Block 3

Statistical Analysis

ignou
THE PEOPLE'S
UNIVERSITY

BLOCK 3 INTRODUCTION

Block 3 "Statistical Analysis" deals with the measures of central tendency and dispersion, sampling distribution and statistical analysis.

Unit 7 "Descriptive Statistics-I" deals with different types of measures of central tendency and measures of dispersion.

Unit 8 "Descriptive Statistics-II" deals with different types of correlations, regression, regression analysis and regression coefficients.

Unit 9 "Sampling Distributions" deals with sampling distribution, the concept of standard error and the Central Limit Theorem.

Unit 10 "Statistical Analysis-I" deals with the concept of hypothesis, the procedure of testing a hypothesis, and also the test of hypothesis for large samples for the population mean and the difference between two population means.

Unit 11 "Statistical Analysis-II" deals with the small sample test procedures for a single population mean and the difference between two population means based on t-distribution. Also, the unit delves into the chi-square test for the goodness of fit and independence of attributes. The test for the equality of two variances using F- distribution is also discussed in the unit.

Unit 12 "Analysis of Variance" delves into the one-way as well as two-way Analysis of Variance.

Objectives

After studying this block, you should be able to:

- explain the different types of measures of central tendency;
- explain the measures of dispersion;
- explain the concept of correlation between two variables and identify different types of correlations;
- explain regression, regression analysis and linear regression;
- explain the salient properties of regression coefficients;
- explain the concept of sampling and sampling distribution;
- explain the concept of standard error and the Central Limit Theorem;
- describe the sampling distribution of the sample mean and the difference between the two sample means;
- describe the procedure of testing a hypothesis;
- describe the small sample test procedures for a single population mean and the difference between two population means based on t-distribution;

- explain the chi-square test for the goodness of fit and independence of attributes;
- describe the test for the equality of two variances using F-distribution; and
- explain the Analysis of Variance technique;

We hope that after studying this block you will acquire an understanding of statistical analysis.

Wishing you success in this endeavour!

UNIT 7 DESCRIPTIVE STATISTICS-I

Structure

7.1 INTRODUCTION

A frequency distribution for a given set of numerical data and different graphs using the raw data in hand; which, in fact, is a method of summarization of data for having a rough or approximate idea about the nature, properties and pattern of variation among the values of the variable. But what, if we need to observe whether there is some tendency in the data set to concentrate around a single numerical figure or, in other words, what would be the appropriate single representative value of the variable, representing most of the part of the data set? Obviously, then we have to find such a numerical figure based upon all the values of the characteristic and not only with only one or few of the values, in order to make it the best

representative figure. The choice of such a single number or summary statistics that we choose to summarize a given data depends upon the particular characteristics we want to describe. The statistical measures which describe such characteristics are called measures of central tendency or measures of location.

In this unit, we are going to discuss about different measures of central tendency. Measures of central tendency in Statistics play a very important and powerful role; since, such a measure reflects most of the features of the entire mass of data as regards to its tendency to concentrate around a single value. Generally, any measure of central tendency is called an "average". Average of a set of data is actually found to locate the value (or point of location of it) to which the average value coincides in the range of values of the characteristic.

Different measures of central tendency provide us an idea about the central value around which most of the values of the data set are concentrated. However, being values of a variable and not of a constant, all of them are not of the same magnitudes; rather they are scattered or dispersed over a given range of values. This scatteredness of variate-values is a peculiar and natural characteristic of all the variables; whether they may be less scattered or may be highly scattered. Thus, even if central values of two different sets of data are same, the scatteredness will be different of all the data sets. In this context, we should have different measures of the second property, namely, property of scatteredness or spreadness of data; or what is popularly called, "property of dispersion".

Section 7.3 highlights the meaning of measure of central tendency along with its significance for the preliminary analysis of the data, important properties which a good measure (average) is supposed to be satisfied. Section 7.4 describes different types of arithmetic mean which can be computed over data sets which are tabulated in different manners along with the merits and demerits of it. Section 7.5 describes the methods of computing median from data sets which are available in different forms of tabular representation. The merits and demerits of it have also been mentioned in the section. Section 7.6 describes the methods of computing it from data sets which are available in different forms of tabular representation. The merits and demerits of it have also been mentioned in the section. Section 7.7 will present a discussion on different measures of dispersion, like, range, mean deviation, standard deviation, etc. Also, we shall highlight the essential properties of such measures which are desirable for a good measure of dispersion. Section 7.8 deals with the simplest kind of measure of dispersion: Range, its calculation and merits and demerits. Section 7.9 is devoted to the description of Mean Deviation as a measure of dispersion along with its computation and its merits and demerits. Section 7.10 describes Standard Deviation and Variance as very important and most popular measures. The methods of computing of both are explained using different kinds of data.

7.2 OBJECTIVES

After studying this unit, you should be able to

- explain the concept of central tendency;
- define an average;
- explain the significance of a measures of central tendency;
- explain the properties of a good average;
- know the different types of measures of central tendency;
- calculate arithmetic mean, median and mode for different types of classification of data; and
- calculate range, mean deviation and standard deviation for different types of data.

7.3 MEASURES OF CENTRAL TENDENCY

It is a bare fact that, whatever be the characteristic under study and the number of values of the characteristic measured; there is always a tendency of the data set, so obtained, to concentrate or cluster around a particular value, which is termed as "central value". This property of the data tending towards the central value is known as "property of central tendency". For example, if the data on age of primary school children is considered, most of the ages would be around 5 to 6 years, whereas, most of the ages of undergraduate students would be around 17-18 years. Thus, central value in the first data set is 5 years and for second data set is about 18 years.

Central values of data sets are generally those figures which might be considered to be representative figures to describe the 'property of central tendency' of the data sets. In this sense, these are also known as "measures of Central tendency" or more popularly the "average".

According to **Professor Bowley**, averages are "statistical constants which enable us to comprehend in a single effort the significance of the whole". They throw light as to how the values are concentrated in the central part of the distribution. In other words, an average is a single value which is considered as the most representative for a given set of data. Measures of central tendency show the tendency of some central value around which data tend to cluster.

7.3.1 Significance of the Measure of Central Tendency

There are two main reasons for studying an average.

1. **To get a single representative**

   A measure of central tendency provides us the idea of that value in the entire range of the data, to which most of the values of the data set

tends to cluster and, thus, provides a representative figure for the mass of data. It is easy to remember this representative figure of any data set whereas all the values of the set are impossible to remember. For example, in a class of 50 students, it is not possible to remember all the heights, but the average height of these students, being a single number, is easy to remember.

2. **To facilitate comparison**

   Measures of central tendency enable us to compare two or more populations by deducing the mass of data in one single figure. The comparison can be made either at the same time or over a period of time. For example, if a subject has been taught in more than two classes, obtaining the average marks of students of those classes, comparison of average performance of students over classes can be made.

7.3.2 Properties of a Good Average

Since, an average or a measure of central tendency is obtained for any data set and is considered to be a true representative of all the values of a data set; it is expected to possess some properties so that one can choose the best measure depending upon the nature of the data in hand. The following are the properties of a good measure of an average:

1. **It should be simple to understand**

   Since we use the measures of central tendency to simplify the complexity of a data so an average should be understandable easily otherwise its use is bound to be very limited.

2. **It should be easy to calculate**

   An average not only should be easy to understand but also should be simple to compute so that it can be used as widely as possible.

3. **It should be rigidly defined**

   A measure of central tendency should be defined properly without any ambiguity. It should have a rigid definition as well as an appropriate interpretation. It should also have an algebraic formula so that if different people compute the average from same figures, they get the same answer.

4. **It should be liable for algebraic manipulations**

   A measure of central tendency should be liable for the algebraic manipulations, that is, it must have some good properties which widen its application in different situations. For example, if there are two or more sets of data and the individual information are ~~is~~ available for these sets, the formula of the concerned measure can be manipulated in such a way that the formula of same measure can be obtained for the combined set.

5. It should be least affected by sampling fluctuations

A measure of central tendency (an average) must be robust (almost stable) in the sense that it should not be affected too much if there is a large variation in samples selected from a particular population. This means that if we select 10 different groups of observations from same population and compute the average of each group, then we should expect to get approximately the same values. However, whatever little difference between the averages values appear must be because of the sampling fluctuation only.

6. It should be based on all the observations

If any measure of central tendency is used to analyze the data, it is desirable that the measure is based upon each and every observation for its calculation.

7. It should be possible to calculate even for open-end class intervals

If a measure of central tendency is computed for a grouped data, that is, for a data set arranged in the form of a frequency distribution, then, if possible, it should be in such a form that it can be calculated for open end classes also.

8. It should not be affected by extremely small or extremely large observations

As described in the property 6 above, an average must be dependent on all the observations of the data and, hence, each and every observation influences the value of the average. As a representative figure, therefore, an average should not be highly affected by some of the very large and very small values which generally occur in any set of data.

7.3.3 Different Measures of Central Tendency

As we described earlier, generally every set of data has a central value around which the values of the variable tend to cluster. If we wish to know the central value in a set of data, some measures of central tendency (also known as measures of location) are computed. These measures are popularly known as "average".

In Statistics, some of the measures of central tendency which are very much popular are as follows:

1. Arithmetic Mean or simply Mean;
2. Median;
3. Mode;
4. Geometric Mean;
5. Harmonic Mean.

We shall discuss Arithmetic Mean, Median and Mode only out of these measures one at a time in next sections.

7.4 ARITHMETIC MEAN

"Arithmetic mean" (also called "mean") of any set of data is defined as the sum of all the observations of the data set, divided by the total number of observations in the set. It is the most popular and widely used measure of central tendency.

As per above definition of mean, mathematically it can be written as

$$\text{Mean} = \frac{\text{Sum of all the observations of the data set}}{\text{Number of observations in the set}} \quad \ldots (7.1)$$

This is the only basic formula for computation of mean. However, you will come across with some other formulae for computing mean in further text; but these formulae in no way provide any new method of computation of mean; rather all these formulae are, in fact, reducible to this basic formula on simplification.

We have seen that in Statistics, generally the raw data are summarized in a number of forms, which are

1. Individual series or "ungrouped data";
2. Discrete frequency distribution and
3. Continuous frequency distribution.

You might be knowing that the data summarized in the forms 2 and 3 are popularly termed as "grouped data". Accordingly, you will see below that the basic formula for the computation of mean, which is used for ungrouped data, is presented in a different form for the grouped data.

7.4.1 Calculation of Simple Arithmetic Mean

(I) For Ungrouped Data

(a) Direct Method

If individual values of a variable are given in the form of individual series, that is, the data are not grouped into classes; the basic formula is directly applicable for the computation of mean.

Thus, for a given set of data with N values $x_1, x_2, \ldots, x_N$; the mean would be

$$\bar{X} = \frac{x_1 + x_2 + x_3 + \ldots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}. \quad \ldots (7.2)$$

where $\bar{X}$ stands for the arithmetic mean or mean of the set; X being the characteristic (variable) under study.

(b) Short-Cut Method

Under this method, we arbitrarily choose two constants, say A and h; A is used to change the origin of the values and h for changing the scale of values.

The constant A is frequently called "assumed mean". This technique is very much used in daily life.

If $x_1, x_2, x_3, ..., x_N$ are the N observations for the variable X and let A and h be the chosen constants, where A be the assumed mean, then let deviations of x -values from A respectively be $d_1, d_2, d_3, ..., d_N$ where $d_i = (x_i - A),\ i = 1, 2, \ldots, N.$

Further, let $u_i = \frac{d_i}{h}$, where u_1, u_2, …, u_N are N values of the transformed variable U. Then, we have

$$\bar{X} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{\sum_{i=1}^{N} (A + d_i)}{N} = \frac{\sum_{i=1}^{N} A + h\sum_{i=1}^{N} u_i}{N}$$

$$= \frac{NA}{N} + h\frac{\sum_{i=1}^{N} u_i}{N} = A + h\bar{U};$$

where, $\bar{U} = \frac{\sum_{i=1}^{N} u_i}{N}$ being the mean of the transformed variable U.

Thus, the relation between the mean of the original variable X and the mean of the transformed variable U is obtained as

$$\bar{X} = A + h\bar{U}. \qquad \ldots (7.3)$$

We see that the mean is affected by change of origin and scale both, since the relation obtained above, involves both the constants, A, used to change the origin and h, the constant used for changing the scale.

Example 1: Calculate the mean of runs scored, as given below, by ten players in 20-20 cricket match:

Player	1	2	3	4	5	6	7	8	9	10
Score	15	20	25	19	12	11	13	17	18	20

Solution: We have been given the scores of 10 players in a 20-20 cricket match. Therefore, we have N = 10.

Now for calculating the arithmetic mean or average score of 10 players, we need to follow the following steps:

1. Calculate the sum, $\sum x$, by adding the scores of 10 players
2. Divide the total by N, the number of observations,

In symbols:

$$\bar{X} = \frac{15 + 20 + 25 + 19 + 12 + 11 + 13 + 17 + 18 + 20}{10} = \frac{170}{10} = 17.$$

Thus, the mean score of the 10 players is 17 runs.

Example 2: Use short-cut method to calculate arithmetic mean for the data given in Example 1.

Solution: We follow the following steps for calculating the arithmetic mean or average score of the 10 players using short-cut method as:

1. Decide the value of the assumed mean A and the constant h.
2. Calculate the values $u_i = \frac{x_i - A}{h}$ (i=1, 2, … , 10) of the transformed variable U.
3. Obtain the sum, $\sum_{i=1}^{10} u_i$ and hence, the value of $\bar{U}$. Use the formula (1.3) to get the mean $\bar{X}$.

In this series of observations, since the values range from 11 to 25; therefore, 18, a neat round value in the middle of 11 and 25, may be taken as assumed mean, that is we take A =18. Further, let h = 1.

With these values of A and h, we have the following table:

Players	Scores (X)	U = (X – 18)/1
1	15	–3
2	20	+2
3	25	+7
4	19	+1
5	12	–6
6	11	–7
7	13	–5
8	17	–1
9	18	0
10	20	+2
		$\sum_{i=1}^{10} u_i = -10$

Using formula (7.3), we get

$$\bar{X} = A + h\bar{U} = A + h \times \frac{\sum_i u_i}{N} = 18 + 1 \times \frac{-10}{10} = 17 \text{ runs.}$$

It can be observed in the example that we have taken h = 1, since, the transformed values are one-digit numbers, which are easy to tackle in the computation process. This is the wise choice of the constant h; otherwise, we take h = 10 or 100, if the transformed values would be two-digit or three-digit figures, which are not so easy to tackle.

(II) For Discrete Frequency Distribution

By a discrete frequency distribution, we mean a grouped data where all the values of the concerned variable are given separately associated with the number of times they occur in the data (you know that the number of times a value occur in the data is called the "frequency" of the value). Thus, in fact, it is also a form of the grouped data. For example, suppose that the k distinct values in the data are $x_1, x_2, x_3, ..., x_k$ with associated frequencies $f_1, f_2, f_3, ..., f_k$ respectively.

Since the value x_i appears f_i times (i = 1, 2, .., k) in the data, the sum of all observations belonging to the set is equal to

$$\sum_{i=1}^{k} x_i = \underbrace{(x_1 + x_1 + \ldots + x_1)}_{f_1 \text{ times}} + \underbrace{(x_2 + x_2 + \ldots + x_2)}_{f_2 \text{ times}} + \ldots + \underbrace{(x_k + x_k + \ldots + x_k)}_{f_k \text{ times}}$$

$$\sum_{i=1}^{k} x_i = f_1 x_1 + f_2 x_2 + \ldots + f_k x_k = \sum_{i=1}^{k} f_i x_i.$$

The total number of observations in the set is obviously given by

$$f_1 + f_2 + \ldots + f_k = \sum_{i=1}^{k} f_i.$$

Therefore, using the basic formula (5.1), we have the formula for computing mean of a discrete frequency distribution as

$$\text{Mean} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}. \qquad \ldots (7.4)$$

Both the methods, namely, direct method and short-cut method for computing the mean are also applicable in this case without any new concept. Therefore, without adding any new thing here, we present Examples 3 and 4 below to illustrate how these methods can be applied for computing mean of a discrete frequency distribution.

(a) Direct Method

Let us consider the following discrete frequency distribution for illustration:

Example 3: The following table gives the data relating to the marks (out of 10) of 100 students in a statistics test:

Score(X)	0	1	2	3	4	5	6	7	8	9	**Total**
Frequency (f)	1	2	4	9	15	21	20	15	9	4	100

Find the average marks of the students:

Solutions: We can use directly the formula (1.4) for the purpose. For this, we prepare the following table for finding the numerator and denominator of the formula:

Score (X)	**Frequency (f)**	**fX**
0	1	0
1	2	2
2	4	8
3	9	27
4	15	60
5	21	105
6	20	120
7	15	105
8	9	72
9	4	36
Total	100	535

Thus, by using the formula we get

$$\bar{X} = \frac{\sum f_i x_i}{\sum f_i} = \frac{535}{100} = 5.35\,;$$

which provides the average marks of the students in statistics test.

(b) Short-cut Method

The short-cut method of calculating the mean, as described in sub-section 7.4.1 (Ib) can also be applied in this case with the same aim to reduce the complexity of the computation process. We may similarly choose appropriate constants A and h, where A is the assumed mean. Defining the transformed variable U through the transformation $U = \frac{(X - A)}{h}$, we can find the u_i values for each x_i values and hence, we can find the mean $\bar{U}$ of the transformed variable in order to apply the formula (1.3).

