BASIC STATISTICS FOR MANAGEMENT

UNIT 1

INTRODUCTION TO STATISTICS

Statistics and Rationale, Frequency Distribution, Classification and Tabulation, Diagrammatical and Graphical Representation.

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 completely ruled out. The word has now acquired a wider meaning.

STATISTICS IN PLURAL

Statistics in plural refer to any set of data or information. The president of a company may call for 'statistics on the sales of northern region' or an MP may quote the statistics on price-rise in agricultural products. More familiar examples for the students will be the marks of students in a class, the ages of children in primary school.

Prof. Secrist defines the word 'Statistics' in the first sense as follows"

"By **Statistics** we mean aggregate of facts affected to a marked extend by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other."

This definition gives all the characteristics of Statistics:

i. Aggregate of Facts: A single isolated figure is not 'Statistics.' Marks of one student in one subject will not called Statistics. But, if we consider the marks of all the students in the class in a particular subject, they will be called 'Statistics.'

ii. Affected by Multiplicity of causes: There are various causes for the changes in the data, the marks of the students depend upon, the intelligence of students, their capacity and desire to work etc.

iii. Numerically expressed: Unless the characteristics have some numerical measurement they will not be called Statistics. The statement 'A student writes very good English' is not Statistics. But if marks of the whole class in 'English' are given they will be called 'Statistics.'

iv. Enumerated or Estimated according to reasonable standards of accuracy: However much a person tries, it is not possible to attain perfect accuracy whether we actually measure or estimate the characteristic. But a certain standard of accuracy should be set up according to the problem under consideration. The estimate for the cost of big project may be correct up to Rs. 1, 000 but for household expenses it should be correct up to a rupee.

v. Collected in a systematic manner: There should be a method in the manner of collection, then only the figures will be reliable and useful.

vi. Collected for a predetermined purpose: Unless we know the purpose, the data collected may not be sufficient. Besides some unnecessary information may be collected which will be a waste of time and money.

vii. Placed in relation to each other: Only when we want to compare characteristics, which have some relation with each other, we collect Statistics. The wages of fathers and ages of sons should not be collected together. But we can have ages and heights of a group of persons, so that we can find the relation between the two.

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 proper collection 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 not be 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.

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 with thousands 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 of the 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.

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.

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 that it 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.

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 on sample 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.

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.

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 houses in 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 individual to 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.

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.

CHAPTER 1
UNIT 2

DATA: COLLECTION AND PRESENTATION

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.

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.

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.

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.

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.

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 errors can be tolerated in the inquiry. The editing of secondary data is simpler than that of primary data.

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.

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 like heights, 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 time of occurrence, such as years, months, weeks, days etc.... For example, Time Series Data.

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.

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.

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 data are placed in row and columns.
3. Classification is a process of statistical analysis where as tabulation is a process of presenting the data in suitable form.

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 the class 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 value 19 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 called the upper class boundary of the class. It is important to note that the upper class boundary of a 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.

CONSTRUCTIN 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.

4. 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.
5. 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.
6. 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

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 called a 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 dos not always prove to be interesting to a layman. Too many figures are often confusing and 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:

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

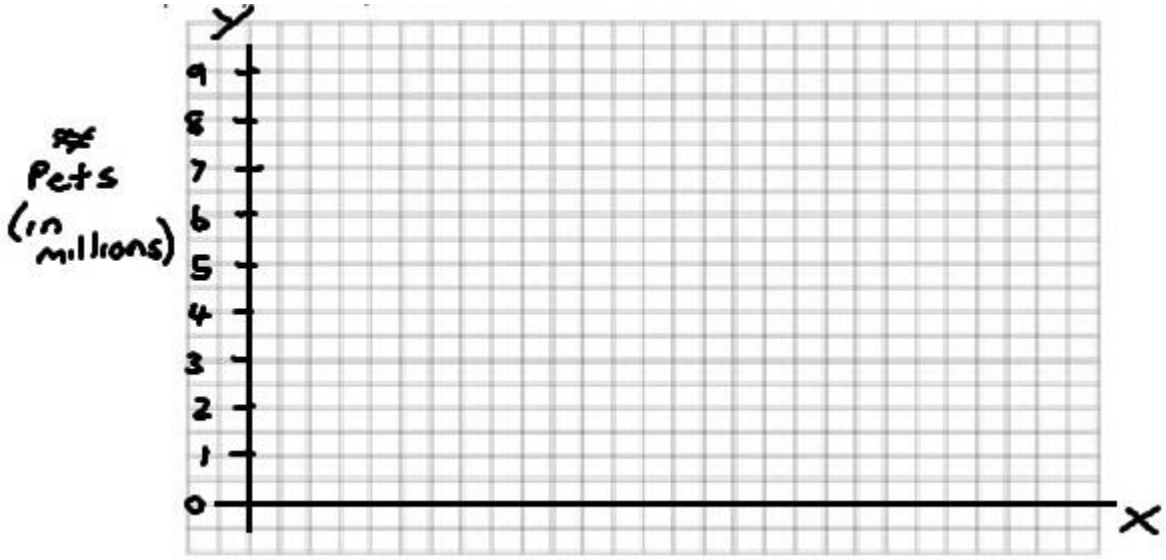
1. SIMPLE BAR CHART

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:

