1 + s_2^2 / n_2}}$$

$$z = \frac{(31.83 - 25.07) - 0}{\sqrt{(7.93)^2 / 132 + (4.80)^2 / 137}} = \frac{6.76}{.8029} = 8.42$$

(6) Statistical decision

Reject H_0 because 8.42 > 2.33.

(7) Conclusion

On the basis of these data, the average persons with disabilities score higher on the BHADPtest than do the nondisabled persons.

Actual valuesz =
8.42
p = 1.91 x 10-17

Paired comparisons

Sometimes data comes from nonindependent samples. An example might be testing "beforeand after" of cosmetics or consumer products. We could use a single random sample and do"before and after" tests on each person. A hypothesis test based on these data would be called a *paired comparisons test*. Since the observations come in pairs, we can study the difference, d, between the samples. The difference between each pair of measurements is called di.

Test statistic

With a population of n pairs of measurements, forming a simple random sample from a

normally distributed population, the mean of the difference, μ_d , is tested using the following implementation of *t*.

$$t = \frac{\bar{d} - \mu_{d_0}}{s_{\bar{d}}}$$

$\bar{d}$ is the sample mean difference

μ_{d_0} is the hypothesized mean difference

$s_{\bar{d}} = \frac{s_d}{\sqrt{n}}$ -- the standard error

n is the number of sample differences

s_d is the standard deviation of the sample diffences

Paired comparisons

Example 7.4.1

Very-low-calorie diet (VLCD) Treatment

Table gives B (before) and A (after) treatment data for obese female patients in a weight-loss program.

		Table of Weight Loss Data for Example 7.4.1							
		Weights (kg) of Obese Women Before and After 12-Week VLCD Treatment							
B:	117.3	111.4	98.6	104.3	105.4	100.4	81.7	89.5	78.2
A:	83.3	85.9	75.8	82.9	82.3	77.7	62.7	69.0	63.9

We calculate di = A-B for each pair of data resulting in negative values meaning that the participants lost weight.

We wish to know if we may conclude, at the 95% confidence level, that the treatment iseffective in causing weight reduction in these people.

(1) Data

Values of di are calculated by subtracting each A from each B to give a negative number. Onthe TI-83 calculator place the A data in L1 and the B data in L2. Then make L3 = L1 - L2 and the calculator does each calculation automatically.

In Microsoft Excel put the A data in column A and the B data in column B, without using column headings so that the first pair of data are on line 1. In cell C1, enter the following formula: =a1-b1. This calculates the difference, di, for B - A. Then copy the formula down column C until the rest of the differences are calculated.

n = 9

= .0

(2) Assumptions

- the observed differences are a simple random sample from a normally distributed population of differences

(3) Hypotheses

: $H_0 \quad \mu_d \geq 0$

: $H_A \quad \mu_d < 0$ (meaning that the patients lost weight)

(4) Test statistic

The test statistic is *t* which is calculated as

$$t = \frac{\bar{d} - \mu_{d_0}}{s_{\bar{d}}}$$

(a) Distribution of test statistic

The test statistic is distributed as Student's t with 8 degrees of freedom

(b) Decision rule

With α = .05 and 8 df the critical value of t is -1.8595. We reject H_0 if t < -1.8595.

(5) Calculation of test statistic

$$\bar{d} = \frac{\sum di}{n} = \frac{-203.3}{9} = -22.5889$$

$$s_d^2 = 28.2961$$

$$t = \frac{\bar{d} - \mu_{d_0}}{s_{\bar{d}}} = \frac{-22.5899 - 0}{\sqrt{28.2961/9}} = -12.7395$$

(6) Statistical decision

Reject H_0 because -12.7395 < -1.8595
p = 6.79 x 10-7

(7) Conclusion

On the basis of these data, we conclude that the diet program is effective.Other

considerations

- a confidence interval for μ_d can be constructed
- z can be used if the variance is known or if the sample is large.

CAUTION WHILE USING T-TEST

While drawing inferences on the basis of t-test it should be remembered that the conclusions arrived at on the basis of the 't-test' are justified only if the assumptions upon which the test is based are true. If the actual distribution is not normally distributed then, strictly speaking, the t-test is not justified for small samples. If it is not a random sample, then the assumption that the observations are statisticallyindependent is not justified and the conclusions based on the t-test may not be correct. The effect of violating the normality assumption is slight when making inference about means provided that the sampling is fairly large when dealing with small samples. However, it is a good idea to check the normality assumption, if possible. A review of similar samples or related research may provide evidence as to whether or not the population is normally distributed.

LIMITATIONS OF THE TESTS OF SIGNIFICANCE

In testing statistical significance the following points must be noted:

1. They should not be used mechanically: Tests of significance are simply the raw materials from which to make decisions, not decisions in themselves. There may be situations where real differences exist but do not produce evidence that they are statistically significant or the other way round. In each case it is absolutely necessary to exercise great care before taking a decision.
2. Conclusions are to be given in terms of probabilities and not certainties: When a test shows that a difference was statistically significant, it suggests that the observed difference is probably not due to chance. Thus statements are not made with certainty but with a knowledge of probability. "Unusual" events do happen once in a while.
3. They do not tell us "why" the difference exists: Though tests can indicate that a difference has statistical significance, they do not tell us why the difference exists. However, they do suggest the need for further investigation in order to reach definite answers.
4. If we have confidence in a hypothesis it must have support beyond the statistical evidence. It must have a rational basis. This phrase suggests two conditions: first, the hypothesis must be 'reasonable' in the sense of concordance with a prior expectation. Secondly, the hypothesismust fit logically into the relevant body of established knowledge.

The above points clearly show that in problems of statistical significance as in other statistical problems, technique must be combined with good judgement and knowledge of the subject-matter.

EXERCISES

Q1. Explain the concept of standard error and discuss its role in the large sample theory.

2. Explain briefly the procedure followed in testing hypothesis.

3. Give some important applications of the t-test and explain how it helps in making business decisions.

4. What is null hypothesis? How is it different from alternative hypothesis?

5. The mean life of a sample of 10 electric light bulbs was found to be 1, 456 hours with standard deviation of 423 hours. A second sample of 17 bulbs chosen from a different batch showed a mean life of 1, 280 hours with standard deviation of 398 hours. Is there a significant difference between the means of the two batches?

6. Test the significance of the correlation $r = 0.5$ from a sample of size 18 against hypothetical correlation $\rho = 0.7$.

7. A correlation coefficient of 0.2 is discovered in a sample of 28 pairs of observations. Use z-test to find out if this is significantly different from zero.

8. How many pairs of observations must be included in a sample in order that an observed correlation coefficient of value 0.42 shall have a calculated value of t greater than 2.72?

9. State the cautions of using t-test.

10. State the limitations of tests of significance.