Let us describe the steps which we follow in short-cut method while applying it to a discrete frequency distribution.

Step I: Determine the appropriate values of assumed mean A and the constant h.

Step II: Find the values $u_i = \frac{(x_i - A)}{h}$ corresponding to each x_i.

Step III: Find the values $f_i u_i$ for each i and hence, find their sum, $\sum_i f_i u_i$.

Step IV: Find the mean of the transformed variable U, using formula $\bar{U} = \frac{\sum_i f_i u_i}{\sum_i f_i}$.

Step V: Use the formula $\bar{X} = A + h\bar{U}$ for computing the mean of the variable X.

Now let us solve the following exercises for illustrating the calculation process:

Example 4: Following table gives the wages paid to 125 workers in a factory. Calculate the arithmetic mean of ways for the data by using short-cut method.

Wage (in Rs)	200	210	220	230	240	250	260
No. of Workers	5	15	32	42	15	12	4

Solution: In this case, assume mean A is taken as 230 because it is neatly at the centre of the range of the variable X. Let us take h = 1.

The other needed calculations are shown in the following table:

Wages (X)	No. of Workers (f)	Deviations U = (X – A)/h	fU
200	5	–30	–150
210	15	–20	–300
220	32	–10	–320
230	42	0	0
240	15	10	150
250	12	20	240
260	4	30	120
	125		$\sum_i f_i u_i = -260$

Using the formula (1.3), the arithmetic mean for the variable X will be

$$\bar{X} = A + h\bar{U} = A + h \times \frac{\sum_i f_i u_i}{\sum_i f_i} = 230 + \frac{-260}{125} \times 1 = 230 - 2.08 = 227.92.$$

(III) For Continuous Frequency Distribution

We can compute the mean of a continuous frequency distribution either using the direct method or by using the short-cut method. The only thing in both the methods is that, given the class intervals and corresponding class frequencies; we obtain first the mid values of each class intervals and consider these mid values as the k values $x_1, x_2, \ldots, x_k$ of the variable X and the given class frequencies $f_1, f_2, \ldots, f_k$ as frequencies associated with the respective values $x_1, x_2, \ldots, x_k$.

In order to illustrate the procedure of calculation, we consider the following two examples:

(a) Direct Method

Let us consider the following example:

Example 5: Using direct method, calculate arithmetic mean for the following data:

Class	20-25	25-30	30-35	35-40	40-45	45-50	50-55
Frequency	8	10	12	20	11	4	5

Solution: For computing arithmetic mean for the given continuous frequency distribution, we first obtain mid values (X) for all the classes. After determining these mid-values, formula (1.4) is used for calculating $\bar{X}$. The calculations needed are shown in the following table:

Classes	Mid-values (X)	Frequency(f)	fX
20-25	22.5	8	180.0
25-30	27.5	10	275.0
30-35	32.5	12	390.0

35-40	37.5	20	750.0
40-45	42.5	11	467.5
45-50	47.5	4	190.0
50-55	52.5	5	262.5
		$\sum_i f_i = 70$	$\sum f_i x_i = 2515$

Using formula (1.4), the Arithmetic Mean will be

$$\bar{X} = \frac{\sum f_i x_i}{\sum f_i} = \frac{2515}{70} = 35.93.$$

(b) Short-Cut Method

Considering the mid values of the given class intervals as the values X_i, follow the process as shown in the Example 4 with suitable choices of A and h.

Consider the following example:

Example 6: Calculate arithmetic mean for the data given in Example 5, using short-cut method.

Solution: In this case, let the assumed mean A = 37.5 and h = 1. Rests of the calculations are presented in the following table:

Classes	**Mid Values (X)**	**Frequencies (f)**	**Deviations U= (X_i – A)/h**	**fU**
20-25	22.5	8	–15	–120
25-30	27.5	10	–10	–100
30-35	32.5	12	–5	–60
35-40	37.5 (A)	20	0	0
40-45	42.5	11	5	55
45-50	47.5	4	10	40
50-55	52.5	5	15	75
		$\sum f_i = 70$		$\sum_i f_i u_i = -110$

Using the formula (7.3), the arithmetic mean can be obtained as

$$\bar{X} = A + h\bar{U} = A + h \times \frac{\sum_i f_i u_i}{\sum_i f_i} = 37.5 + \frac{-110}{70} = 37.5 - 1.57 = 35.93.$$

7.4.2 Combined Arithmetic Mean

If the arithmetic means and the number of observations of two groups are known, we can calculate the combined arithmetic mean of these groups. The combined mean formula for two groups is as under

$$\bar{X}_{12} = \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2}{N_1 + N_2}$$

Here, $\bar{X}_{12}$ = combined mean of two groups;

(N_1, $\bar{X}_1$) : Number of observations and mean of the first group;

(N_2, $\bar{X}_2$) : Number of observations and mean of the second group;

Similarly for the three groups, the formula becomes

$$\bar{X}_{123} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2 + N_3\bar{X}_3}{N_1 + N_2 + N_3}.$$

The formula can also be extended for k – groups

$$\bar{X}_{12...k} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2 + ... + N_k\bar{X}_k}{N_1 + N_2 + ... + N_k}.$$

Let us consider the following examples:

Example 7: The mean marks of the 120 students in section A is 80 and mean marks of 80 students in section B is 90. Find the combined mean of both the sections.

Solutions: Here $N_1 = 120, \quad N_2 = 80, \quad \bar{X}_1 = 80$ and $\bar{X}_2 = 90.$

Using formula, the combined mean of all the 200 students will be

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2} = \frac{120 \times 80 + 80 \times 90}{120 + 80}$$

$$\Rightarrow \bar{X}_{12} = \frac{9600 + 7200}{200} = \frac{16800}{200} = 84 \text{ marks}$$

CHECK YOUR PROGRESS 1

Note: i) Check your answers with those given at the end of the unit.

1) Find the arithmetic mean of the observations 5, 8, 12, 15, 20, 30.

2) For the following discrete frequency distribution, find arithmetic mean.

Wages (in Rs)	20	25	30	35	40
No. of workers	4	8	20	12	6

3) Find arithmetic mean of the distribution of marks given below.

Marks	0-10	10-20	20-30	30-40	40-50
No. of students	6	9	17	10	8

7.5 MEDIAN

Generally, in a raw data set, values of the study variable are given in a haphazard way, that is, in the order in which values are collected in the experiment or survey and, therefore, it does not follow any rule of arrangement. However, for any statistical analysis, it is sometimes advisable to arrange the data in the form of an 'array', that is, arranging the values either in an ascending or descending order of magnitude. In an array, every observation holds a certain rank. For example, when arranged in ascending order of magnitude, the very first value is the smallest one, the second value is the second smallest value, and so on; but when arranged in the descending order, the very first value is the largest value, the second value is the second largest value and so on. Thus, it is the way of arrangement which changes the rank of values in the series.

"Median" of a set of data is a positional value. The Median is defined as the middle-most value of the variable, when values are arranged either in ascending or descending order of magnitude. According to Connor, "The median is the value of the variable which divides the group or series into two equal parts, one part comprising all values greater and the other part, value lesser than the Median". Since the median denotes the central position in a series of values, it is called a position average.

7.5.1 Finding the Median for a Set of Data

(I) For Ungrouped Data

Let N values of a variable X be given in the form of raw data, that is, each individual value is given separately in the order they are recorded. In order to find the median of the series of values, we shall consider the following two cases:

Case I: When there are odd number of values

Let N be an odd number, that is, we can represent N as $N = 2k + 1$, where $k = 0, 1, 2, \ldots$. Then, let all the N values be arranged in the ascending order of magnitude and, thus, we have the values as

$$x_1 \leq x_2 \leq x_3 \leq \ldots \leq x_k \leq x_{k+1} \leq x_{k+2} \leq x_{k+3} \leq \ldots \leq x_{2k+1}.$$

Clearly, in this series the middle-most value is x_{k+1}, and there are equal number of values on both the sides of it. According to the definition of median, therefore, x_{k+1} is the median of the series. Thus, if there are 2k+1 values (an odd number of values) in the data, the median is the value which occupies the $(k+1)^{th}$ place or, equivalently, $[(N + 1)/2]^{th}$ position, when values are arranged either in ascending or descending order of magnitude.

Example 8: Let us consider the following set of raw data:

5, 12, 45, 8, 21, 32, 5, 36, 16, 8, 20, 24, 54, 18, 44, 56, 20, 30, 44

Solution: Here, N = 19 = 2x9 +1, means k = 9.

Arranging these values in ascending order of magnitude, we get the series as:

5, 5, 8, 8, 12, 16, 18, 20, 20, 21, 24, 30, 32, 36, 44, 44, 45, 54, 56

Then, since the value 21 occupies the middle-most position (that is, 10^{th} position), it is the median of the set.

Case II: When there are even number of values

Let there be an even number of observations. Then, N can be written as N = 2k for k = 1, 2, 3, … . If these values are arranged in ascending order of magnitude as

$$x_1 \leq x_2 \leq x_3 \leq \ldots \leq x_k \leq x_{k+1} \leq x_{k+2} \leq x_{k+3} \leq \ldots \leq x_{2k}.$$

In such situations, we observe that there is no single middle-most value in the series, rather, we will get two middle-most values which would be x_k and x_{k+1} occupying the $(N/2)^{th}$ and $[(N/2)+1]^{th}$ positions. Following the definition of the median, therefore, we consider neither of the values as the median, but we take the arithmetic mean of both the middle-most values as the median of the series. In other words, in such situations the median of the series, denoted by M_d, will be given by

$$M_d = \frac{\left(\frac{N}{2}\right)^{th} \text{value} + \left(\frac{N}{2}+1\right)^{th} \text{value}}{2}$$

Example 9: Calculate the median for the following data:

7, 8, 9, 3, 4, 10

Solution: First we arrange given data in ascending order as 3, 4, 7, 8, 9, 10

Here, (N) = 6 = 2x3 (even). So, here k = 3. Therefore, we get the median as

$$M_d = \frac{3^{rd}\text{observation} + 4^{th}\text{observation}}{2} = \frac{7+8}{2} = \frac{15}{2} = 7.5.$$

You can see that in such cases, median does not correspond to any of the values of the series, rather, it is the mean of two middle-most consecutive values of the series.

(II) For Discrete Frequency Distribution

In case of discrete frequency distribution, the procedure of determining median is fundamentally the same as that for an individual series. The procedure consists of the following steps:

(i) Let the number of distinct values of the variable be k. Arrange these values in ascending or descending order of magnitude, each associated with their corresponding frequency. Let f_i (i = 1, 2, …, k) denotes the frequency associated with the i^{th} value. Further, we have $f_0 = 0$ and $f_{k+1}=0$;

(ii) Obtain the "less than type" cumulative frequencies $F_1 = f_1$, $F_2 = f_1+f_2$, … , $F_k = f_1+f_2+\ldots+f_k$ corresponding to each value, and list them in a separate column. We have, $F_0 = 0$ and $F_{k+1} = F_k$. The last cumulative frequency will coincide with the total number of values, that is, N; Now, as before, we shall consider two cases: N is odd and N is even.

(iii) If **N is odd**, then, find the value of [(N +1)/2]. If the inequality $F_{i-1} \leq (N+1)/2 \leq F_i$ is satisfied for the i^{th} value (i = 1, 2, …, k), then the i^{th} value will be the median of the data.

(iv) If **N is even**, then we know that there will be two middle-most values, the mean of which would be the median. Find the values of (N/2) and [(N/2)+1]. The first middle-most value (i^{th} value) will be obtained using the inequality

$F_{i-1} < \left(\frac{N}{2}\right) \leq F_i$. Now, the second middle-most value (j^{th} value) will be obtained by using the inequality $F_{j-1} < [\left(\frac{N}{2}\right) + 1] \leq F_j$, where j = 1, 2, … , k.

Note: Since, for the case of N even, there would be two middle most values, therefore, care must be taken in finding the second middle-most value $(i+1)^{th}$ value after finding the first middle-most value (the i^{th} value). It may happen that both the j^{th} and i^{th} values are the same value, that is, j = i or j= i + 1.

Example 10: Find the median size of the following series:

Size (X)	4	5	6	7	8	9	10
Frequency (f)	06	12	15	28	20	14	05

Solution: Values are already arranged in ascending order of magnitude, therefore, we compute less than cumulative frequencies as shown in the table:

Size X	**f**	**Cumulative Frequency (F_i)**
4	6	6
5	12	18
6	15	33
7	28	61
8	20	81
9	14	95
10	5	100
	100	

Here, N is even, so we compute (N/2) and [(N/2) +1], which are 50 and 51. The first value (N/2) satisfies the condition $F_{i-1} < \left(\frac{N}{2}\right) \leq F_i$ for $F_3 < \left(\frac{N}{2}\right) \leq F_4$; F_3 and F_4 are respectively, 33 and 61. Also, we see that the value [(N/2) + 1] satisfies the condition for $F_3 < \left(\frac{N}{2}\right) \leq F_4$. This implies that both (N/2) and [(N/2) +1], correspond to the value 7. There mean will also be 7. Thus, median is 7.

(III) For Continuous Frequency Distribution

In the case of continuous frequency distribution, we first locate the median class, that is, the particular class which contains the exact median which would have been obtained if the data was given in the form of ungrouped series of values. Obviously, after locating the median class, finally the approximate value of median within the class is determined by using an interpolation formula. The procedure of computation involves the following steps:

(i) Compute less than type cumulative frequency for each of classes given. Let there be k classes in all.

(ii) Find the value (N/2).

(iii) Now, find the class for which the condition $F_{i-1} < (N/2) \leq F_i$; where F_i is the cumulative frequency corresponding to i^{th} class (i =1, 2, …, k), is satisfied. Then, the i^{th} class is said to be "Median Class".

(iv) Obtain the median value by applying the formula;

$$\text{Median} = L + \frac{\frac{N}{2} - C}{f} \times h$$

where, L = Lower class limit of the Median class

C = cumulative frequency of preceding class of the median class

f = frequency of the median class

h = class interval (class width) of the median class.

Example 11: 100 salesmen employed by a company have booked the following number of orders for a newly introduced FAX machine during the last six months:

No. of Orders Booked	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of Salesmen:	4	12	25	30	15	8	6

Calculate median of the above data and account for the difference, if any.

Solution: We compute the cumulative frequencies, which are:

No. of Orders Booked	**No. of Salesmen (f)**	**Cumulative Frequency (F)**
10-20	4	4
20-30	12	16
30-40	25	41
40-50	30	71
50-60	15	86
60-70	8	94
70-80	6	100
	$\sum f = 100$	

Here, (N/2) = 50; so we see that the condition $F_{i-1} < (N/2) \leq F_i$ is satisfied for the values $F_{i-1}=41$ and $F_i=71$. Therefore, median class is 40 – 50.

Now, we use the formulae to calculate the median as:

$$\text{Median} = L + \frac{\frac{N}{2} - C}{f} \times h = 40 + \frac{50 - 41}{30} \times 10 = 40 + \frac{9}{30} \times 10 = 43.$$

CHECK YOUR PROGRESS 2

Note: i) Check your answers with those given at the end of the unit.

4) Find the median of the following series of values:

(i) 10, 6, 15, 2, 3, 12, 8

(ii) 10, 6, 2, 3, 8, 15, 12, 5

5) Find the median of the following frequency distribution

Marks	5	10	15	20	25	30	35	40	45	50
No. of Students	2	5	10	14	16	20	13	9	7	4

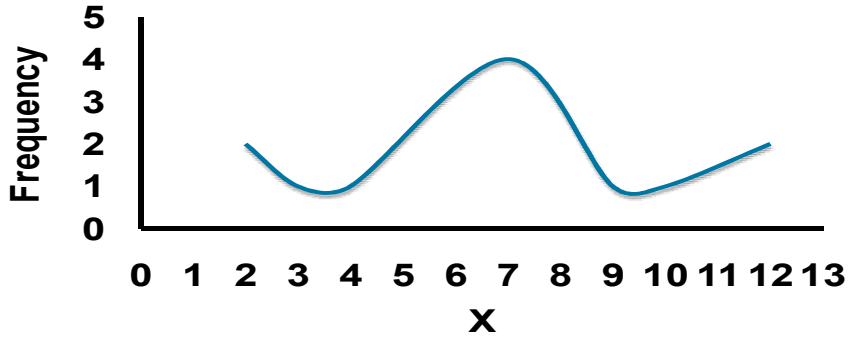
6) Find Median for the following frequency distribution

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of students	5	10	15	20	12	10	8

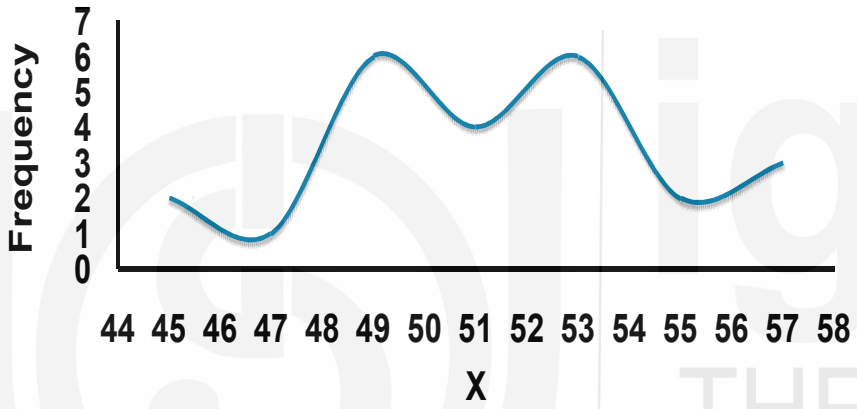
7.6 MODE

The Word 'Mode' comes from the French word 'La Mode' which means the fashion. In statistical language, the Mode is that observation in a distribution which occurs most often, i.e., the value which is most typical. According to Croxton and Cowden, the mode of a distribution is the value around which the items tend to be most heavily concentrated. Thus, mode or the modal value is the value in a series of observations which occurs with highest frequency. For example, when we say that the average size of shoes which are sold in maximum number in a shop is 7, we talk about the modal size of shoes.