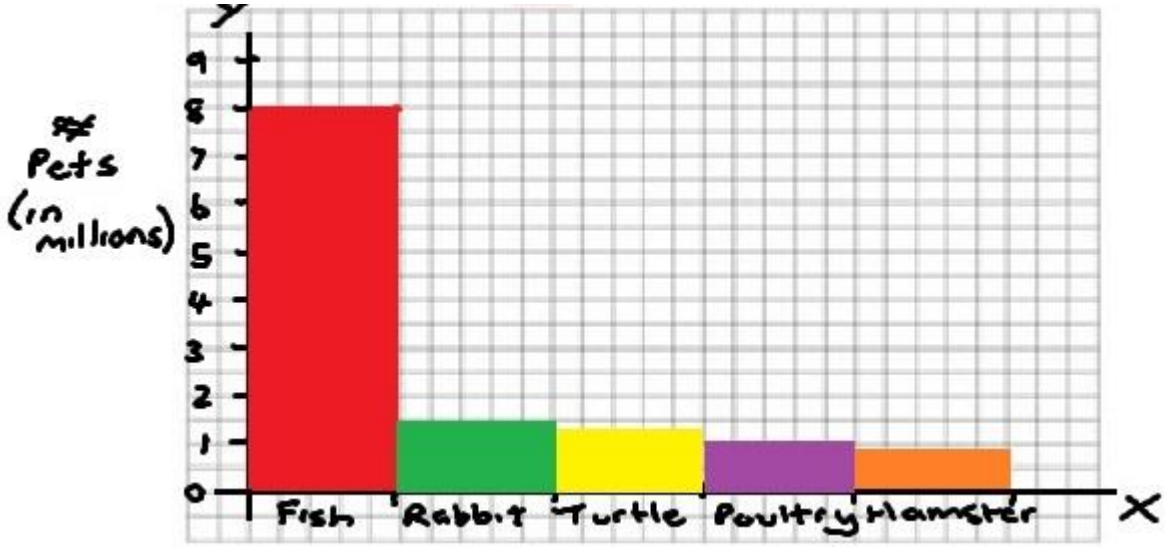
- Draw two perpendicular lines one horizontally and the other vertically at an appropriate place of the paper.
- Take the basis of classification along horizontal line (X-axis) and the observed variable along vertical line (Y-axis) or vice versa.
- Mark signs of equal breadth for each class and leave 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.*

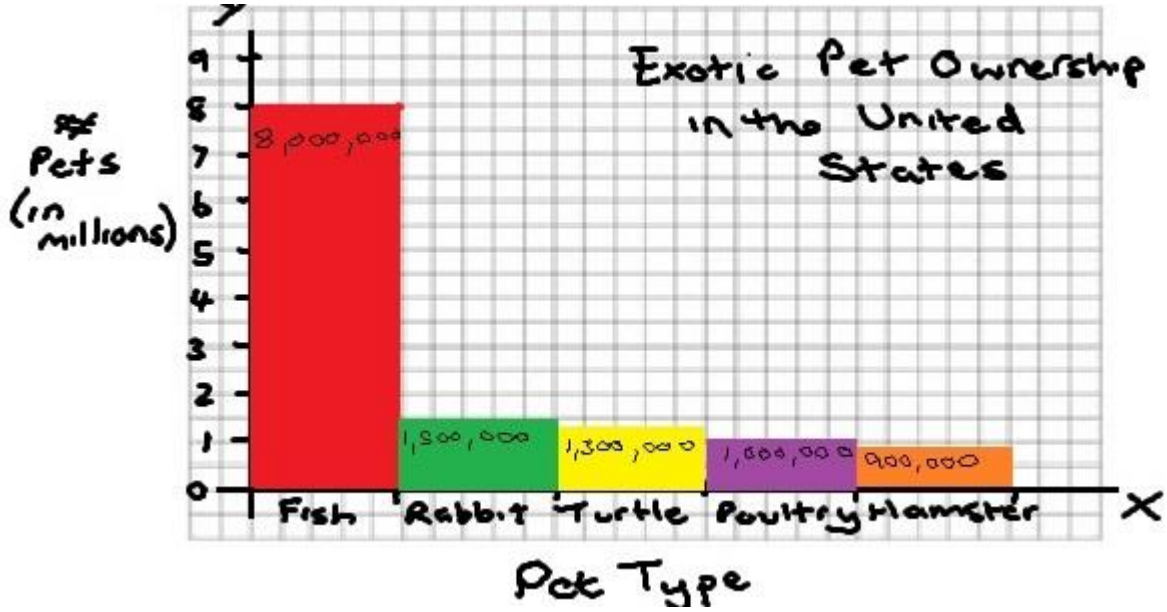
Step 1: **Number the Y-axis** with the dependent variable. The dependent variable is the one being tested in an experiment. In this sample question, the study wanted to know how many pets were in U.S. households. So the number of pets is the dependent variable. The highest number in the study is 8,000,000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