In real life situations, generally we observe data with three types of mode. A data set may be either (i) unimodal that is, having only one mode; (ii) bimodal, that is, having two modes and (iii) multi-modal, that is, more than two modes. To illustrate this, let the scores of 10 batsmen be 40, 45, 58, 46, 48, 48, 58, 58, 70 and 81. Then the only mode for the data set is 58 runs, because it appears three times. So, it is a unimodal data set or unimodal distribution. Again, let the runs scored by 10 batsmen of other team be 42, 47, 58, 79, 72, 81, 95, 50, 56 and 69. Then, there is no mode in the above series of scores as every score has equal frequency.



Consider then a third series of scores, say: 86, 62, 58, 58, 48, 47, 48, 81, 59 and 50. Here, we observe that there are two modes 48 and 58, as they both occur twice while other values occur only once. In such a case the distribution is called bimodal. But still, in case of a bimodal distribution, the two modes need not always have the same frequency as shown in the figure given below:



However, this is clear from the above discussion that in some cases mode may be absent while in others there may be one or more modes. If the frequency curve of the frequency distribution is drawn, the mode is the value of the variable at which curve reaches its maximum, i.e., it is the value where the concentration of observations is maximum. In case of a bimodal distribution, the concentration of observation occurs at two points and thus, we have two modes.

Here, it is important that the occurrence of more than one mode in a distribution may be useful for further statistical analysis, but the mode as a measure of central tendency has little significance in the case of bimodal or a trimodal data.

Mode is an important 'average' in many situations specially in data related to marketing studies. For example, when we talk about, a brand of shirt which is most popular amongst younger generation, a song which is listened by the largest section of societies, etc.

7.6.1 Calculation of Mode

(I) For Individual Series

For an individual series the mode can be located simply by inspection or observing the values or observations. Here the values of variable under

consideration may first be arranged in the form of a discrete frequency distribution and then, by looking the frequencies, we can find the value possessing highest frequency. The value with highest frequency would be the mode of the data.

Mathematically, if x_1, x_2,…, x_n are the n observations and if some of the observation are repeated in the data, say x_i is repeated highest times then we can say that x_i would be the mode value.

Thus, in fact, it would be necessary to list the individual values of the variable with their corresponding frequencies. Only then, it would be possible to locate the modal value of the data. We present here an example to illustrate how to find the mode of an individual series of values:

Example 12: Find the mode of the following observations given in an individual series of scores:

9, 9, 10, 11, 11, 11, 12, 14, 14, 17

Solution: Here there are only ten observations in the series we need not put the data in the form of a frequency distribution and mode can be determined by just inspection only. In this case, the observation 11 has the maximum frequency 3. Therefore, 11 is the mode.

(II) For Discrete Frequency distribution

In case of a discrete frequency distribution; mode can be located simply by inspection. Here the variable having the maximum frequency will be taken as mode.

Example 13: Determine mode for the following distribution:

X	10	12	14	16	18	20	22
f	4	6	10	11	21	10	5

Solution: In the above discrete frequency distribution the value 18 has the maximum frequency 21. Therefore, 18 is the mode of the given discrete distribution.

CHECK YOUR PROGRESS 3

Note: i) Check your answers with those given at the end of the unit.

7) Find the mode for the following items

4, 7, 6, 5, 4, 7, 8, 3, 7, 9, 2, 7, 6, 1, 2, 5, 10, 8, 4, 7, 1, 4, 2, 1, 3, 7, 9, 3, 7, 6.

8) Find mode value for the given data

2, 2, 3, 4, 9, 7, 9, 10, 12, 12, 10, 9, 12, 9, 7, 10, 9, 7, 9, 12, 7, 3, 2, 7, 9, 9, 7, 4, 12, 10, 9, 7, 12.

(III) For Continuous Frequency Distribution

We know that in continuous frequency distribution, instead of listing individual values of the variable, values are grouped into various class intervals along with the number of values lying within the class. Therefore, frequency of each and every value in such distributions are not known, hence, the previously stated methods of finding mode is not applicable here. Due to this difficulty, the exact modal value cannot be obtained in such distribution. What then we can do is simply to find an approximate modal value for the distribution using some technique or using some formula for finding approximate mode. There are two approaches for this, (i) using the graphical method, or (ii) using some approximate formula developed from an appropriate diagram under certain assumptions. For this purpose, it is seen that histogram is the appropriate figure which can be utilized. However, if class intervals are made narrower and narrower, the histogram is can be used to locate the modal value by joining mid values of all such classes. Since, the graphical method is practically not a simple method to adopt, mode of continuous frequency distributions is computed using a mathematical formula, which is derived from a histogram under certain assumptions. We shall not provide the proof of the formula here, rather, we shall illustrate its application through examples.

The formula is as follows:

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

where L = lower limit of the modal class,

f_1 = frequency of the modal class,

f_0 = frequency of the pre-modal class,

f_2 = frequency of the post-modal class, and

h = width of the modal class.

Modal class is that class which has the maximum frequency.

Remark 1: For determining mode, the distribution must have continuous classes.

Example 14: Obtain mode of the following distribution:

Classes:	10-20	20-30	30-40	40-50	50-60	60-70
Frequency:	08	12	25	45	11	09

Solution: In this example, the highest frequency 45 lies in the class (40-50) hence, this is the modal class. For determining the mode value in the modal class, we use the formula as:

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

Here, $f_0 = 25,\ f_1 = 45,\ f_2 = 11,\ L = 40$ and $h = 10$.

Thus $$\text{Mode} = 40 + \frac{45-25}{2\times 45 - 25 - 11} \times 10 = 40 + 3.70 = 43.70$$

CHECK YOUR PROGRESS 4

Note: i) Check your answers with those given at the end of the unit.

9) For the following data, calculate mode:

Class Interval	Frequency
0-10	3
10-20	5
20-30	7
30-40	9
40-50	4

7.7 MEASURES OF DISPERSION

According to Spiegel, "the degree to which numerical data tend to spread about an average value is called the variation or dispersion of data". Since, Statistics mostly deals with those characteristics which are variable in nature, that is, varies over unit to unit; dispersion is always present in any set of data and, therefore, in order to explore the nature of the data, the degree of dispersion should be measured by some means. The degree of spreadness or scatteredness of data about an average value is generally explored through some measures of dispersion. Obviously, intuitively it can be thought that such measures must assume value zero if there is no variation in the values of the variable or in other words the characteristic is a constant instead of a variable. The measures of dispersion may be of two types which are mentioned below:

(i) Absolute Measures of Dispersion

Absolute measures of dispersion are those measures which depend upon the unit(s) in which the values of the variable are expressed. This means that if two variables are recorded in different units, then corresponding absolute measures of dispersion are not helpful in comparing their degree of scatteredness. However, if units of these two or more variables are same, they can be compared with each other in respect of their dispersion. Thus, absolute measures of dispersion are useful for comparing variation in two or more distribution where units of measurement are the same. Such measures are not suitable for comparing the variability of the distribution expressed in different units of measurement.

(ii) Relative Measures of Dispersion

Owing to the problem of comparison of two or more variables measured in different units, as described above, we may think of developing some relative measures of dispersion which are free from units of the variables. Such measures are popularly known as "relative measures of dispersion". Relative measures of dispersion are expressed as ratio or percentage or the coefficient of the absolute measures of dispersion. In this way, relative measures are pure unit free ~~less~~ numbers. Relative measures are useful for comparing variability between two or more distributions where units of measurement may be different.

Following are some important measures of dispersion:

1. **Range**
2. **Mean Deviation (M. D.)**
3. **Standard Deviation (S. D.) and Variance**
4. **Coefficient of Variation (C. V.)**

Different measures of dispersion are discussed in following sections:

7.8 RANGE

Range is the simplest measure of dispersion. It is defined as the difference between the largest value (L) and the smallest value (S) of the variable in the distribution. Its merit lies in its simplicity. It can be defined as

$$\text{Range}(R) = L - S$$

where, L : Largest value of the variable, and
S : Smallest value of the variable, as give in the data set.

It should be mentioned here that whether the data are given in the form of a series of the values or in the form of grouped data, in both the cases range is well defined and, therefore, can be computed easily.

Example 15: Find the range of the distribution 6, 8, 2, 10, 15, 5, 1, 13.

Solution: For the given distribution, the largest value of variable is 15 and the smallest value is 1.

Hence, Range (R) = Largest value (L) - Smallest value (S) =15-1= 14.

Example 16: Marks of 10 students in Mathematics and Statistics are given below:

Marks in Mathematics	25	40	30	35	21	45	23	33	10	29
Marks in Statistics	30	39	23	42	20	40	25	30	18	19

Compare the range of marks in the two subjects.

Solution: The lowest and the highest marks in mathematics are 10 and 45 respectively. Also, the lowest and highest marks in statistics are 18 and 42 respectively.

Thus, Range (R) of marks in mathematics = L – S = 45 – 10 = 35

Range (R) of marks in Statistics = L – S = 42 – 18 = 24

CHECK YOUR PROGRESS 5

Note: i) Check your answers with those given at the end of the unit.

10) Calculate the range for the following data:

60, 65, 70, 12, 52, 40, 48

11) Calculate range for the following frequency distribution:

Class	0-10	10-20	20-30	30-40	40-50
Frequency	2	6	12	7	3

7.9 MEAN DEVIATION

As observed, the range is positional measure of dispersion as its computation is not based on all the observations. More so, it is not measure of dispersion in the strict sense of the term as they do not measure scatteredness in observations around an average. As a first measure, mean deviation (M. D.) is defined for meeting the objective. M. D. of a distribution is the arithmetic mean of the absolute deviations of various values from some central value, such as mean, median or mode. Since, the sum of deviations of values from their mean is always zero (an algebraic property of mean); the mean deviation would be always zero even if there is a greater variability in the data. That is why, in defining M. D. the average of the absolute deviations [ignoring plus (+) and (–) minus signs of actual deviations] from some central value are taken. In general, the central value can be any measure of central tendency or even some arbitrarily chosen constant, but usually; being the mostly used measure of central tendency, mean is taken as the central value for taking deviations.

The deviation of an observation x_i from the constant A is defined as $(x_i - A)$, which might be either positive, negative or zero. In order to define M. D., therefore, we ignore the sign of the deviations and consider the absolute values of the deviations which are denoted by $|x_i - A|$.

Thus, the mean deviation (M. D.) defined as

$$\text{M. D.} = \frac{1}{N}\sum_{i=1}^{N}|x_i - A|$$

where, $\sum_{i=1}^{N}|x_i - A|$ = Sum of absolute deviations of values taken from the constant A.

As mentioned above, it can be shown that if we take A = median of the data, the quantity $\sum |x_i - A|$ is least as compared to any other choice of A. But in practice, the mean deviation is computed by taking the constant A as the

mean of the data, since, mean is the most frequently used measure of central tendency. If A is the mean of the data, the M. D. would be given by

$$\text{M. D.} = \frac{\sum_{i=1}^{N} |x_i - \bar{X}|}{N},$$

where $\sum_{i=1}^{N} |x_i - \bar{X}|$ = Sum of absolute deviations taken from mean.

However, if A is the median, M_d; M. D. would be given by

$$\text{M. D.} = \frac{\sum_{i=1}^{N} |x_i - M_d|}{N}$$

Where $\sum_{i=1}^{N} |x_i - M_d|$ = Sum of absolute deviation taken from the median.

In case the data are given in the form of frequency distribution, the formula for M. D. can be put as:

$$\text{M. D.} = \frac{\sum_{i=1}^{k} f_i |x_i - \bar{X}|}{\sum_{i=1}^{k} f_i} \quad \text{and} \quad \text{M. D.} = \frac{\sum_{i=1}^{k} f_i |x_i - M_d|}{\sum_{i=1}^{k} f_i}$$

Without any confusion, in such situations, X_i represents the mid-value of the i^{th} class interval and f_i, the frequency of this class, where there are k such class intervals.

Example 17: Find mean deviation for the given data

1, 2, 3, 4, 5, 6, 7

Solution: First of all we find Mean

$$\bar{X} = \frac{1+2+3+4+5+6+7}{7} = \frac{28}{7} = 4$$

So, we have $|x_i - \bar{X}|$ as 3, 2, 1, 0, 1, 2, 3 and $\sum |x_i - \bar{X}| = 12$.

Therefore, $MD = \frac{12}{7} = 1.71$

Example 18: Find mean deviation for the following data:

X	1	2	3	4	5	6	7
F	3	5	8	12	10	7	5

Solution: We do the necessary calculations in the following table:

X	**f**	**fX**	$\lvert x - \bar{X} \rvert$	$f \lvert x - \bar{X} \rvert$
1	3	3	3.24	9.72
2	5	10	2.24	11.20

3	8	24	1.24	9.92
4	12	48	0.24	2.88
5	10	50	0.76	7.60
6	7	42	1.76	12.32
7	5	35	2.76	13.80
Total	50	212	12.24	67.44

From the values obtained in the columns of the table, we have

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{212}{50} = 4.24$$

and $$M.D = \frac{\sum_{i=1}^{k} f_i \left| x_i - \bar{X} \right|}{\sum_{i=1}^{k} f_i} = \frac{67.44}{50} = 1.348$$

Example 19: Compute the mean deviation (M.D.) from the following data.

Classes:	0–20	20–40	40–60	60–80	80–100	100–120
Frequency:	5	50	84	32	10	6

Solution: We use the technique of change of origin and scale for calculating the mean of the data. We choose the assumed mean as 50 and the constant H for changing the scale, as 20. The transformed values, denoted by d' are then computed and presented in the fourth column of the table. We do other necessary calculations in the following table:

Classes	**X**	**f**	$d' = \frac{x - 50}{20}$	fd'	$\left\|x - \bar{X}\right\|$	$f\left\|x - \bar{X}\right\|$
0–20	10	5	–2	–10	41.07	205.35
20–40	30	50	–1	–50	21.07	1053.50
40–60	50	84	0	0	1.07	89.88
60–80	70	32	1	32	18.93	605.76
80–100	90	10	2	20	38.93	389.30
100–120	110	6	3	18	58.93	353.58
		N = 187		$\sum fd' = 10$		$\sum f\left\|x - \bar{X}\right\| = 2697.37$

$$\therefore\ \bar{X} = A + h\frac{\sum f d'}{N} = 50 + 20\ \frac{10}{187} = 50 + 1.07 = 51.07$$

$$\text{Mean Deviation} = \frac{1}{N}\sum f\left|x - \bar{X}\right| = \frac{2697.37}{187} = 14.42$$

CHECK YOUR PROGRESS 6

Note: i) Check your answers with those given at the end of the unit.

12) Following are the marks of 7 students in statistics. Find the mean deviation.

16, 24, 13, 18, 15, 10, 23

13) Find the mean deviation for the following distribution:

Class	0-10	10-20	20-30	30-40	40-50
Frequency	5	8	15	16	6

7.10 STANDARD DEVIATION AND VARIANCE

In the previous section, we have seen that while calculating the mean deviation, negative deviations are straightaway made positive by finding the absolute value of all the negative deviations. But, since the absolute values are generally difficult to operate and are not amenable to algebraic treatments easily, another way may be thought of converting negative values to positive values. One method may be thought to convert all the deviations into positive quantities by considering the squared values of deviations, instead of taking their absolute values. In this section, we shall use this method for defining some other measures of dispersion.

7.10.1 Root Mean Square Deviation (RMSD)

Before describing other measures of dispersion, such as, Standard Deviation (S.D.), Variance and Coefficient of Variation (C.V.), we shall describe first a general form of measures of dispersion, which is called "Root Mean Square Deviation (RMSD)". The RMSD is defined as the "square root of the mean of squared deviations of all the variate-values taken from an arbitrarily chosen constant, say A".

Thus, if x_i is the value of the variable as recorded on the i^{th} unit (i = 1, 2, …, N), with corresponding mean $\bar{X}$, then RMSD is defined as

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - A)^2}$$

Where, A may be any number. Since, deviations are taken from the constant A, we call this formula as RMSD with respect to the value A.

7.10.2 Standard Deviation (S.D.)

Standard deviation (S. D.), popularly denoted by the letter σ, is a special case of RMSD. It is defined as

$$S.D.(\sigma) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2}$$

Where, $\bar{X}$ is the mean of the data set. It is defined as the positive square root of the mean of the squared deviations taken from the respective mean of the data. You can observe that S.D. is same as RMSD when A = mean of the data (a constant).

The question is that why the constant A has been taken as mean of the distribution and not any other measure of central tendency. The reason is that RMSD is the least when $A = \bar{X}$. Let us show this fact mathematically as follows:

We have

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - A)^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[(x_i - \bar{X}) + (\bar{X} - A)\right]^2}$$

$$= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2} + \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\bar{X} - A)^2} + 2\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})(\bar{X} - A)}$$

Now, if $A = \bar{X}$, it reduces to $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2}$, which is defined as S.D. So, we see that RMSD attains its least value when constant A is taken as the mean of the data, and then RMSD is called Standard Deviation (σ).

The concept of standard deviation was first introduced by Karl Pearson. It satisfies most of the properties of a good measure of dispersion. However, it can be seen that S.D. is not unit free. Its unit is same as that of x-values.

Alternative formula for computation of Standard Deviation:

In order to simplify the calculation process and also to reduce the error of rounding to the minimum extent, we can use an alternative formula for the calculation purpose of Standard Deviation, which is obtained below:

We have

$$\text{Standard Deviation } (\sigma) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2} = \sqrt{\frac{1}{N}\left(\sum_{i=1}^{N}x_i^2 + \sum_{i=1}^{N}\bar{X}^2 - 2\sum_{i=1}^{N}x_i.\bar{X}\right)}$$

Thus, we have an alternative form of the formula of S.D. as

$$\text{Standard Deviation } (\sigma) = \sqrt{\frac{1}{N}\left(\sum_{i=1}^{N}x_i^2 - N\bar{X}^2\right)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}x_i^2 - \bar{X}^2}$$

If the data are given in the grouped data form, either in the discrete or continuous frequency distribution, the formula for S. D. becomes:

$$\text{Standard Deviation}(\sigma) = \sqrt{\frac{1}{N}\sum_{i=1}^{k} f_i(x_i - \bar{X})^2} \qquad \forall i = 1,2,\ldots,k$$

where, there are k class intervals; $N = \sum_{i=1}^{k} f_i$ and f_i is the frequency of the i^{th} class interval.

Proceeding in the same manner as we did above to find the alternative formula of standard deviation; we can derive alternative formula for standard deviation when data are available in the grouped form, which is given as

$$\text{Standard Deviation}(\sigma) = \sqrt{\frac{1}{N}\sum_{i=1}^{k} f_i x_i^2 - \bar{X}^2} \quad \text{where, } N = \sum_{i=1}^{k} f_i$$

7.10.3 Variance

Variance is another frequently used measure of dispersion. Actually, variance is nothing but the square of the standard deviation and hence, is denoted by σ^2.

Variance is the average of the square of deviations of the values taken from the respective mean. Thus, variance is defined as

$$\text{Variance}(\sigma^2) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2$$

and for a frequency distribution, the formula is

$$\text{Variance}(\sigma^2) = \frac{1}{N}\sum_{i=1}^{k} f_i(x_i - \bar{X})^2$$

where, all symbols have their usual meanings as given for the formula of standard deviation above. The unit of the variance will be the square of the variable X.

Alternative formula for computation of Variance:

Since, variance is the square of standard deviation; the same problem arises for variance also. Therefore, we have to use an alternative formula for variance also which must be used for computation purpose.

Squaring the alternative formula for standard deviation, we have alternative formula for calculation of the variance, which is given by

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{X}^2$$

Similarly, the alternative formula for computing σ^2 when the data are given in the grouped form, will be

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} f_i x_i^2 - \bar{X}^2 \qquad \text{where, } \bar{X} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} \quad \text{and } N = \sum_{i=1}^{k} f_i$$

k being the number of class intervals.

Example 20: Calculate S.D. for the following set of observations:

X: 10 11 17 25 7 13 21 10 12 14

Solution: There are 10 values of the variable, therefore, n = 10.

Here, we have $\sum_{i=1}^{10} x_i^2 = 10^2+11^2+ \ldots.+14^2 = 2234$. Also, $\sum_{i=1}^{10} x_i = 140$.

Therefore, $\bar{X} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{140}{10} = 14$;

and hence, $S.D.(\sigma) = \sqrt{\frac{2234}{10} - (14)^2} = \sqrt{223.4 - 196} = \sqrt{27.4} = 5.2345$

Example 21: Calculate the S.D. of the following discrete frequency distribution:

Size (X)	4	5	6	7	8	9	10
Frequency	6	12	15	28	20	14	5

Solution: For rest of the computation, we prepare the following table:

Size (X)	4	5	6	7	8	9	10	Total
Frequency (f)	6	12	15	28	20	14	5	100
fX	24	60	90	196	160	126	50	706
fX^2	96	300	540	1372	1280	1134	500	5222

Therefore, standard deviation is

$$S.D.(\sigma) = \sqrt{\frac{5222}{100} - (7.06)^2} = \sqrt{52.22 - 49.8436} = \sqrt{2.3764} = 1.542$$

Example 22: Calculate standard deviation from the following data:

Age (in years)	4–6	6–8	8–10	10–12	12–14	14–16	16–18
No. of Students	30	90	120	150	80	60	20

Solution: Let us use the short-cut method for the calculation. We choose A = 11 and H = 2. Then, the calculations are shown in the table given below:

Age	**Frequency f**	**Mid – Value**	$d = \frac{X-11}{2}$	fd	fd^2
4–6	30	5	–3	–90	270
6–8	90	7	–2	–180	360
8–10	120	9	–1	–120	120
10–12	150	11	0	0	0
12–14	80	13	1	80	80

14–16	60	15	2	120	240
16–18	20	17	3	60	180
	N = 550			$\sum fd = -130$	$\sum fd^2 = 1250$

We know that $(\sigma_X) = (h\sigma_D)$. So, the S. D. of X (the original variable) will be

$$\sigma(X) = h\sqrt{\left[\frac{\sum f d^2}{N} - \left(\frac{\sum f d}{N}\right)^2\right]}$$

$$= 2\sqrt{\frac{1250}{550} - \left(\frac{-130}{550}\right)^2} = 2 \times \sqrt{2.272 - 0.055} = 2.977$$

CHECK YOUR PROGRESS 7

Note: i) Check your answers with those given at the end of the unit.

14) Find the standard deviation for the following data.

10 27 40 60 33 30 10

15) Calculate standard deviation for the following data:

Class	0-10	10-20	20-30	30-40	40-50
Frequency	5	8	15	16	6

7.11 COEFFICIENT OF VARIATION

It is now clear that the standard deviation or the variance as measures of dispersion, give us an idea about the extent to which observations are scattered around their mean. Therefore, two or more distributions can be compared directly for their variability of such characteristics of two or more distributions, which are measured in same unit with the help of corresponding standard deviations and a distribution having the largest (smallest) value of S. D. or variance as compared to the same of the other distributions, may be said to be most dispersed (least dispersed) distribution than other distributions.

However, mostly, when we wish to compare the variability of such characteristics of two or more distributions, which are measured in different units or even to compare two or more characteristics in the same distribution such that the characteristics are measured in different units; a problem arise due to the non-conversion of units. We know that any unit of height measurement is not convertible to any unit of weight measurement. So, variability of heights is not directly comparable on the basis of S. D. or variance. Similar may be the case with other characteristics also. Then, the question is how to compare variability's of such characteristics which are measured in different units.

As before, we define a relative measure of variability with the help of S. D. for this purpose. For making comparisons in the above two situations, we use

a relative measure of dispersion, called coefficient of variation (C.V.) which is defined as

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

It is easy to see that the unit of the numerator is same as that of the denominator and, hence, it is a unit free number and, therefore, it is called to be a coefficient. In this sense it is a relative measure of variability. If we are comparing the two data series, the data series having smaller CV will be more consistent. The distribution having greater C.V. is considered to be more dispersed than the other distribution with lesser C.V.

Example 23: Suppose batsman A has mean 50 with SD 10 of runs he scored and Batsman B has mean 30 with SD 3 for the runs he scored. What do you infer about their performance?

Solution: Obviously, in cricket, a batsman is said to be more consistent or reliable in his performance if the variability of runs he scored from match to match is less.

A batsman is not said to be a good performer if the variability of scores is larger. However, since A has higher mean of runs than B, this means A is a better run maker.

However, B has lower CV (3/30 = 0.1) than A (10/50 = 0.2) it means that B is more consistent.

Example 24: Two workers on the same job show the following results over a long period of time:

	Worker A	**Worker B**
Mean time of completing the job (minutes)	25	20
Variance	36	16

i) Which worker appears to be more consistent in the time he requires to complete the job?

ii) Which worker appears to be faster in completing the job?

Solution: (i) Given that $\bar{X}_1 = 25,\ \bar{X}_2 = 20,\ \sigma_1^2 = 36, \sigma_2^2 = 16$.

For worker A, Coefficient of Variation $= \frac{\sigma_1}{\bar{X}_1} \times 100 = \frac{6}{25} \times 100 = 24\%$

For worker B, Coefficient of Variation $= \frac{\sigma_2}{\bar{X}_2} \times 100 = \frac{4}{20} \times 100 = 20\%$

i) The comparison of C.V.s led to the conclusion that worker B appears more consistent in the time needed for completing the job.

ii) Since the mean time taken to complete the job is less for worker B, so worker B appears to be faster in completing the job.

Note: i) Check your answers with those given at the end of the unit.

16) The following table shows the results of an analysis relating to the prices of two shares X and Y, that were recorded during the year 2000:

Share	Arithmetic mean (Rs.)	Standard Deviation (Rs.)
X	45.00	9.00
Y	62.50	18.75

Which of the two shares, X or Y, have more variability in prices?

17) If n=10, $\bar{x} = 24$, $\sum x^2 = 200$, find the coefficient of variation.

7.12 LET US SUM UP

In this unit, we have discussed:

1. Some basic measures of central tendency and methods of calculation of the Arithmetic mean, Median and Mode for different kinds of data.
2. properties of a good measure of measures of central tendency and dispersion which are supposed to be satisfied.
3. different types of measures of dispersion; like, Range, Root Mean Square Deviation, Standard deviation, Variance and Coefficient of Variation.
4. methods of calculation of Range, Standard deviation, Variance and Coefficient of Variation for different kinds of data.

7.13 KEY WORDS

Measures of Central Tendency : A general term for the various averages that attempt to describe the middle or typical value in a distribution.

Measures of Variability : A general term for various measures of the amount by which scores are dispersed or scattered.

7.14 SUGGESTED FURTHER READING/ REFERENCES

Witte, R., & Witte, J. (2017). *Statistics*. Hoboken, NJ: John Wiley & Sons.

7.15 ANSWERS TO CHECK YOUR PROGRESS

1) For calculating the arithmetic mean, we add all the observations and divide by 6 as follows:

$$\bar{X} = \frac{\sum x_i}{n} = \frac{5+8+12+15+20+30}{6} = 15$$

2) For calculation of mean, the following table is made:

Wages	No. of Workers (f)	fX
20	4	80
25	8	200
30	20	600
35	12	420
40	6	240
	$\sum_{i=1}^{5} f_i = 50$	$\sum f_i x_i = 1540$

$$\bar{X} = \frac{\sum_{i=1}^{5} f_i x_i}{5} = \frac{1540}{50} = 30.8$$

3) Using short cut method

Marks	f	X	$d = \frac{X - 25}{10}$	fd
0-10	6	5	–2	–12
10-20	9	15	–1	–9
20-30	17	25= A	0	0
30-40	10	35	+1	+10
40-50	8	45	+2	+16
	$\sum f_i = 50$			$\sum f_i d_i = 5$

$$\text{Now } \bar{X} = A + \frac{\sum f_i d_i}{N} \times h \quad \Rightarrow \quad \bar{X} = 25 + \frac{5}{50} \times 10 = 26$$

4) (1) Here, N = 7, an odd number. After arranging the values in ascending order, we get 2, 3, 6, 8, 10, 12, 15 and, therefore, the Median value will be

Median = value of $[(N+1)/2]^{th}$ item= value of 4^{th} item = 8.

(2) Here N= 8, an even number, so arranging values in ascending order we get the values as 2, 3, 5, 6, 8, 10, 12, 15 and, therefore,

Median = Mean of $[N/2]^{th}$ and $[(N/2)+1]^{th}$ values = (6+8)/2 = 7.

5) For the given data, we compute less than cumulative frequencies as follows:

Marks	5	10	15	20	25	30	35	40	45	50
No. of Students	2	5	10	14	16	20	13	9	7	4
Cumulative Frequency	2	7	17	31	47	67		80	89	96
	100									

Since, here N is even, so there will be two middle-most values, the first is $(N/2)^{th}$ value and the second is $[(N/2) + 1]^{th}$ value.

Here, N/2 = 50 and [(N/2) + 1] = 51. Thus, median is equal to the mean of the 50^{th} and 51^{st} values. We see that $(N/2)^{th}$ value satisfies the condition $F_5 < (N/2) < F_6$.

So, $(N/2)^{th}$ value will be equal to 30. Also, [(N/2)+1] value is 30. Therefore, their mean is also 30. So, median is 30.

6) First we shall calculate the cumulative frequency distribution

Marks	**f**	**Cumulative Frequency**
0-10	5	5
10-20	10	15
20-30	15	30=C
30-40	**20=f**	**50**
40-50	12	62
50-60	10	72
60-70	8	80
	N= 80	

Here N/2 = 40. Since, 40 is not in the cumulative frequency so, the class corresponding to the next cumulative frequency 50 is median class. Thus 30-40 is median class.

$$\text{Median} = L + \frac{\left(\frac{N}{2} - C\right)}{f} \times h = 30 + \frac{40-30}{20} \times 10 = 35$$

7) For calculation of mode, the following table is made:

Size	1	2	3	4	5	6	7	8	9	10
Frequency	3	3	3	4	2	3	7	2	2	1

From the frequency distribution we see that size 7 occurs with maximum frequency of 7 hence mode is 7.

8) First we prepare frequency table as

X	2	3	4	7	9	10	12
f	3	2	2	7	9	4	6

This table shows that 9 have the maximum frequency. Thus, mode is 9.

9) From the given frequency distribution corresponding to highest frequency 9 modal class is 30-40 and, therefore, we have L=30, f_1 =9, $f_o = 7$, $f_2 = 4$, $h = 10$ Applying the formula,

$$\text{Mode} = 30 + \frac{9-7}{2\times 9-7-4}\times 10 = 30 + 2.86 = 32.86$$

10) Range $R = X_{Max} - X_{Min}$ = 70-12 = 58

11) First convert the given inclusive class limits continuous classes (Exclusive classes) as shown in the table:

Inclusive Classes	**Continuous Classes**	**Frequency**
6–10	5.5–10.5	7
11–15	10.5–15.5	8
16–20	15.5–20.5	15
21–25	20.5–25.5	35
26–30	25.5–30.5	18
31–35	30.5–35.5	7
36–40	35.5–40.5	5

Here, L = 40.5 and S = 5.5

Then we have R = 40.5 – 5.5 = 35

12) For calculation of mean deviation, the following table is made:

x	**(x-17)**	$\lvert x - \bar{X}\rvert$
16	-1	1
24	+7	7
13	-4	4
18	+1	1
15	-2	2
10	-7	7
23	+6	6
$\sum x = 119$		$\sum \lvert x_i - \bar{X}\rvert = 28$

$$\bar{X} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{119}{7} = 17 \quad \text{And} \quad M.D = \frac{\sum_{i=1}^{N} |x_i - \bar{X}|}{N} = \frac{28}{7} = 4$$

13) For calculation of mean deviation, the following table is made:

Class	**x**	**f**	**xf**	$(x - \bar{X})$	$\lvert x - \bar{X} \rvert$	$f\lvert x - \bar{X} \rvert$
0-10	5	5	25	-22	22	110
10-20	15	8	120	-12	12	96
20-30	25	15	375	-2	2	30f
30-40	35	16	560	8	8	128
40-50	45	6	270	18	18	108
		50	1350			$\sum f\lvert(x - \bar{X})\rvert = 472$

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{1350}{50} = 27 \text{ and } M.D = \frac{\sum_{i=1}^{k} f_i |x_i - \bar{X}|}{\sum_{i=1}^{k} f_i} = \frac{472}{50} = 9.44$$

14) For calculation of standard deviation, the following table is made:

X	**X^2**
10	100
27	729
40	1600
60	3600
33	1089
30	900
10	100
$\sum x = 210$	$\sum x^2 = 8118$

$$\bar{X} = \frac{\sum x}{N} = \frac{210}{7} = 30$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x^2 - (\bar{X})^2} = \sqrt{\frac{8118}{7} - 900} = \sqrt{1159.7} = 34.61$$

15) Let us take A=25

Class	**x**	**f**	**d = x-A**	**f d**	**$f d^2$**
0-10	5	5	-20	-100	2000
10-	15	8	-10	-80	800

20 20-30 30-40 40-50	25 35 45	15 16 6	0 10 20	0 160 120	0 1600 2400
		N=50		$\sum fd = 100$	$\sum fd^2 = 6800$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{k} f_i d_i^2 - \left(\frac{\sum_{i=1}^{k} f_i d_i}{N}\right)^2} = \sqrt{\frac{1}{50}(6800) - \left(\frac{100}{50}\right)^2}$$

$$= \sqrt{136 - 4} = \sqrt{132} = 11.49$$

16) We are given that $\bar{X} = 45, \bar{Y} = 62.50, \sigma_X = 9, \sigma_Y = 18.75$

C.V. for share $X = \frac{\sigma_X}{X} \times 100 = \frac{9}{45} \times 100 = 20\%$

C.V. for share $Y = \frac{\sigma_Y}{Y} \times 100 = \frac{18.75}{62.50} \times 100 = 30\%$

On comparing the coefficients of variation for two shares, we find that share Y shows greater variation in prices.