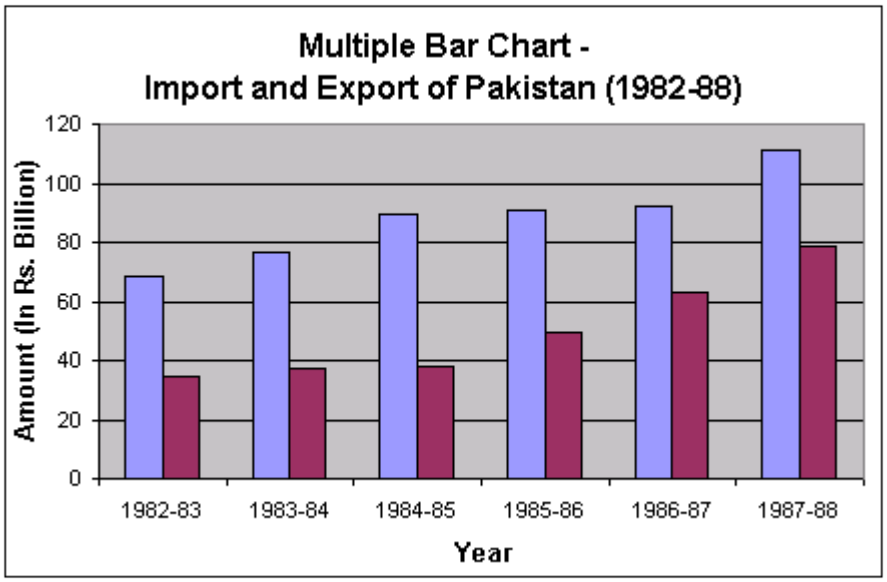
2. MULTIPLE BAR CHART

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



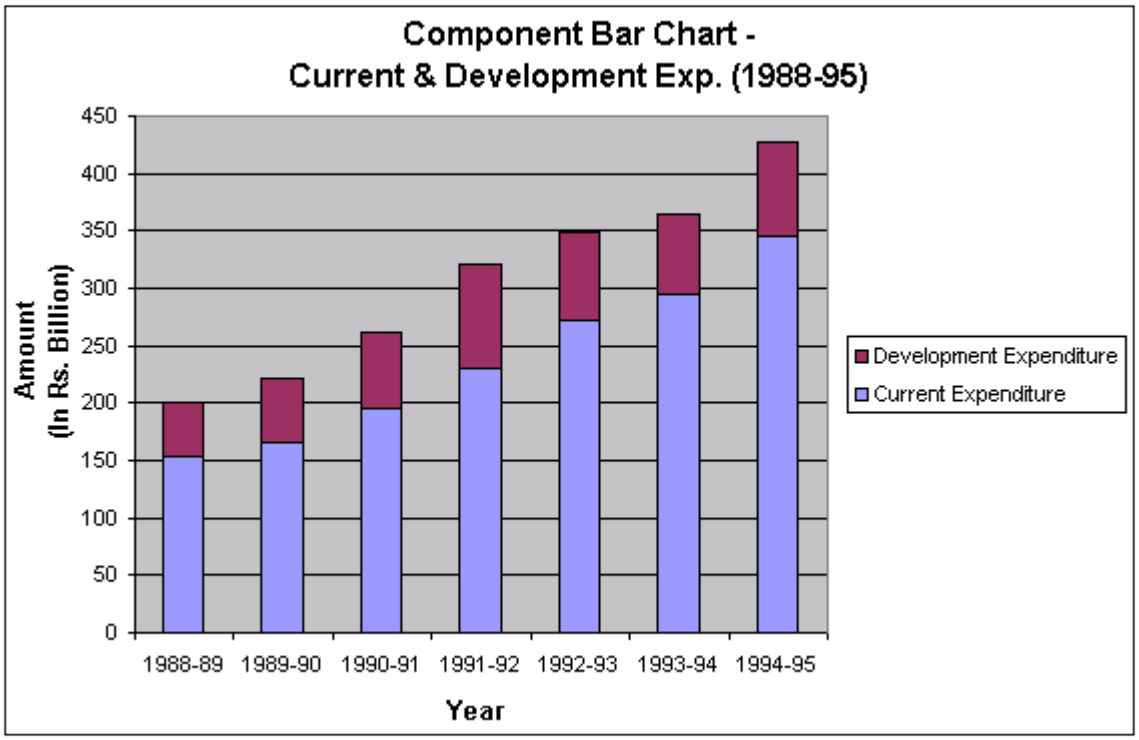
3. COMPONENT BAR CHART

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 step and 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.