17) We have

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{X}^2 = \frac{1}{10} \times 200 - (4)^2 = 20 - 16 = 4$$

$$\Rightarrow \sigma^2 = 4 \Rightarrow \sigma = 2$$

Therefore, coefficient of variation

$$C.V.(X) = \frac{\sigma}{\bar{X}} \times 100 \Rightarrow \frac{2}{4} \times 100 = 50\%$$

UNIT 8 DESCRIPTIVE STATISTICS-II

Structure

8.1 INTRODUCTION

In the previous unit, you have studied various measures, such as, measures of central tendency and measures of dispersion which are virtually used for analyzing the nature of a single variable, say X, and its distribution in many aspects. These measures, in fact, disclosemany such aspects of the variable, which are considered to be basic properties of the variable for its further study. However, when we come across with simultaneous study of two or more variables, besides the separate study of these variables, one thing is more important and necessary to note. This is, whether the variables exhibit some kind of relationship between them or whether they do not exert any kind of impact on others, that is, variables are supposed to be independent to each other. For example, take the case of price, demand and supply of 50 commodities in a market. We know that, according to laws of economics, variables 'price' and 'demand' exhibit one kind of direct relationship, since,

increasing demand results to increase in price also. On the other hand, 'supply' and 'price' have an indirect relationship, since, more the supply, less would the price of commodities. In statisticalterminology, if there exist any type of relationship between two variables, we say that they are mutually 'correlated' to each other. Thus, the word 'correlation', in fact, describes a specific property of two variables.

If we state that the relationship between the two variables under consideration is linear, it means that the this relation may be represented in the mathematical form $Y = b_0 + b_1X$; which is the equation of a straight line, where Y and X are respectively the dependent and independent variables and b_0 and b_1 are some suitable constants (we shall use the word 'parameter' for such constants in further text) so chosen that the straight line passes through maximum number of points (x_i, y_i) for $i = 1, 2, \ldots, N$; on the scatter diagram. This is same as to say that if our purpose is to estimate or to find the value of the dependent variable Y for a given value of the independent variable X through the straight line obtained, then the actual value Y and the corresponding estimated value of it must be as close to each other as possible, that is, the difference of actual and estimated values must be minimum for each Y-values. Such a line used for this purpose is generally called a "Regression Line" and the theory behind regression lines and their applications to practical problems is termed "Linear Regression Analysis".

Section 8.3 of this unit discusses the conceptsand different type ofcorrelationssupported with a number of practical examples that describe the situationsunder which they might be observed.Section 8.4 describes the 'scatter diagram' which helps us diagrammatically to explore the nature of the relationship between the variables. Section 8.5 describes a measure of correlation, popularly known as "correlation Coefficient", which provides the type of correlation and the magnitude of the correlation between two variables. Some of the salient properties of it are mentioned and theoretically proved. Section 8.6 discusses rank correlation coefficient, which provides the magnitude of the correlation between two attributes. Section 8.7 here explains the basic meaning of the word 'regression'. Section 8.8 describes the mathematical procedure of obtaining all the linear regression lines in case of two variables. Some of the properties of linear regression lines with one independent variable are also discussed. Section 8.9 provides the meaning of regression coefficients in regression analysis and their significance. Some important properties of regression coefficients are also discussed.

8.2 OBJECTIVES

After studying this unit, you should be able to:

- explain the concept of correlation between two variables and identify different types of correlations;
- describe the scatter diagram and its use in correlation analysis;

- explain the computation method and describe the important properties of correlation coefficient;
- explain regression, regression analysis and of linear regression;
- derive expressions of lines of regression of Y on X and X on Y mathematically on the basis of least squares principle;
- explain how to apply both the regression lines in practical problems for predicting the value of one variable given the value of the other; and
- explain the salient properties of regression coefficients.

8.3 CORRELATION ANALYSIS

In previous units we were confined to the study of nature and properties of only one variable. What kind of data in such case are given to us were values of the variable as measured over a number of objects, that is, data were univariate data. However, in many practical situations, we may have data in which more than one variable can be associated with each unit of observation. For example, for providing the information about the marks obtained by the students of a class in two subjects, say,Statistics and Mathematics; the values associated with each student are the marks obtained by them in Mathematicsas well as in Statistics. A practical example of such a data is depicted in the following table:

Number of Student	1	2	3	4	5	6	7	8
Marks in Statistics (X)	90	88	52	48	67	85	56	65
Marks in Mathematics (Y)	95	70	64	75	38	69	86	54

Here, each student assumes values on two variables X (marks in Statistics) and variable Y (marks in Mathematics) simultaneously and, as such, we get a 'bivariate distribution'. You may also recall here that in a bivariate distribution, we have two variables of observation on which values are recorded for each unit of observation. Similarly, the distributions involving more than two variables are termed as "multivariate distributions". In this unit we confine ourselves to the study of bivariate distributions only in which two variables may be inter-dependent.

Let us consider the example of sales revenue and expenditure on advertising in business. A natural question arises in mind that is there any connection between sales revenue and expenditure on advertising? Does sales revenue increase or decrease as expenditure on advertising increases or decreases?Similarly, if we see the example of time spent on study and marks obtained by students, a natural question appears whether marks increase or decrease as time spent on study increases or decreases.

In all these situations and some other situations, we try to find out a type of relation, if any, between two variables and its magnitude. Then, correlation analysis comes up in the way for answering the question, if there is any relationship between one variable and the other.

For instance, generally, the weight of a person is seen to increase with height up to a certain age; prices of commodities seem to vary with supply; pressure exerted on a flexible body generally reduces its volume and, thus, are seen to have been inversely related; cost of industrial production varies with the cost of raw material and so on. In all these cases, the change in the value of one variable appears to be accompanied by a change in the values of other variable. If it happens, variables are said to be correlated, and then this relationship is called correlation or covariation. The study and measurement of the extent or degree of relationship between the two or more than two variables along with the nature of relationshipbetween them is called correlation analysis.

8.3.1 Types of Correlation

Depending upon the nature of different types of characteristics (variables) seen in practice, correlation can be categorized into following three types of cases:

i) Positive, Zero and Negative Correlation

ii) Simple, Multiple and Partial Correlation

iii) Linear and Non-Linear Correlation.

Let us discuss these types of correlation one at a time as follows:

i) Positive, Zero and Negative Correlation

When we compute the magnitude of relationship between two variables using some measure, we come across with two types of values; one, positive value of correlation and, two, negative value of the correlation. We should to understand why this happens.

Since, as mentioned above, the coefficient of correlation measures the magnitude of the linear relationship, the magnitude may be either a positive or a negative quantity. Accordingly, the correlation may be positive or negative according to the direction of change in the two variables. Positive correlation refers to the change in the values of both the variables in the same direction, that is, the values of both the variables increase or decrease in the same direction. One such relationship is seen between variables ‘income’ and ‘expenditure’. Generally, with increasing income, expenditure of families also seen to increase but with decreasing income, expenditure is also seen to decrease.

Similarly, Negative correlation refers to the change of the values of the variables in opposite directions, that is, if the values of one variable increase the values of the other variable decrease (on an average) and if the values of

one variable decrease, then the values of other variable increase. For example, the price of a product and demand in the market, volume and pressure of perfect gas, etc.

Sometimes, when we find correlation between two variables, we get it very close to zero or exactly zero. It is an indication that the two variables are not affecting each other at all or it is because of the reason that they are not linearly related. This situation is termed as the case of zero correlation.

ii) Simple, Multiple and Partial Correlation

Based on the number of variables, correlation may be classified in the following three types of classes:

a) Simple Correlation: Simple correlation is the study of correlation between only two variables, that is, if we measure the degree or intensity of the relationship between two variables only, it is called simple correlation. For example, when we study the correlation between the price and demand of a product, it is a problem of simple correlation.

b) Multiple Correlation: Multiple correlation is defined and computed when we consider more than two variables at a time. Let there be (n + 1) variables Y and X_1, X_2, X_3, …, X_n and we use a relationship between Y and all other n variables as

$$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + \ldots + a_nX_n;$$

where a_i for i = 0, 1, 2,…, n are (n + 1) unknown constants.Since, in this relationship Y seems to be dependent on n independent variables X_i for i = 1, 2, …, n;obviously, Y- variable can be estimated on the basis of these n variables, which may be denoted by $\hat{Y}$ .Then, if we find the simple correlation between the two variables; namely, variable Y and the estimated variable $\hat{Y}$; it is termed as "Multiple correlation". In this sense, multiple correlation is also a simple correlation. For example, if we study the relationship between the observed yield of rice(Y),observed on n fields and the estimated yield ($\hat{Y}$), when the amount of rainfall (X_1), fertility of soil of fields (X_2), amount of fertilizers used(X_3), etc., were also recorded and taken into consideration in the relationship, for estimating the yield, then it becomes the case of multiple correlation. We shall discuss it separately afterwards.

c) Partial Correlation: In b) above, we observed that sometimes we may consider more than two variables for finding a particular type of correlation, which is multiple correlation. Now, think the situation in another way for the relationship given between Y and n X-variables there. Since, yield of a crop is always thought to be affected by the variables-amount of rainfall (X_1), soil fertility(X_2), amount of fertilizer used (X_3) and other variables; we can say that the observed value of the yield (Y) is the effect of a number of multiplicity of causes. Therefore, if anyhow we can find that value of the yield (say, $\tilde{Y}$) , which could be obtained by eliminating the effect of these variables on the yield and find the

correlation between Y and $\tilde{Y}$, it would have produced another kind of correlation. This type of correlation is termed as Partial correlation. The study of it is again a separate subject matter, which we shall consider afterwards.

iii) Linear and Non-Linear Correlation

a) Linear correlation: In the section 10.2 above, we have mentioned the meaning of linear and non-linear relationships between two variables. A relation between the variables Y and X is said to be 'linear' if points of values (y_i, x_i); i = 1, 2, 3, …, N, when plotted on a graph paper, exhibit a straight line. For example, the following set of Y and X values represent a linear relationship:

X:	10	20	30	40	50
Y:	45	85	125	165	205

We see that the relationship between Y and X may be expressed as Y = 4X +5, which is a straight-line representation of the above values. Now, if the values of the variables Y and X exhibit an exact linear relationship or are very near to a linear relationship; the correlation between the variables is called to be 'linear correlation'. We have already pointed out that this unit is devoted to the study of correlation analysis when we have a linear relationship. In other words, when the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable, the variables are said to have linear correlation.

b) Non-Linear Correlation: When the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable, it is called a non-linear relationship, since the graph of such a relation is not a straight line. Some examples of such relations are, $Y = a + bX + cX^2 + dX^3$; $Y = ab^X$ and $Y = ae^{bX}$; etc.

Whenever, the values of the variables follow a non-linear relationship, the correlation between the two variables is a non-linear correlation. For example, if the data are given as

X:	0	2	3	6	5
Y:	5	45	135	3645	1215

the relationship cannot be expressed by a straight line; rather it is represented by the relation $Y = 5(3^X)$, which is non-linear. The curve drawn along with these values it will show a curvilinear graph.

There are available some methods in the literature which measure the correlation for curvilinear relationship. However, these will not be discussed in this unit.

8.3.2 Measures of Correlation

In order to know have an idea of the type of relationship between two variables and to find the extent of correlation between them, there are some methods, which are listed below:

i) Scatter Diagram Method

ii) Karl Pearson's Coefficient of Correlation

iii) Spearman's Coefficient of Rank Correlation

We shall be discussing these methods one by one in the following sections.

CHECK YOUR PROGRESS 1

Note: i) Use the space given below for your answers.
ii) Check your answers with those given at the end of the unit.

1. What do you mean by Correlation?

……

……

……

……

2. Mention few examples of different types of correlations.

……

……

……

……

8.4 SCATTER DIAGRAM

Scatter diagram is a statistical tool for determining the potentiality of correlation between dependent and independent variables. However, Scatter diagram does not tellus about the exact relationship between them, neither it is helpful in providing the degree of relationship of correlation. It is simply helpful to observe whether they are correlated or not.

Scatter Diagram is graphical device for drawing certain conclusions about the correlation between two variables. In preparing a scatter diagram, the observed pairs of observations are plotted on a graph paper in a two-dimensional space by taking the measurements on variable X along the horizontal axis and that on the variable Y along the vertical axis. The pairs of values are represented by dots on the graph. The diagram of dots so obtained is known as "scatterdiagram". The pattern and the direction of the scatteredness of points provides some rough idea about the nature and degree of the relationship.

Let $(x_i, y_i); (i = 1, 2, ..., N)$ be the bivariate distribution under consideration. By plotting the N pairs of observations $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$ on a graph paper, we have a scatter diagram from which we can draw certain rough conclusions concerning the extent to which the points cluster about a straight line as follows:

1. If all the plotted points, representing a given data, lie on a straight line, or seems to be very closer to certain straight line, then it indicates that the relationship is linear and not curvilinear.
2. If dots are in the shape of a line and line rises from left bottom to the right top it indicates that the variablesdo have a linear relationship and both the variables are having a positive correlation.
3. If all the points lie on a straight line or seems to be very close to certain straight line such that it has a negative slope, then it indicates again a linear relationship but a negative correlation, that is, an inverse relationship between the variables.
4. In case of linear relationship, described in 2 and 3, if all the points exactly lie on a certain straight line, the degree of correlation is said to be 'perfect'. If it is point2, the variables are said to have 'positive perfect correlation' and if it is point 3, the variables are said to have 'negative perfect correlation'.
5. If plotted points on scatter diagram do not show any trend,it is an indication thatthere is almost no correlation or a weak correlation between the variables.

A discussion on the usefulness of scatter diagram in correlation analysis was made in this section, which makes it clear that such a diagram is useful only to have a rough idea about the trend of the values and particularly to know whether the relationship is linear or not, so that the measurement of degree of correlation can be made with fruitful results. The diagram is not helpful in any way to state that the relationship is strong, moderately strong, average or poor.

8.5 KARL PEARSON'S COEFFICIENT OF CORRELATION

We mentioned in the above sections a number of times that when we study two variables simultaneously, we are interested to know whether they have some correlation between them and if yes, what is the nature of relationship and how much they are related, that is, what is the degree of correlation. Scatter diagram, though, tells us whether variables are correlated or not;but it does not indicate the extent to which they are correlated.

No doubt, in order to know the degree of correlation (relationship), we need some measure which can reflect the degree of relationship in terms of some numerical value. Then, the value of the measure may numerically provide us the idea of closest relationship, high relationship, moderate relationship, weak relationship or no relationship between the variables. Fortunately, in literature of Statistics, such a measure exists. It is known as "Coefficient of Correlation".

Coefficient of correlation given by British biometrician Karl Pearson (1867-1936) gives the exact idea of the extent to which two variables are

correlated.The Karl Pearson's coefficient of correlation is also known as "product moment correlation coefficient". It is the most widely used method of determining theextent and nature of linear relationship between two variables.Coefficient of correlation measures, in numerical terms, the intensity or degree of linear relationship between two variables.

Notationally, if X and Y are two random variables,then "Karl Pearson's correlation coefficient"or"Product Moment Correlation Coefficient" between X and Y is generally denoted byletter r orby Corr(X, Y) and is defined as

$$r = \mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{V(X)V(Y)}}; \qquad \ldots(1)$$

where, $\mathrm{Cov}(X,Y)$ stands for the covariance between X and Y which is definedas:

$$\mathrm{Cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})$$

and $V(X)$, V(Y) stand for the variance of X, Y respectively. Here, V(X) is defined as:

$$V(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2.$$

Similarly, $V(Y)$ is defined as

$$V(Y) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^2$$

where, N is number of paired observations.

Using the definitions of covariance, and variance as given above,the correlation coefficient "r" may be defined as:

$$r = \mathrm{Corr}(X,Y) = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2\right)\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^2\right)}}. \qquad \ldots(2)$$

If V(X) and V(Y) are denoted by the notations σ_X^2 and σ_Y^2 respectively, then the formula of coefficient of correlation is reduces to

$$r = \mathrm{Corr}(X,Y) = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})}{\sigma_x \sigma_y}. \qquad \ldots(3)$$

It can be seen that given the N values of the variable Y, that is, $y_1, y_2, \ldots, y_N$ and corresponding values of the variable X; $x_1, x_2, \ldots, x_N$; it is easy to compute the quantities $\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})$, $\sum_{i=1}^{N}(y_i - \bar{Y})^2$ and $\sum_{i=1}^{N}(x_i - \bar{X})^2$ and hence to compute the value of r.

The Karl Pearson's correlation coefficient r is seen to be a unit free quantity, the reason being that the numerator and denominator have same units of measurement. If X is measures in kilograms and Y in meters, then the unit of

numerator will be "kilogram meter" whereas in the denominator, the unit of standard deviation, σ_x, would be kilogram and that of σ_y would be meter. Hence, unit of numerator is same as that of denominator and hence cancels each other. Because r does not depend on any unit, it is called a coefficient.

CHECK YOUR PROGRESS 2

Note: i) Use the space given below for your answers.
ii) Check your answers with those given at the end of the unit.

3. Define product moment correlation coefficient.

……

……

……

……

8.5.1 Properties of Correlation Coefficient

Before going into the problem of computing r using different methods,we shall discuss here some of the salient properties of it.

Property 1 (Limits of r): The value of product moment correlation coefficient(Karl Pearson's coefficient of correlation) always lies in the range [–1 to 1], where–1 and 1 are inclusive in the range.

Proof: Consider the inequality

$$\frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{x_i-\bar{X}}{\sigma_x}\right)\pm\left(\frac{y_i-\bar{Y}}{\sigma_y}\right)\right]^2\geq 0,$$

which is always true for all values of x_i and y_i, since, the left-hand side (L.H.S.) expression is a squared quantity. Now, considering first the plus sign between the two terms and expanding the L.H.S., we get

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i-\bar{X}}{\sigma_x}\right)^2+\frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i-\bar{Y}}{\sigma_y}\right)^2+\frac{2}{N}\sum_{i=1}^{N}\left(\frac{x_i-\bar{X}}{\sigma_x}\right)\left(\frac{y_i-\bar{Y}}{\sigma_y}\right)\geq 0$$

$$\text{or,}\quad \frac{1}{N}\frac{\sum_{i=1}^{N}(x_i-\bar{X})^2}{\sigma_x^2}+\frac{1}{N}\frac{\sum_{i=1}^{N}(y_i-\bar{Y})^2}{\sigma_y^2}+\frac{2}{N}\sum_{i=1}^{N}\left(\frac{x_i-\bar{X}}{\sigma_x}\right)\left(\frac{y_i-\bar{Y}}{\sigma_y}\right)\geq 0. \quad \ldots (4)$$

But, we know that

$$\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{X})^2=V(X)=\sigma_x^2;\ \text{and}\qquad \frac{1}{N}\sum_{i=1}^{N}(y_i-\bar{Y})^2=V(Y)=\sigma_y^2$$

$$\text{andalso}\ \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{X})(y_i-\bar{Y})}{\sigma_x\sigma_y}=\ r,\ \text{from (3).}$$

Therefore, (4) becomes

$$1+1+2r\geq 0\ \ 2+2r\geq 0\ \ r\geq -1. \qquad \ldots (5)$$

Now, considering the negative sign in the expression

$$\frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{x_i-\bar{X}}{\sigma_x}\right)\pm\left(\frac{y_i-\bar{Y}}{\sigma_y}\right)\right]^2 \geq 0 \ ,$$

we have

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i-\bar{X}}{\sigma_x}\right)^2+\frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i-\bar{Y}}{\sigma_y}\right)^2-\frac{2}{N}\sum_{i=1}^{N}\left(\frac{x_i-\bar{X}}{\sigma_x}\right)\left(\frac{y_i-\bar{Y}}{\sigma_y}\right)\geq 0$$

or, $1+1-2r \geq 0$ that is $2-2r \geq 0$ that is, $r \leq 1$. … (6)

Therefore, combining (5) and (6), we get the result

$$-1 \leq r \leq 1. \qquad \ldots (7)$$

which gives the range of the coefficient of correlation.This completes the proof of property 1.

Some specific values of r:

We observe that$r \in [-1,+1]$, this means that depending upon the values of the two variables, it may assume any value in between –1 to +1, both values inclusive. Three important values of r are important to consider in this range. These are –1, 0 and +1.

(i) It assumes value –1, when there is a perfect negative linear relationship between the variables, that is, asone variable increases in its values, the other variable decreases in its values through an exact linear relationship rule.

(ii) Similarly, it assumes value +1 when there is a perfect positive linear relationship between the variables, that is, as one variable increases in its values, the other variable also increases in its values through an exact linear rule.

(iii) The r assumes value zero, when there is no linear relationship between the variables. However, if there exists a linear relationship, then a zero value of r indicates that the variables are not related at all.

(iv) However, other values of r also indicate the extent of the linear relationship.

Being maximum values of r; –1 and +1 on extreme left side and extreme right side of the range respectively and zero at the middle of the range; it seems that r becomes weaker and weaker as it approaches to middle most point of the range from both the sides. Depending upon the values of r in its range, we, therefore, define weak, moderate and strong relationships as follows:

(a) The linear relationship is said to be a "weak positive (negative) relationship" if the value of r lies between 0.0 and 0.3 (–0.0 and - 0.3). This happens when the linear rule, showing the relationship is represented by a shaky linear rule.

(b) Values between 0.3 and 0.7 (–0.3 and –0.7) indicate a "moderate positive (negative) linear relationship" through a fuzzy-firm linear rule.

(c) Lastly, values between 0.7 and 1.0 (–0.7 and –1.0) is an indication of a "strong positive (negative) linear relationship" through a perfect linear rule.

Property 2: The coefficient of correlation, r, is independent of change of origin and scale of the variables. This indicates that if r_{XY}be the coefficient of correlation between two variables X and Y and r_{UV} be the same between the variables U and V which are respectively obtained by changing the origin and scale of the variables X and Y, then we haver$_{UV}$= r_{XY}.

Proof: We know that $r_{XY}= \frac{Cov.(X, Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})}{\sigma_x \sigma_y}$.

Now, let us define the variable $U = \frac{X-a}{b}$, which is obtained by changing the origin and scale of the variable X; constant a changes the origin of X and constant b is used to change the scale of X. Similarly, we define the variable

$V = \frac{Y-c}{d}$,which is obtained by changing the origin and scale of the variable Y. From these relations, we get X= a + bU and Y = c + dV.

Hence, we see that $\bar{X} = a + b\bar{U}$ and $\bar{Y} = c + d\bar{V}$. Therefore, substituting these values of X, Y, $\bar{X}$ and $\bar{Y}$ in the formula of $r_{XY,,}$we have

$$r_{XY} = \frac{\frac{1}{N}\sum_{i=1}^{N}\{b(u_i - \bar{U})\}\{d(v_i - \bar{V})\}}{b\sigma_U . d\sigma_V} \quad \ldots (8)$$

The denominator is obtained due to the results that $\sigma_X = b\sigma_U$ and $\sigma_Y = d\sigma_V$; which are properties of standard deviation(already proved in the unit discussed the measures of dispersion).

Now, (8) becomes

$$r_{XY} = \frac{\frac{1}{N}\, bd\sum_{i=1}^{N}\{(u_i - \bar{U})\}\{(v_i - \bar{V})\}}{bd\sigma_U\sigma_V} = \frac{\frac{1}{N}\sum_{i=1}^{N}(u_i - \bar{U})(v_i - \bar{V})}{\sigma_U\sigma_V};$$

which, by definition, is the coefficient of correlation between variables U and V, denoted by r_{UV}. So, we have the relation $r_{XY} = r_{UV}$. Hence the proof is complete.

This property is very much useful in computingthe coefficient of correlation when X and Y values are large enough in magnitudes. Using suitably chosen constants a, b, c and d; the magnitudes of X and Y values can be reduced in magnitude so that U and V values are comparatively smaller in magnitudes. Thus, the calculations steps may be made simpler.

8.5.2 Calculation of Correlation Coefficient

In this sub-section, we shall describe how to calculate the value of r using different methods of calculations:

(a) Using the direct formula of r:

We shall show here how to calculate the Karl Pearson's Coefficient of Correlation for ungrouped data using the actual formula of r.

The direct (actual) formulaof r given by (2) is

$$r = \text{Corr}(X,Y) = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2\right)\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^2\right)}};$$

which needs the computation of

$\text{Cov}(X, Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y}); V(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2$ and

$V(Y) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^2$ .For a given set of pairs of values (x_i, y_i), these expressions can be obtained in the form of a table. By taking the following example, we will show how these quantities can be calculated:

Example 1: Find the correlation coefficient between advertisement expenditure and profit for the following data:

Advertisement expenditure	30	44	45	43	34	44
Profit	56	55	60	64	62	63

Solution: In order to find the value of r from this formula, it is clear that we need the values of the quantities $\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})$, $\sum_{i=1}^{N}(y_i - \bar{Y})^2$ and $\sum_{i=1}^{N}(x_i - \bar{X})^2$. For this, we first need the values of $\bar{X}$ and $\bar{Y}$. The following table provides rest of the necessary calculations:

Taking Advertisement expenditure as variable X and Profit as variable Y, we see that $\bar{X} = \frac{1}{6}\sum_{i=1}^{6}x_i = \frac{240}{6} = 40$ and $\bar{Y} = \frac{1}{6}\sum_{i=1}^{6}y_i = \frac{360}{6} = 60.$

X	**Y**	$(X - \bar{X})$	$(X - \bar{X})^2$	$(Y - \bar{Y})$	$(Y - \bar{Y})^2$	$(X - \bar{X})(Y - \bar{Y})$
30	56	−10	100	−4	16	40
44	55	4	16	−5	25	−20
45	60	5	25	0	0	0
43	64	3	9	4	16	12
34	62	−6	36	2	4	−12
44	63	4	16	3	9	12
$\sum x_i = 240$	$\sum_{i=1}^{6}y_i = 360$	$\sum_{i=1}^{6}(x_i - \bar{x}) = 0$	$\sum_{i=1}^{6}(x_i - \bar{x})^2 = 202$	$\sum_{i=1}^{6}(y_i - \bar{y}) = 0$	$\sum_{i=1}^{7}(y_i - \bar{y})^2 = 70$	$\sum_{i=1}^{6}(x_i - \bar{x})(y_i - \bar{y}) = 32$

Substituting values from the table in the formula of r, we get

$$r = \frac{32}{\sqrt{202 \times 70}} = \frac{32}{\sqrt{14140}} = \frac{32}{118.91} = 0.27 .$$

Hence, the correlation coefficient between expenditure on advertisement and profit is 0.27. Since the value of r is 0.27, the relationship between the two variables is positive but weak.

(b) Using the extended formula of r:

In the above examples, we used the formula of r as it was originally defined in the formula (2). However, we mentioned a number of times in earlier units that in the process of calculations, we need to avoid terms involving deviations of values of the variables from its respective means, since, calculation of these terms istroublesome on the one hand and generate rounding errors which affect the actual value of the quantity. We, therefore, used formulae for variance, standard deviation and many such terms by expanding the deviation-type terms in order to simplify the calculations as well as minimizing the rounding errors as far as possible. We should develop a formula for it by expanding these terms which would be preferable over the direct formula for calculation purpose. Let us find it.

We have already expanded formulae of standard deviations σ_X and σ_Y under the previous unit discussing measures of dispersion above. These are

$$\sigma_X = \sqrt{\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{X}^2\right)} \qquad \text{and} \qquad \sigma_Y = \sqrt{\left(\frac{1}{N}\sum_{i=1}^{N} y_i^2 - \bar{Y}^2\right)}$$

which are the terms in the denominator of (2).

Now, the numerator of (2) is

$$\begin{aligned}\text{Cov}(X, Y) &= \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{X}\right)\left(y_i - \bar{Y}\right)\\ &= \frac{1}{N}\left[\sum_{i=1}^{N} x_i y_i - \bar{Y}\sum_{i=1}^{N} x_i - \bar{X}\sum_{i=1}^{N} y_i + \sum_{i=1}^{N}\bar{X}\bar{Y}\right]\\ &= \frac{1}{N}\left[\sum_{i=1}^{N} x_i y_i - N\bar{X}\bar{Y} - N\bar{X}\bar{Y} + N\bar{X}\bar{Y}\right];\end{aligned}$$

since, $\sum_{i=1}^{N} x_i = N\bar{X}$, $\sum_{i=1}^{N} y_i = N\bar{Y}$ and $\sum_{i=1}^{N}\bar{X}\bar{Y} - N\bar{X}\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{X}\bar{Y}$.

Thus, the extended formula for r is given by

$$r = \frac{\frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{X}\bar{Y}}{\sqrt{\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{X}^2\right)\left(\frac{1}{N}\sum_{i=1}^{N} y_i^2 - \bar{Y}^2\right)}}. \qquad \ldots (9)$$

It can further be reduced to the formula

$$r = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sqrt{\left\{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2\right\}\left\{N\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2\right\}}}. \qquad \ldots (10)$$

It can be seen that in this formula of r, all the deviation-type terms have been removed and all the terms can be calculated using the exact values of x_i and

y_i. Thus, this form of the formula is suitable for calculation purpose, since, it minimizes the rounding error in each the terms.

Let us now illustrate the steps which should be followed for computing the value of r using the extended formula as obtained in (10). We use the following example for this purpose:

Example3: Calculate Karl Pearson's coefficient of correlation between price and demand for the following data:

Price	17	18	19	20	22	24	26	28	30
Demand	40	38	35	30	28	25	22	21	20

Solution: Since, formula for r is

$$r = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sqrt{\left\{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2\right\}\left\{N\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2\right\}}};$$

therefore, we will compute the values of $\sum_{i=1}^{N} x_i$, $\sum_{i=1}^{N} y_i$, $\sum_{i=1}^{N} x_i y_i$, $\sum_{i=1}^{N} x_i^2$ and $\sum_{i=1}^{N} y_i^2$ from the given values of x_i and y_i. We present the calculations for these terms in the following table:

X	**Y**	**X^2**	**Y^2**	**XY**
17	40	289	1600	680
18	38	324	1444	684
19	35	361	1225	665
20	30	400	900	600
22	28	484	784	616
24	25	576	625	600
26	22	676	484	572
28	21	784	441	588
30	20	900	400	600
$\sum x_i = 204$	$\sum y_i = 259$	$\sum x_i^2 = 4794$	$\sum y_i^2 = 7903$	$\sum x_i y_i = 5605$

Therefore,

$$r = \frac{(9\times 5605) - (204)(259)}{\sqrt{\{(9\times 4794) - (204\times 204)\}\{(9\times 7903) - (259\times 259)\}}}$$

$$r = \frac{50445 - 52836}{\sqrt{(43146 - 41616)\times(71127 - 67081)}}$$

$$r = \frac{-2391}{\sqrt{1530\times 4046}} = \frac{-2391}{2488.0474} = -0.96\,.$$

CHECK YOUR PROGRESS 3

Note: i) Check your answers with those given at the end of the unit.

4) Calculate coefficient of correlation between X and Y for the following data:

X	1	2	3	4	5
Y	2	4	6	8	10

5) Find the coefficient of correlation for the following ages of husband and wife.

Husband's age	23	27	28	29	30	31
Wife's age	18	22	23	24	25	26

8.6 SPEARMAN'S RANK CORRELATION COEFFICIENT

For the calculation of product moment correlation coefficient characters must be measurable. In many practical situations, characters are not measurable. Sometimes, we are given a distribution of items where no numerical measure can be made but where best and worst or most favoured and least favoured can be identified. They are quality characteristics and individuals or items can be ranked in order of their merits. This type of situation occurs when we deal with the qualitative study such as honesty, beauty, voice, etc. Rankings are often applied in these situations to put the series into an order. For example, contestants of a singing competition may be ranked by judge according to their performance. In another example, students may be ranked in different subjects according to their performance in tests.

Arrangement of individuals or items in order of merit or proficiency in the possession of a certain characteristic is called ranking and the number indicating the position of individuals or items is known as rank. If ranks of individuals or items are available for two characteristics then correlation between ranks of these two characteristics is known as rank correlation. If we have a group of individuals ranked according to two different characteristics, it is natural to ask that is there any association between the ranks? To answer this question, we need to use the spearman's rank correlation coefficient.

The formula for the determining the degree of relationship between two ranked characteristics is given as follows:

$$r = 1 - \frac{6\sum d_i^2}{N(N^2 - 1)}$$

where d_i is the difference between the two ranks given to each individual and N is the number of observations.

With the help of rank correlation, we find the association between two quality characteristics. As we know that the Karl Pearson's correlation coefficient gives the intensity of linear relationship between two variables, Spearman's rank correlation coefficient gives the concentration of association between two quality characteristics. In fact, Spearman's rank correlation coefficient measures the strength of association between two ranked variables.

8.6.1 Method of Calculation of Rank Correlation

Let us discuss some problems on rank correlation coefficient. We shall consider three cases to compute the rank correlation coefficient as follows:

Case I: When Actual Ranks are given

Case II: When Ranks are not given

Case III: When Ranks are repeated.

Case I: When Actual Ranks are given

In this case the following steps are involved:

i) Compute $d_i = R_X - R_Y$, i.e., the difference between the two ranks given to each individual

ii) Compute d_i^2, i.e., the squares of differences for each individual.

iii) Apply the formula of the rank correlation coefficient by putting the value of sum of the squares of the difference of the each individuals and N:

Example 4: Suppose we have ranks of 8 students of B.Sc. in Statistics and Mathematics. On the basis of rank we would like to know that to what extent the knowledge of the student in statistics and mathematics is related.

Rank in Statistics	1	2	3	4	5	6	7	8
Rank in Mathematics	2	4	1	5	3	8	7	6

Solution: Let us denote the rank of students in statistics by R_x and rank in mathematics by R_y. For the calculation of rank correlation coefficient we have to find $\sum_{i=1}^{N} d_i^2$ which is obtained through the following table:

Rank inStatistics R_x	**Rank inMathematics** R_y	**Difference of ranks** $d_i = R_x - R_y$	d_i^2
1	2	−1	1
2	4	−2	4
3	1	2	4

4	5	−1	1
5	3	2	4
6	8	−2	4
7	7	0	0
8	6	2	4
			$\sum d_i^2 = 22$

Here, N = number of paired observations =8

$$r_s = 1 - \frac{6\sum d_i^2}{N(N^2-1)} = 1 - \frac{6 \text{ X } 22}{8 \text{ X } 63} = 1 - \frac{132}{504} = \frac{372}{504} = 0.74$$

Thus, there is a positive association between ranks of Statistics and Mathematics.