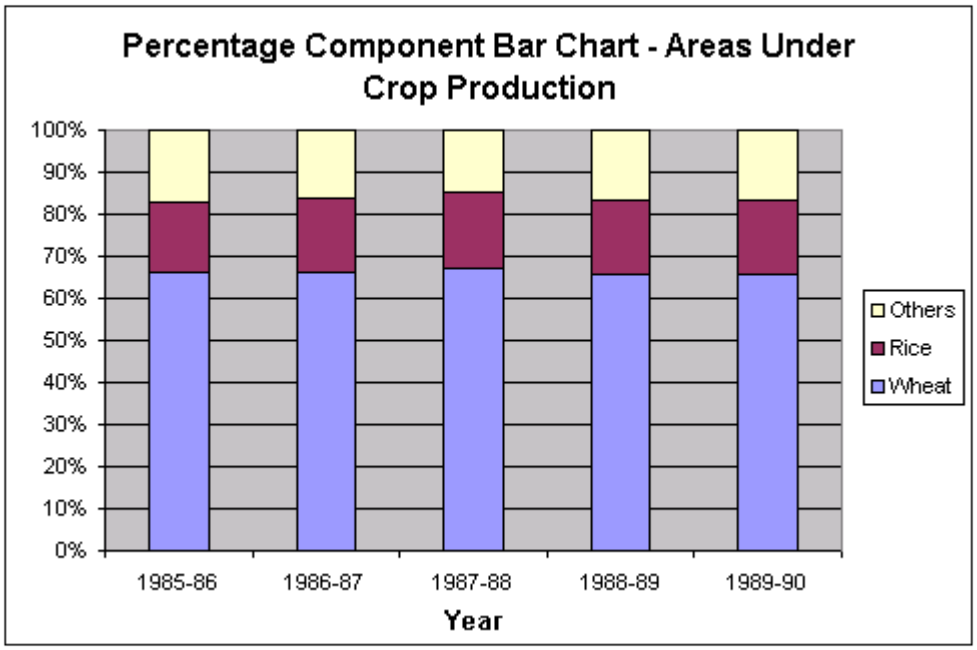
Areas Under Crop Production (1985-90)

('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 are drawn with radii proportional to the square root of the quantities because the area of a circle is $A = 2\pi r^2$

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:

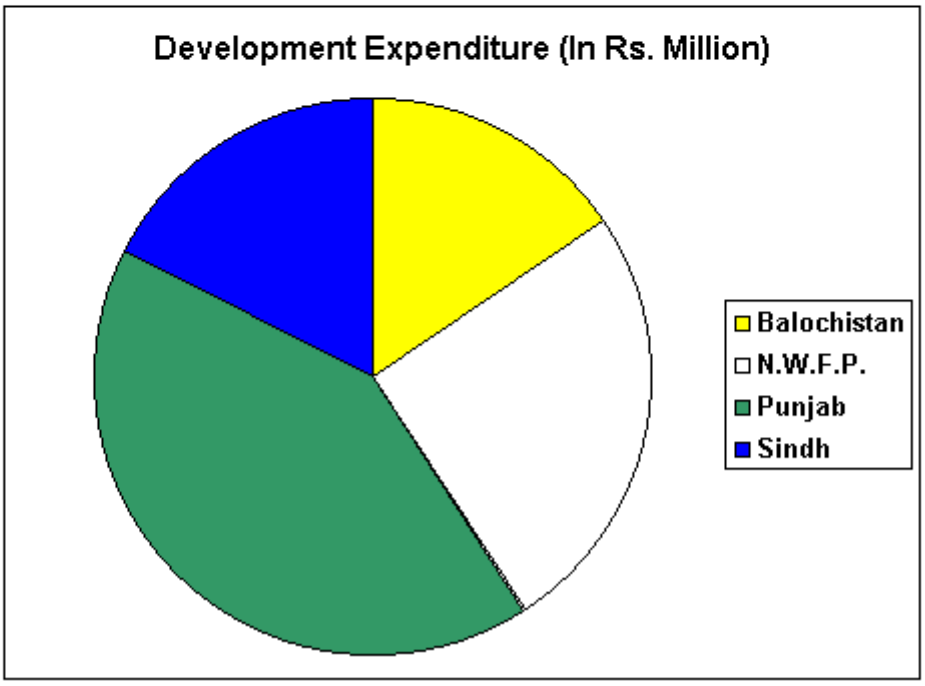
Angle of Sector = $\frac{Component\ Part}{Total} \times 360°$

These angles are made in the circle by means of a protractor to show different components. 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	



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

Chapter 3

MEASURES OF DISPERSION

Introduction to Measures of Dispersion, Various methods of Dispersion like Range, Mean Deviation, Variance, Standard Deviation and Coefficient of Variation, Measurement of Shapes like Skewness and Kurtosis.