Case-II: When Ranks are not given

Sometimes not have rank not given but actual values of both variables are available. If we are interested in rank correlation coefficient it is necessary to have ranks so, we a sign ranks to the given values. Ranks can be assigned by taking either the highest value as 1 or the lowest value as 1. The next highest or the next lowest value is given the rank 2 and so on. But whether we start with the highest value or lowest value we must follow the same method in case of both the variables. Considering this case, we are taking a problem and try to solve it.

Example 5: Calculate rank correlation coefficient from the following data:

X	78	89	97	69	59	79	68
Y	125	137	156	112	107	136	124

Solution:

X	**Y**	**Rank of X (R_x)**	**Rank of Y (R_y)**	**$d = R_x - R_y$**	**d^2**
78	125	4	4	0	0
89	137	2	2	0	0
97	156	1	1	0	0
69	112	5	6	-1	1
59	107	7	7	0	0
79	136	3	3	0	0
68	124	6	5	1	1
					$\sum_{i=1}^{N} d_i^2 = 2$

Spearman's Rank correlation formula is

$$r_s = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} = 1 - \frac{6 \times 2}{7(49-1)} = 1 - \frac{12}{7 \times 48} = 0.96$$

Case-III: When Ranks are repeated

In previous section it was assumed that two or more individuals or units do not have same rank. But there might be a situation when two or more individuals have same rank in one or both characteristics, then this situation is said to be tied.

If two or more individuals have same value, in this case common ranks are assigned to the repeated items. This common rank is the average of ranks they would have received if there were no repetition. For example, we have a series 50, 70, 80, 80, 85, 90 then 1st rank is assigned to 90 because it is the biggest value then 2nd to 85, now there is a repetition of 80 twice. Since both values are same so the same rank will be assigned which would be average of the ranks that we would have assigned if there were no repetition. Thus, both 80 will receive the average of 3 and 4 i.e. (Average of 3& 4 i.e. (3+4)/2= 3.5) 3.5 then 5th rank is given to 70 and 6th rank to 50. Thus, the series and ranks of items are

Series:	50	70	80	80	85	90
Ranks:	6	5	3.5	3.5	2	1

In the above example 80 was repeated twice. It may also happen that two or more values are repeated twice or more than that.

For example, in the following series there is a repetition of 80 and 110. You observe the values, assign ranks and check with following:

50	70	80	90	80	120	110	110	110	100
10	9	7.5	6	7.5	1	3	3	3	5

When there is a repetition of ranks, a correction factor $\frac{m(m^2-1)}{12}$ is added to $\sum d^2$ in the Spearman's rank correlation coefficient formula, where m is the number of times a rank is repeated. It is very important to know that this correction factor is added for every repetition of rank in both characters.

In the first example correction factor is added once which is 2(4-1)/12=0.5, while in the second example correction factors are 2(4-1)/12=0.5 and 3(9-1)/12=2 which are aided to $\sum d^2$.

Thus, in case of tied or repeated rank Spearman's rank correlation coefficient formula is

$$r_s = 1 - \frac{6\left\{\sum d^2 + \frac{m(m^2-1)}{12} + \ldots\right\}}{N(N^2-1)}$$

Example 6: Calculate rank correlation coefficient from the following data:

Expenditure on advertisement	10	15	14	25	14	14	20	22
Profit	6	25	12	18	25	40	10	7

Solution: Let us denote the expenditure on advertisement by X and profit by Y

X	**Rank of X (R_X)**	**Y**	**Rank of Y (R_Y)**	**$d = R_x - R_y$**	**d^2**
10	8	6	8	0	0
15	4	25	2.5	1.5	2.25
14	6	12	5	1	1
25	1	18	4	−3	9
14	6	25	2.5	3.5	12.25
14	6	40	1	5	25
20	3	10	6	−3	9
22	2	7	7	−5	25
					$\sum d^2 = 83.50$

$$r_s = 1 - \frac{6\left\{\sum d^2 + \frac{m(m^2-1)}{12} + \ldots\right\}}{N(N^2-1)}$$

Here rank 6 is repeated three times in rank of X and rank 2.5 is repeated twice in rank of Y so the correction factor is

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

Hence, rank correlation coefficient is

$$r_s = 1 - \frac{6\left\{83.50 + \frac{3(3^2-1)}{12} + \frac{2(2^2-1)}{12}\right\}}{8(64-1)} = 1 - \frac{6\left\{83.50 + \frac{3\times 8}{12} + \frac{2\times 3}{12}\right\}}{8\,X\,63}$$

$$r_s = 1 - \frac{6(83.50 + 2.50)}{504} = 1 - 1.024 = -0.024$$

There is a negative association between expenditure on advertisement and profit.

Note: i) Check your answers with those given at the end of the unit.

6) Calculate Spearman's rank correlation coefficient from the following data:

x	20	38	30	40	50	55
y	17	45	30	35	40	25

7) Calculate rank correlation coefficient from the following data:

x	70	70	80	80	80	90	100
y	90	90	90	80	70	60	50

8.7 CONCEPT OF REGRESSION

Before we proceed to discuss the concept and application of regression analysis mainly used for prediction purpose, let us describe first the literal meaning of this word and how it was used for the first time in the subject.

Literally, regression means to return back to a previous and less advanced or less developed or worse state/condition. It is an act of going or coming back to or moving backward to a previous worse or more primitive state or condition.

'Regress' is the opposite of 'progress'. Therefore, to regress is not a good thing in many situations whereas it may be good in other situations. For example, if the economy of a country regresses, it is not a good sign for the economic condition of the country; whereas on the other hand, if the mental performance of mentally retarded patients starts improving, we can hope that they are not regressing, which is a good thing for such patients.

The term "regression" was first coined in the nineteenth century by Sir Francis Galton (1822-1911), a British mathematician, statistician and biometrician. He used this word while he was working on a biological phenomenon regarding the height of parents and their offsprings. In fact, during experimentation he observed the particular biological phenomenon that if parents in families were very tall, their children tended to be tall but shorter than their parents. On the other hand, if parents were very short, their children tended to be short but taller than their parents were. This phenomenon he named as "regression to the mean," with the word "regression" meaning to come back to. Although, Galton used the word regression only for this biological phenomenon, later on, his work was extended by some other statisticians, like, Udny Yule, Karl Pearson and R. A. Fisher to a more general statistical context.

CHECK YOUR PROGRESS 5

Note: i) Use the space given below for your answers.
ii) Check your answers with those given at the end of the unit.

8. What do you understand the word 'regression'?

……

……

……

……

8.8 LINES OF REGRESSION

Let us now discuss the Linear Regression Models. As described above, in linear regression model, we will have some straight-line equations as regression models.

Considering first the simplest case of linear regression, that is, **single variable linear regression model**, our aim here would be to explain the method of finding the best-fitted regression lines which would be the most appropriate straight lines expressing the average relationship between variables.

8.8.1 Single Variable linear Regression Lines

As you know, by single variable regression model we mean in the linear regression, only one independent variable is included other that one dependent variable. Such a regression model may be expressed as $Y = a + bX$ or $Y = b_0 + b_1X_1$, where Y represents the dependent variable or 'output' variable; X or X_1 represent the only independent variable or 'predictor' and a and b or alternatively b_0 and b_1 are the constants (parameters) of the corresponding regression line. For generalization purpose later on, we shall write our model as

$$Y = b_0 + b_1X_1. \quad \ldots (11)$$

i) Regression Line of Y on X_1:

Let us first consider the single variable linear regression model given by equation (11). This model, as explained earlier, is used to estimate (predict) the dependent variable Y on the basis of independent variable X_1, where we assume that variables are linearly related to each other with some degree of correlation. Such a straight-line equation which is used to predict the output variable Y for given predictor values $X_1$is termed as **Regression Line of Y on X_1**. We shall show how we can get the straight line (11) which fits best to a set of pairs of values (y_i, x_{1i}); i = 1, 2, …, N.

As far as the problem of finding the most appropriate line which fits best to a set of data (y_i, x_{1i}); i = 1, 2, …, N is concerned; it seems to be equivalent to obtaining best-fitted straight-line equation using least square principle. We

know that the least square principle provides us a curve/line which when fitted to any set of bivariate data; (y_i, x_{1i}); it passes through maximum number of points and rest of the points which do not lie exactly on the curve/line, the difference between actual and estimated values of the independent variable are minimum. No doubt, thus, least square principle provides us a method of finding most appropriate curve/line which is intuitively a scientifically-justified method and suffices our purpose.

We shall, therefore, use the same principle here for finding the Regression Line of Y on X_1. If we follow the method of least squares, we can write the sum of squares of residuals as

$$U = \sum_{i=1}^{N}\left(y_i - b_0 - b_1 x_{1i}\right)^2 = \sum_{i=1}^{N}\left[y_i - \left(b_0 + b_1 x_{1i}\right)\right]^2 = \sum_{i=1}^{N}\left[y_i - \left(\hat{y}_i\right)\right]^2 \ldots (12)$$

where y_i and $\hat{y}_i$ respectively denote the actual and predicted values of the variable Y.

The two normal equations, obtained by minimizing U simultaneously with respect to parameters b_0 and b_1 are

$$\sum_{i=1}^{N} y_i = Nb_0 + b_1 \sum_{i=1}^{N} x_{1i}; \qquad \ldots (13)$$

$$\sum_{i=1}^{N} y_i x_{1i} = b_0 \sum_{i=1}^{N} x_{1i} + b_1 \sum_{i=1}^{N} x_{1i}^2 \qquad \ldots (14)$$

Multiplying the equation (13) by $\sum_{i=1}^{N} x_{1i}$ and the equation (14) by n and then subtracting the resultant second equation from the resultant first equation, we have

$$\left(\sum_{i=1}^{N} y_i\right)\left(\sum_{i=1}^{N} x_{1i}\right) - N\sum_{i=1}^{N} y_i x_{1i} = b_1\left[\left(\sum_{i=1}^{N} x_{1i}\right)^2 - N\sum_{i=1}^{N} x_{1i}^2\right]$$

$$\text{or, } b_1 = \frac{N\sum_{i=1}^{N} y_i x_{1i} - \left(\sum_{i=1}^{N} y_i\right)\left(\sum_{i=1}^{N} x_{1i}\right)}{N\sum_{i=1}^{N} x_{1i}^2 - \left(\sum_{i=1}^{N} x_{1i}\right)^2} \qquad \ldots (15)$$

This is the estimated value of the parameter b_1, denoted by $\hat{b}_1$, obtained on the basis of given (y_i, x_{1i}) values.

Now, since we know that $\sum_{i=1}^{N} y_i = N\bar{Y}; \sum_{i=1}^{N} x_{1i} = N\bar{X}_1$, we have

$$\hat{b}_1 = \frac{\frac{1}{N}\sum_{i=1}^{N} y_i x_{1i} - \bar{X}_1\bar{Y}}{\frac{1}{N}\sum_{i=1}^{N} x_{1i}^2 - \bar{X}_1^2} = \frac{\mathrm{Cov}(X_1, Y)}{V(X_1)} = \frac{r\sigma_Y\sigma_{X_1}}{\sigma_{X_1}^2}$$

or, $\hat{b}_1 = r\frac{\sigma_Y}{\sigma_{X_1}}$; r being the product moment correlation between X_1 and Y and σ_Y and σ_{X_1}, respectively, the standard deviation of Y and X_1 variables.

Substituting the value of $\hat{b}_1$ in equation (13) and solving for b_0, we get

$$\hat{b}_0 = \bar{Y} - r\frac{\sigma_Y}{\sigma_{X_1}}\bar{X}_1.$$

Finally, substituting these estimated values of constants b_0 and b_1 in the regression model (11), we get

$$\left(Y - \bar{Y}\right) = r\frac{\sigma_Y}{\sigma_{X_1}}\left(X_1 - \bar{X}_1\right). \qquad \ldots (16)$$

This is the single variable linear regression model, including the single independent variable X_1 and the dependent variable Y. It represents the linear relationship between the variables Y and X_1 where the degree of correlation is given by r. The equation can be converted into a linear equation as

$$Y = \left(\bar{Y} - r\frac{\sigma_Y}{\sigma_{X_1}}\bar{X}_1\right) + r\frac{\sigma_Y}{\sigma_{X_1}}X_1 = \hat{b}_0 + \hat{b}_1 X_1; \qquad \ldots (17)$$

where $\hat{b}_0$ and $\hat{b}_1$ are the estimated values of the parameters b_0 and b_1 respectively, obtained through the given data set (x_i, y_i), such that the regression line (17) provides the best fit to the data, in the sense that the predicted values of Y would be closest to its corresponding actual values.

Since, through this regression, we predict the values of the dependent variable Y for given values of the variable X_1; it is popularly termed as the **"Regression Line of Y on X_1"**.

ii) Regression Line of X_1 on Y:

As described above, regression line of Y on X_1 is used to predict the values of the dependent variable Y when values of the independent variable X_1 are given. You may argue that for predicting the values of the X_1 variable, given Y values, the same regression line, (12) can be used if we convert the line in the form

$$X_1 = \frac{1}{\hat{b}_1}Y - \frac{\hat{b}_0}{\hat{b}_1} \quad \text{or} \quad X_1 = b_0^* + b_1^* Y;$$

where $b_0^* = \left(-\frac{\hat{b}_0}{\hat{b}_1}\right)$ and $b_1^* = \frac{1}{\hat{b}_1}$.

However, mathematically it is not incorrect, but statistically this method is not justified from the point of view of least square principle. You know that in order to obtain the best-fitted line using this principle, we use the concept of minimizing the sum of squares of residuals given in (12) as

$$U = \sum_{i=1}^{N}\left[y_i - \left(b_0 + b_1 x_{1i}\right)\right]^2 = \sum_{i=1}^{N}\left[y_i - \left(\hat{y}_i\right)\right]^2.$$

If you observe, in fact, $\left[y_i - \left(\hat{y}_i\right)\right]$s are the vertical distances of points for given X-values. Thus, U is the sum of squares of all such vertical distances of points between the actual and predicted Y-values, which has been minimized with respect to all the parameters.

Now, if instead of Y values, X_1-values are to be predicted for known Y-values, then, according to least square principle, we should minimize the sum

of squares of residuals given by $\left[x_{1i} - (\hat{x}_{1i})\right]$ s which would be horizontal distances between the actual and predicted X_1-values for all the points. Certainly then, we will get a new regression line which will be different than $X_1 = b_0^* + b_1^* Y$, because the method of estimation of parameters b_0 and b_1 in the two methods are quite different.

Now, since we wish to treat X_1-variable as dependent and Y-variable as independent in order to obtain regression line of X_1 on Y using least square principle, we write the single variable linear regression line of X_1 on Y as

$$X_1 = c_0 + c_1 Y \qquad \ldots (18)$$

Using the principle of least squares, the sum of squares of residuals

$$V = \sum_{i=1}^{N}\left[x_{1i} - (c_0 + c_1 y_i)\right]^2 = \sum_{i=1}^{N}\left[x_{1i} - (\hat{x}_{1i})\right]^2$$

is to be minimized with respect to parameters c_0 and c_1. The corresponding two normal equations will be

$$\sum_{i=1}^{N} x_{1i} = Nc_0 + c_1 \sum_{i=1}^{N} y_i; \qquad \ldots (19)$$

$$\sum_{i=1}^{N} y_i x_{1i} = c_0 \sum_{i=1}^{N} y_i + c_1 \sum_{i=1}^{N} y_i^2 \qquad \ldots (20)$$

Finally, solving the above two equations, we get

$$c_1 = \frac{N\sum_{i=1}^{N} y_i x_{1i} - \left(\sum_{i=1}^{N} y_i\right)\left(\sum_{i=1}^{N} x_{1i}\right)}{N\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2} \qquad \ldots (21)$$

This is the estimate of the parameter c_1, say $\hat{c}_1$, as obtained on the basis of the given set of data.

The expression (16) can further be written as

$$\hat{c}_1 = \frac{\mathrm{Cov}(X_1, Y)}{\sigma_Y^2} = \frac{r\sigma_{X_1}\sigma_Y}{\sigma_Y^2} = r\frac{\sigma_{X_1}}{\sigma_Y}.$$

Further, substituting the value of $\hat{c}_1$ in the equation (19) and solving for c_0, we get

$$\hat{c}_0 = \bar{X}_1 - r\frac{\sigma_{X_1}}{\sigma_Y}\bar{Y}$$

Hence, the regression line of X_1 on Y is given by

$$(X_1 - \bar{X}_1) = r\frac{\sigma_{X_1}}{\sigma_Y}(Y - \bar{Y}). \qquad \ldots (22)$$

It can alternatively be written as

$$X_1 = \left(\bar{X}_1 - r\frac{\sigma_{X_1}}{\sigma_Y}\bar{Y}\right) + r\frac{\sigma_{X_1}}{\sigma_Y}Y = \hat{c}_0 + \hat{c}_1 Y \qquad \ldots (23)$$

which is the regression line of X_1 on Y.

Thus, there are always two lines of regression, one, regression line of Y on X_1, for predicting the values of dependent variable Y given the values of independent variable X_1 and the other, regression line of X_1 on Y, for predicting the values of dependent variable X_1 given the values of independent variable Y.