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 of the 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 something connected with human life and study is very important for mankind.

DISPERSION

The word dispersion has a technical meaning in Statistics. The average measures the centre of the data. It is one aspect observations. Another feature of the observations is as to how the observations are spread about the centre. The observation may be close to the centre or they may be spread away from the centre. If the observation are close to the centre (usually the arithmetic mean or median), we say that dispersion, 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	$\bar{X}_A = 50$
Group B: 30, 40, 50, 60, 70	$\bar{X}_B = 50$
Group C; 40, 50, 60, 70, 80	$\bar{X}_C = 60$

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 which is 50. Thus in group B and C the means are different but their dispersion is the same. In group A and C the means are different and their dispersions are also different. Dispersion is an important feature of the observations and it is measured with the help of the measures of dispersion, scatter or variation. The word variability is also used for this idea of dispersion.

The study of dispersion is very important in statistical data. If in a certain factory there is consistence in the wages of workers, the workers will be satisfied. But if 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.

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 commonly used are:

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

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).

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 = $\frac{Xm - Xo}{Xm + Xo}$.

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 English are 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 = $\frac{Xm - Xo}{Xm + Xo} = \frac{1575 - 1380}{1575 + 1380} = \frac{195}{2955} = 0.66$

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 = 1

Range = X_m - X_o = 10 - 1 = 9.

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

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 of the lowest class = 59.5

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

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

Method 2

Here X_m = Mid value of the highest class = 73; X_o = Mid Value of the lowest class = 61

Range = X_m - X_o = 73 - 61 = 12 Kilogram.

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

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 [$\frac{n}{4}$] th item

= Value of [$\frac{60}{4}$] th item

= 15th item

Q_1 lies in the class 10.25 - 10.75

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

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 $[\frac{3\ X\ 60}{4}]$ th item

= 45th item

Q_3 lies in the class 11.25 - 11.75

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

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 = $\frac{Q3-Q1}{Q3+Q1}$

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

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

The mean deviation about the mode is

M.D (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 deviation is 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 deviation to 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 get

Marks 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 = $\frac{\Sigma|X-Median|}{n} = \frac{21}{9}$ = 2.33

M.D from Mode = $\frac{\Sigma f|X-Mode,|}{n} = \frac{23}{9}$ = 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 = $\frac{\Sigma|X-\bar{X}|}{n} = \frac{198.53}{217}$ = 0.915

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

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

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

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

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

For frequency distribution the formulas become

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

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
$S = \sqrt{\frac{\Sigma (X - \overline{X})_2}{n}}$	$S = \sqrt{\frac{\Sigma f (X - \overline{X})^2}{\Sigma f}}$

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
$S = \sqrt{\frac{\Sigma D2}{n} - \left(\frac{\Sigma D}{n}\right)2}$	$S = \sqrt{\frac{\Sigma fD2}{\Sigma f} - \left(\frac{\Sigma fD}{\Sigma f}\right)2}$

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
$S = \sqrt{\frac{\Sigma X2}{n}} - (\frac{\Sigma x}{n})2$	$S = \sqrt{\frac{\Sigma fX2}{\Sigma f} - \left(\frac{\Sigma fX}{\Sigma f}\right)2}$

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 = $\sqrt{\frac{\Sigma fu2}{\Sigma f}} - \left(\frac{\Sigma fu}{\Sigma f}\right)$ 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	$(12 - 7)^2 = 25$
ΣX = 42	**$\Sigma(X - \bar{X})^2 = 70$**

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

S = $\sqrt{\frac{\Sigma (X - \overline{X})_2}{n}}$

S = $\sqrt{\frac{70}{6}} = \sqrt{\frac{35}{3}} = 3.42$

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	**ΣD = 6**	**$\Sigma D^2 = 76$**

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

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

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**

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

$$S = \sqrt{\frac{364}{6}} - \left(\frac{42}{n6}\right)^2$$

$$S = \sqrt{\frac{70}{6}} = \sqrt{\frac{35}{3}} = 3.42$$

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**

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

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

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

Example

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

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**

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

$S = \sqrt{\frac{\Sigma f\ (X - \overline{X})\ 2}{\Sigma f}}$

$S = \sqrt{\frac{400}{100}} = \sqrt{4} = 2$ 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**

$S = \sqrt{\frac{\Sigma fD2}{\Sigma f} - \left(\frac{\Sigma fD}{\Sigma f}\right)^2}$

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

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

$S = \sqrt{\frac{\Sigma fX2}{\Sigma f} - \left(\frac{\Sigma fX}{\Sigma f}\right)^2}$

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

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

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

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

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

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 = $\frac{S}{X}$

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. The coefficient of variation (CV) is defined as

Coefficient of Variation (C.V) = $\frac{S}{X}$ 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 dispersion in different sets of data particularly the data which differ in their means or differ in the units of 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 use very 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 meat are 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	$(12 - 7)^2 = 25$
$\Sigma X = 42$	**$\Sigma(X - \bar{X})^2 = 70$**

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

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

$S = \sqrt{\frac{70}{6}} = \sqrt{\frac{35}{3}} = 3.42$

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

Coefficient of Variation (C.V) = $\frac{S}{\bar{X}}$ X 100 = $\frac{3.42}{7}$ X 100 = 48 .86%

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.

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

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

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 population variance.
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

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

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

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

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**	**$(X - \overline{X})^2$**	**f $(X - \overline{X})^2$**
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**

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

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

Variance $S^2 = 4$.

SKEWNESS AND KURTOSIS

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 of time, the average may remain the same, but the structure of the distribution may change. Two distributions 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 data includes 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:

$$g1=\sum Ni=1(Yi-\bar{Y})3/Ns3$$

where $\bar{Y}$ 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

$$G1=N(N-1)\overline{\qquad\qquad}\sqrt{N-2}\sum Ni=1(Yi-\bar{Y})3/Ns3$$

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

$$\text{Galton skewness}=Q1+Q3-2Q2Q3-Q1$$

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

$$Sk2=3(\bar{Y}-\tilde{Y})s$$

where $\tilde{Y}$ is the sample median.

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

KURTOSIS

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

$$\text{kurtosis}=\sum Ni=1(Yi-\bar{Y})4/Ns4$$

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 is computed using N in the denominator rather than N - 1.

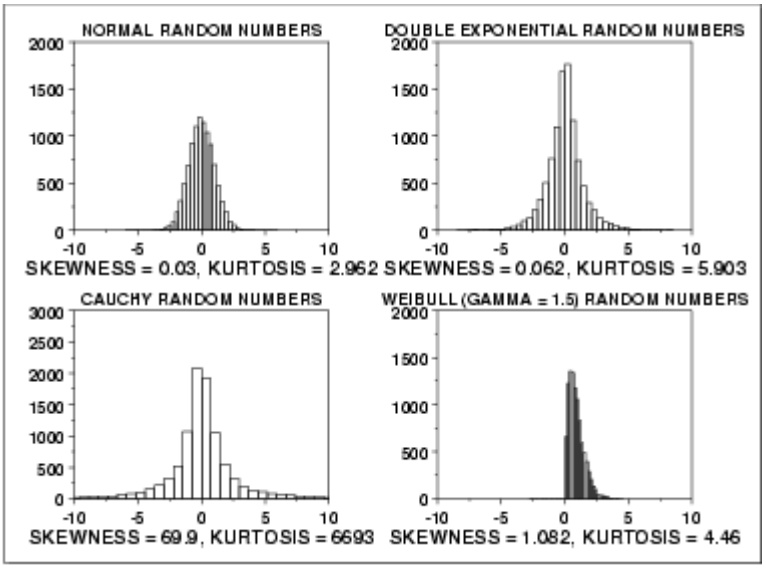
The kurtosis for a standard normal distribution is three. For this reason, some sources use the following definition of kurtosis (often referred to as "excess kurtosis"):

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

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, it has 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 kurtosis is 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 indicated by 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.

EXERCISES

Q1. Calculate the range and quartile deviation for wages: Also calculate coefficient of quartile 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: $\sigma = \sqrt{\sqrt{\frac{\Sigma fd12}{N}} - \left(\frac{\Sigma fd1}{N}\right)^2}$ X C

$= \sqrt{\sqrt{\frac{109}{60}} - \left(\frac{5}{60}\right)^2}$ 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

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

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

Q5. Write short notes on skewness and kurtosis.

UNIT 6

TESTING OF HYPOTHESIS - ONE SAMPLE

Unit 1

SYLLABUS

Introduction to Hypothesis testing, Hypothesis Testing Procedure, Two tail and One tail of Hypothesis, Type I and Type II Errors, Concept of t-test and z-test, Hypothesis testing for Population Proportion.

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. According to Prof. Morris Hamburg, "A Hypothesis in statistics is simply a quantitative statement about a 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, using our 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.

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 fewer than 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 semester is 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 students have 0-8 drinks per week.

EXAMPLE RESEARCH PROBLEM CONTINUED

The authors of "Alcohol Consumption and College Life" wanted to test the hypothesis that the 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 Consumption

Figure 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 calls promptly?

0. No 1. Yes

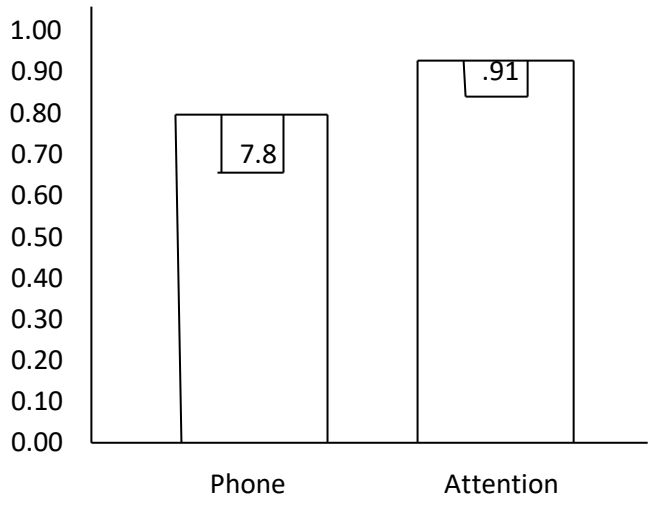
2. Does Mr. Skemp pay adequate attention to your case?

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 results and discuss them in terms of whether or not they are favourable for Mr. Skemp.

Statistician with Mr. Skemp



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 but while check testing light indicator states it is not, it is replaced even though thruster was fine and 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 at landing, 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 the vehicle 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 and stops 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 difference is 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.

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 = $\frac{\overline{x-\mu}}{s}$ of R.A. Fisher later on defined t = $\frac{\overline{x-\mu}}{s}\sqrt{n}$ 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 = $\frac{\overline{x-\mu}}{s}\sqrt{n} = \frac{\overline{x-\mu}}{\sigma s}\sqrt{(n-1)}$ or t = $\frac{\overline{x-\mu}}{s} - \sqrt{(v+1)}$

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 of t is given by

y = $\frac{y0}{1+\frac{t2}{v}v+1/2}$

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 with the 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.

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 correspond to 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).

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

CHAPTER 6

Unit 2
TESTING OF HYPOTHESIS - TWO SAMPLES (Related and Independent)

SYLLABUS: Introduction, Hypothesis testing for difference between Two population means using z-statistic, Hypothesis testing for difference between Two population means using t-statistic, Statistical Inferences about the differences between the Means of Two-related Populations, Hypothesis testing for the difference in Two Population Proportions.

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 samples are 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 made while 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

S. E. $\bar{X} = \frac{\sigma p}{\sqrt{n}}$

where 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

S. E. $\bar{X} = \frac{\sigma(sample)}{\sqrt{n}}$

where σ denotes standard deviation of the sample.

It should be noted that if standard deviation of both sample as well as population are available then 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

$$\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

99% fiducial limits of population mean are

$$\bar{X} \pm 2.58\frac{\sigma}{\sqrt{n}}$$

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 is calculated:

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. $\bar{X} = \frac{\sigma}{\sqrt{n}}$

Calculation of Standard Deviation

Mid-value m	**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

$\sigma = \frac{\sqrt{\Sigma\ fd\prime 2}}{N}$ - $(\frac{\Sigma\ fd\prime}{N})^2$ X C = $\frac{\sqrt{236}}{100}$ - $(\frac{80}{100})^2$ X 100 = $\sqrt{2.36} -$ 0.64 X 10 = 1.311 X 10 = 13.11

S.E. $\bar{X} = \frac{13.11}{\sqrt{100}} = \frac{13.11}{10}$ = 1.311.

STANDARD ERROR OF THE DIFFERENCE BETWEEN THE MEANS OF TWO SAMPLES

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

$= \sqrt{\sigma 2 \frac{1}{n1+n2}}$

If σ is unknown, sample standard deviation for combined samples must be substituted.

ii. If two random samples with $\overline{X1}$, σ_1, n_1 and $\overline{X2}$, 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 obtained by boys and girls.

S.E.$(\overline{X1} - \overline{X2}) = \sqrt{\frac{\sigma 12}{n1} + \frac{\sigma 22}{n2}}$ where $\sigma_1 = 15$, $\sigma_2 = 20$, $n_1 = 150$ and $n_2 = 250$

Substituting the values

S.E.$(\overline{X1} - \overline{X2}) = \sqrt{\frac{(15)2}{150} + \frac{(20)2}{250}} = \sqrt{1.5} + 1.6 = 1.781$

$\frac{Difference}{S.E} = \frac{75-70}{1.781} = 2.84$

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.$(\overline{X1} - \overline{X2}) = \sqrt{\frac{\sigma 12}{n1} + \frac{\sigma 22}{n2}}$ where $\sigma_1 = 10$, $\sigma_2 = 12$, $n_1 = 121$ and $n_2 = 81$
Substituting the values
S.E.$(\overline{X1} - \overline{X2}) = \sqrt{\frac{(10)2}{121} + \frac{(12)2}{81}} = \sqrt{100/121 + 144/81} = \sqrt{2.604} = 1.61$
Difference of means (84 - 81) = 3
$\frac{Difference}{S.E} = \frac{3}{161} = 1.86$
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 $(\sigma_1 - \sigma_2) = \sqrt{\frac{\sigma 12 + \sigma 22}{2n1 + 2n2}}$ 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: $H_a: \mu_1 - \mu_2 \neq 0$ $z = \dfrac{28-33-0}{\sqrt{\dfrac{14.1^2}{75}+\dfrac{9.5^2}{50}}} = \dfrac{-5}{\sqrt{2.65+1.81}} = -2.37$

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— $\bar{x}_1$ could just as well have been the female sample mean and $\bar{x}_2$ 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 and the *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 the corresponding 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 *each population*)

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?

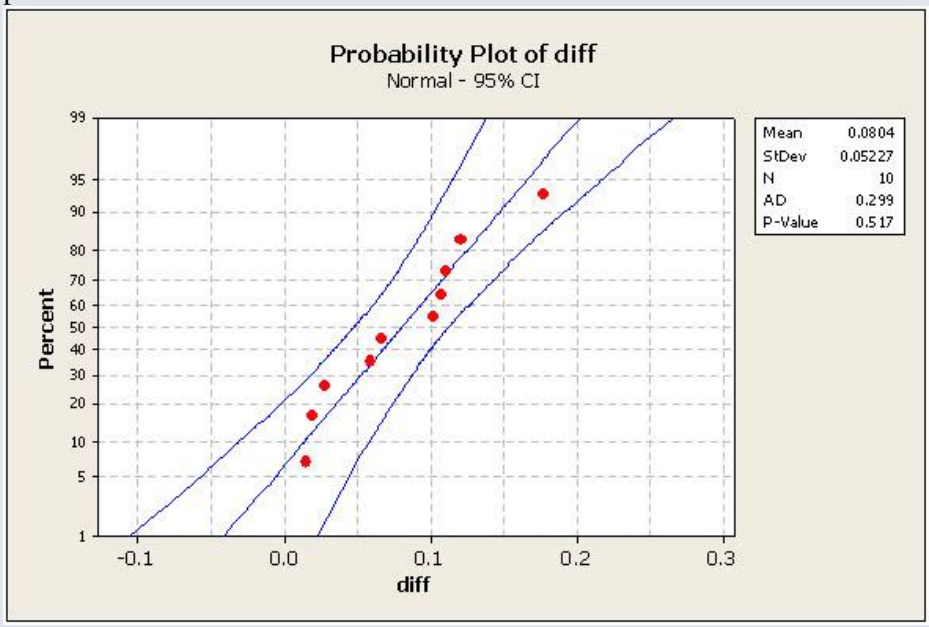
	Location									
	1	2	3	4	5	6	7	8	9	10
Zinc concentration in bottom water	.430	.266	.567	.531	.707	.716	.651	.589	.469	.723
Zinc concentration in surface water	.415	.238	.390	.410	.605	.609	.632	.523	.411	.612

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:

$$t_* = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{0.0804}{0.052310/\sqrt{4.86}}$$

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 hypothesis you 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 nonparametric 1-sample test on the difference.

HYPOTHESIS TESTING OF THE DIFFERENCE BETWEEN TWO POPULATION MEANS

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. $H_0 : \mu_1 = \mu_2$ $\qquad$ $H_A : \mu_1 \neq \mu_2$

2. $H_0 : \mu_1 \geq \mu_2$ $\qquad$ $H_A : \mu_1 < \mu_2$

3. $H_0 : \mu_1 \leq \mu_2$ $\qquad$ $H_A : \mu_1 > \mu_2$

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 and normal individuals?

(1) Data

$\bar{x}_1 = 4.5 \quad n_1 = 12 \quad \sigma_1^2 = 1$

$\bar{x}_2 = 3.4 \quad n_2 = 15 \quad \sigma_2^2 = 1.5$

$\alpha = .05$

(2) Assumptions

- two independent random samples

- each drawn from a normally distributed population

(3) Hypotheses

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

(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

$$H_0 : \mu_1 \leq \mu_2$$

$$H_A : \mu_1 > \mu_2$$

(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 values
t = 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 BHADP test than do the nondisabled persons.

Actual values
z = 8.42
p = 1.91 x 10-17

Paired comparisons

Sometimes data comes from nonindependent samples. An example might be testing "before and 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 is effective 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. On the 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
α = .05

(2) Assumptions

- the observed differences are a simple random sample from a normally distributed population of differences

(3) Hypotheses

H_0: $\mu_d \geq 0$

H_A: $\mu_d < 0$ (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 statistically independent 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 hypothesis must 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.