8.8.2 Calculation of Regression Lines

After obtaining the two regression lines (16) and (22), the next step would be to fit these lines to a given set of data (y_i, x_{1i}) for i = 1, 2,.., N; and then consequently to use them for prediction purposes. We observe that for this purpose, we have to find the values of $\bar{Y}$, $\bar{X}_1$, σ_Y, σ_{X_1} and r. Given y_i and x_{1i} values, it is easy to compute the values of means, $\bar{Y}$, $\bar{X}_1$. However, although the values of σ_Y, σ_{X_1} and r can be computed using their respective direct formulae, but we avoid to use them for computation purpose since they are not free from error of approximations. We have mentioned there that alternative formulae for them should be used which are derived from their direct formulae. For readiness of the text in this unit, we present below these formulae which should be used for calculation

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i;\ \bar{X}_1 = \frac{1}{N}\sum_{i=1}^{N} x_{1i};\ r = \frac{\frac{1}{N}\sum_{i=1}^{N} x_{1i}y_i - \bar{X}_1\,\bar{Y}}{\sqrt{\left(\frac{1}{N}\sum_{i=1}^{N} x_{1i}^2 - \bar{X}_1^2\right)\left(\frac{1}{N}\sum_{i=1}^{N} y_i^2 - \bar{Y}^2\right)}};$$

$$\sigma_Y = \sqrt{\frac{1}{N}\sum_{i=1}^{N} y_i^2 - \bar{Y}^2}\ ;\ \sigma_{X_1} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_{1i}^2 - \bar{X}_1^{\,2}}\ .$$

CHECK YOUR PROGRESS 6

Note: i) Use the space given below for your answers.
ii) Check your answers with those given at the end of the unit.

9. What are lines of regression? What are their purposes?

……
……
……
……

10. Explain why there are two lines of regression. Obtain the expression of regression line of Y on X_1 where Y and X_1 are respectively the dependent and independent variables.

……
……
……
……

8.9 REGRESSION COEFFICIENTS

We have seen that the regression line of Y on X_1 is given by

$$Y = \left(\bar{Y} - r\frac{\sigma_Y}{\sigma_{X_1}}\bar{X}_1 \right) + r\frac{\sigma_Y}{\sigma_{X_1}}X_1$$

Where as that of X_1 on Y is

$$X_1 = \left(\bar{X}_1 - r\frac{\sigma_{X_1}}{\sigma_Y}\bar{Y} \right) + r\frac{\sigma_{X_1}}{\sigma_Y}Y .$$

The multipliers $\left(r\frac{\sigma_Y}{\sigma_{X_1}} \right)$ of X_1 in the line of Y on X_1 and $\left(r\frac{\sigma_{X_1}}{\sigma_Y} \right)$ of Y in the line of X_1 on Y are respectively called the "Regression Coefficient of Y on X_1" and "Regression Coefficient of X_1 on Y. Symbolically, these are respectively denoted by letters b_{YX_1} and $b_{X_1Y.}$; that is, $b_{YX_1} = r\frac{\sigma_Y}{\sigma_{X_1}}$,the regression coefficient of Y on X_1 and $b_{X_1Y.} = r\frac{\sigma_{X_1}}{\sigma_Y}$, the regression coefficient of X_1 on Y.

It can be seen that, in fact, regression coefficients give us the slopes of the respective two regression lines. Therefore, regression coefficient $b_{YX_1} = r\frac{\sigma_Y}{\sigma_{X_1}}$ gives the change in the dependent variable Y as a unit change in the independent variable X_1. As for example, if we are considering the production of a crop by Y and amount of rainfall by X_1, then regression coefficient of Y on X_1 represents the change in production of crop which occurs due to a unit change in rainfall.

Similar interpretation can be given for the regression coefficient $b_{X_1Y} = r\frac{\sigma_{X_1}}{\sigma_Y}$ of X_1 on Y.

8.9.1 Properties of Regression Coefficients

Property 1: The geometric mean of the two regression coefficients is the coefficient of correlation r.

Proof: We know that $b_{YX_1} = r\frac{\sigma_Y}{\sigma_{X_1}}$ and $b_{X_1Y} = r\frac{\sigma_{X_1}}{\sigma_Y}$. Therefore, we have

$$b_{YX_1} . b_{X_1Y} = r^2 .$$

This indicates that $r = \pm\sqrt{b_{YX_1} . b_{X_1Y}}$;

that is, the coefficient of correlation, r is the geometric mean of the two regression coefficients. Thus, given the magnitude of regression coefficients, the value of the correlation coefficient can be obtained. However, as far as

the sign of the correlation coefficient is concerned, we can observe that whenever r is negative, both the coefficients b_{YX_1} and b_{X_1Y} would be negative, whereas if r is positive, both the coefficients would be positive. Therefore, we take r positive if both the regression coefficients are positive, otherwise r would be negative.

Property 2: If one of the regression coefficients is greater than one, then other must be less than one.

Proof: Let b_{YX_1} the regression coefficient of Y on X_1 is greater than one, that is,

$$b_{YX_1} > 1 \quad \text{or} \quad \frac{1}{b_{YX_1}} < 1.$$

We know that

$$b_{YX_1} . b_{X_1Y} = r^2 \leq 1 \text{ (From Property 1)}$$

$$\Rightarrow b_{X_1Y} \leq \frac{1}{b_{YX_1}} < 1.$$

Thus, if b_{YX_1} is greater than one then b_{X_1Y} is less than one.

Property 3: Arithmetic mean of the regression coefficients is greater than the correlation coefficient, that is, $\frac{b_{YX_1} + b_{X_1Y}}{2} \geq r$ subject to the condition r > 0.

Proof: Suppose that arithmetic mean of regression coefficients is greater than the correlation coefficient. Then, we have

$$\frac{b_{YX_1} + b_{X_1Y}}{2} \geq r \Rightarrow \left(b_{YX_1} + b_{X_1Y}\right) \geq 2r$$

$$\Rightarrow \left(b_{YX_1} + b_{X_1Y}\right) \geq 2\left(\pm\sqrt{b_{YX_1} . b_{X_1Y}}\right) \qquad \text{since, } r = \pm\sqrt{b_{YX_1} . b_{X_1Y}}$$

Therefore, we have

$$\left(b_{YX_1} + b_{X_1Y}\right) \mp 2\left(\sqrt{b_{YX_1} . b_{X_1Y}}\right) \geq 0$$

$$\Rightarrow \left(\sqrt{b_{YX_1}} \mp \sqrt{b_{X_1Y}}\right)^2 \geq 0,$$

which,being a squared quantity, is always true.Therefore, assuming that

$$\frac{b_{YX_1} + b_{X_1Y}}{2} \geq r;$$

that is, arithmetic mean of regression coefficients is greater than the correlation coefficient, is always true.

Let us now solve some numerical problems on what we have discussed in this unit:

Example 7: Height of fathers and sons in inches are given below:

Height of Father	65	66	67	67	68	69	70	71
Height of Son	66	68	65	69	74	73	72	70

Find two lines of regression and calculate the estimated average height of son when the height of father is 68.5 inches.

Solution: Let us denote the father's height by X_1 and son's height by Y, then for finding the two lines of regression, we do necessary calculations in the following table:

$\mathbf{X_1}$	**Y**	X_1^2	$\mathbf{Y^2}$	$\mathbf{X_1Y}$
65	66	4225	4356	4290
66	68	4356	4624	4488
67	65	4489	4225	4355
67	69	4489	4761	4623
68	74	4624	5476	5032
69	73	4761	5329	5037
70	72	4900	5184	5040
71	70	5041	4900	4970
$\sum x_{1i} = 543$	$\sum y = 557$	$\sum x_{1i}^2 = 36885$	$\sum y_i^2 = 38855$	$\sum x_{1i}y_i = 37835$

From the values obtained in thetable, we have

$$\bar{X}_1 = \frac{543}{8} = 67.88; \quad \bar{Y} = \frac{557}{8} = 69.62.$$

Standard Deviation of $X = \sigma_{X_1} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{1i} - \bar{X}_1)^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}x_{1i}^2 - \bar{X}_1^2}$

$$= \sqrt{\frac{1}{8} \times 36885 - (67.88)^2}$$

$$= \sqrt{4610.62 - 4607.69} = \sqrt{2.93} = 1.71$$

Similarly, Standard deviation of $Y = \sigma_Y = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}y_i^2 - \bar{Y}^2}$

$$= \sqrt{\frac{1}{8} \times 38855 - (69.62)^2}$$

$$= \sqrt{4856.88 - 4846.94} = \sqrt{9.94} = 3.15$$

Now correlation coefficient

$$r = \text{Corr}(X_1, Y) = \frac{N\sum_{i=1}^{N} x_{1i}y_i - (\sum_{i=1}^{N} x_{1i})(\sum_{i=1}^{N} y_i)}{\sqrt{\left\{N\sum_{i=1}^{N} x_{1i}^2 - (\sum_{i=1}^{N} x_{1i})^2\right\}\left\{N\sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2\right\}}}$$

$$= \frac{8 \times 37835 - (543) \times (557)}{\sqrt{\{(8 \times 36885 - (543)^2\}(8 \times 38855) - (557)^2}}$$

$$= \frac{302680 - 302451}{\sqrt{\{(295080 - 294849\}\{310840 - 310249\}}} = \frac{229}{\sqrt{231 \times 591}}$$

$$= \frac{229}{\sqrt{136521}} = \frac{229}{369.49} = 0.62$$

Substituting the value of $\bar{X}_1$, $\bar{Y}$, σ_{X_1}, σ_Y and r in regression equations, we get regression line of Y on X_1 as

$$(Y - 69.62) = 0.62 \times \frac{3.15}{1.71}(X_1 - 67.88)$$

$$Y = 1.14X_1 - 77.38 + 69.62$$

$$Y = 1.14X_1 - 7.76$$

and regression line of X_1 on Y

$$(X_1 - 67.88) = 0.62 \times \frac{1.71}{3.15}(Y - 69.62)$$

$$X_1 = 0.34\,Y - 23.67 + 67.88$$

$$X_1 = 0.34\,Y + 44.21$$

Obviously, estimate of height of son for the height of father = 68.5 inch is obtained by the regression line of Y on X_1

$$Y = 1.14X_1 - 7.76\,.$$

Putting $X_1 = 68.5$ in the above regression line, we have

$$Y = 1.14 \times 68.50 - 7.76 = 78.09 - 7.76 = 70.33$$

Thus, the estimate of son's height for the father's height 68.5 inch is 70.33 inch.

CHECK YOUR PROGRESS 7

Note: i) Check your answers with those given at the end of the unit.

11) We have data on variables X_1 and Y as

X_1	5	4	3	2	1
Y	9	8	10	11	12

Calculate

(i) both the regression coefficients,

(ii) correlation coefficient,

(iii) regression lines of Y on X_1 and X_1 on Y, and

(iv) the estimated value of Y for X_1 =4.5.

8.10 LET US SUM UP

In this unit, we discussed

1. the concept of correlationbetween two variables and its importance in the simultaneous analysis of them;
2. the meanings of different types of correlationsuch as, linear correlation, multiple correlation, partial correlation, non-linear correlation;
3. scatter diagramand how and why it is helpful to have an approximate idea about the trend of the relationship between the variables;
4. how the formula of Karl Pearson's coefficient of correlationis defined and how the formula for computing the coefficient is obtained from it;
5. the different properties of correlation coefficient, like, the range of it, some significant values of it, effect of change of origin and scale on it;
6. What the word "regression" means and in what sense it was introduced inthe subject of Statistics;
7. What is the importance and meaning of regression analysis in Statistics;
8. How the two single variable linear regression models are theoretically obtained using least squares principle. What are regression lines of Y on X_1 and X_1 on Y. How the two regression lines can be applied to any set of data for predicting the dependent variable;
9. What are the important properties of regression lines;
10. What do we mean by regression coefficients in the regression analysis and what is the significance of these coefficients; and
11. What are some of the important properties of regression coefficients and how these are useful in order to calculate some parameters of the regression models.

8.11 KEY WORDS

Scatterplot : A special graph containing a cluster of dots that represents all pairs of observations.

Pearson Correlation Coefficient (r) : A number between –1.00 and 1.00 that describes the linear relationship between pairs of quantitative variables.

8.12 SUGGESTED FURTHER READING/ REFERENCES

Witte, R., & Witte, J. (2017). *Statistics*. Hoboken, NJ: John Wiley & Sons.

8.13 ANSWERS TO CHECK YOUR PROGRESS

1) Please refer to section 8.3.

2) Please refer to section 8.3.1.

3) Please refer to section 8.5.

4) First, we form the table showing needed calculations.

X	**Y**	$(X-\bar{X})$	$(X-\bar{X})^2$	$(Y-\bar{Y})$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
1	2	–2	4	–4	16	8
2	4	–1	1	–2	4	2
3	6	0	0	0	0	0
4	8	1	1	2	4	2
5	10	2	4	4	16	8
15	30	0	10	0	40	20

Here $\sum x_i = 15 \Rightarrow \bar{X} = 3$ and $\sum y_i = 30 \Rightarrow \bar{Y} = 6$

From the calculation table we observe that

$$\sum(x_i - \bar{X})^2 = 10,\ \sum(y_i - \bar{Y})^2 = 40 \text{ and } \sum(x_i - \bar{X})(y_i - \bar{Y}) = 20$$

Substituting these values in the formula

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{X})^2 \sum_{i=1}^{N}(y_i - \bar{Y})^2}} = \frac{20}{\sqrt{10 \times 40}} \frac{20}{20} = 1$$

Hence, there is perfect positive correlation between X and Y.

5) Let us denote the husband's age as X and wife's age by Y

X	**Y**	**X**	$\mathbf{Y^2}$	**XY**
23	18	529	324	414
27	22	729	484	594
28	23	784	529	644
29	24	841	576	696
30	25	900	625	750
31	26	961	676	806
$\sum x_i = 168$	$\sum y_i = 138$	$\sum x_i^2 = 4744$	$\sum y_i^2 = 3214$	$\sum x_i y_i = 3904$

We use the formula

$$r = \frac{(6\times 3904)-(168)(138)}{\sqrt{\{(6\times 4744)-(168\times 168)\}\{(6\times 3214)-(138\times 138)\}}} = \frac{23424\text{-}23184}{\sqrt{240\times 240}} = 1$$

Hence there is perfect positive correlation between X and Y.Perfect positive correlation between X and Y is because of the perfect linear relationship, given by Y = X – 5, exists.

6) First, we form the table showing needed calculations.

X	**Y**	**Rank of x (R_x)**	**Rank of y (R_y)**	**d = R_X - R_Y**	**d^2**
20	17	6	6	0	0
38	45	4	1	3	9
30	30	5	4	1	1
40	35	3	3	0	0
50	40	2	2	0	0
55	25	1	5	−4	16
					$\sum_{i=1}^{n} d_i^2 = 26$

$$\Rightarrow r_s = 1 - \frac{6 \text{ X } 26}{6(36-1)} = 1 - \frac{26}{35} = \frac{9}{35} = 0.26$$

7) First, we form the table showing needed calculations.

x	**Rank of x (R_x)**	**y**	**Rank of y (R_y)**	**d = R_x-R_y**	**d^2**
70	6.5	90	2	4.5	20.25
70	6.5	90	2	4.5	20.25
80	4	90	2	2	4
80	4	80	4	0	0
80	4	70	5	−1	1
90	2	60	6	−4	16
100	1	50	7	−6	36
					$\sum d^2$ = 97.5

Here, rank 4 and 6.5 is repeated thrice and twice respectively in rank of X and rank 2 is repeated thrice in rank of Y, so the correction factor is

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

and therefore, rank correlation coefficient is

$$r_s = 1-\frac{6\left\{97.5+\frac{3(3^2-1)}{12}+\frac{2(2^2-1)}{12}+\frac{3(3^2-1)}{12}\right\}}{7(49-1)} = 1-\frac{6(102)}{336} = -0.82$$

8) Please refer to section 8.6.

9) Please refer to section 8.7.

10) Please refer to section 8.7.1.

11) For getting necessary values for computation purpose, we present the calculations in the following table:

$$\bar{X}_1 = \frac{\sum x_{1i}}{n} = \frac{15}{5} = 3 \text{ and } \quad \bar{Y} = \frac{\sum y_i}{n} = \frac{50}{5} = 10$$

S. No.	X_1	**Y**	$(X_1-\bar{X}_1)$	$(X_1-\bar{X}_1)^2$	$(Y-\bar{Y})$	$(Y-\bar{Y})^2$	$(X_1-\bar{X}_1)(Y-\bar{Y})$
1	5	9	2	4	–1	1	–2
2	4	8	1	1	–2	4	–2
3	3	10	0	0	0	0	0
4	2	11	–1	1	1	1	–1
5	1	12	–2	4	2	4	–4
Total	15	50	0	10	0	10	–9

(i) Regression coefficient of X_1 on Y,

$$b_{X_1Y} = \frac{\sum(x_{1i}-\bar{X}_1)(y_i-\bar{Y})}{\sum(y_i-\bar{Y})^2} = \frac{-9}{10} = -0.9$$

Regression coefficient of Y on X_1,

$$b_{YX_1} = \frac{\sum(x_{1i}-\bar{X}_1)(y_i-\bar{Y})}{\sum(x_{1i}-\bar{X}_1)^2} = \frac{-9}{10} = -0.9\,.$$

(ii) $r = \pm\sqrt{b_{YX_1} \times b_{X_1Y}} \Rightarrow \pm\sqrt{(-0.9)\times(-0.9)} = -0.90$

Note: We are taking "-"sign because correlation coefficient and regression coefficients have same sign.

(i) To find regression lines we need $\bar{Y}, \bar{X}_1$ b_{YX_1} and b_{X_1Y}. From the calculation in table above we have

Regression line of Y on X_1 is $(Y-10) = -0.90(X_1 - 3)$ and

Regression line of X_1 on Y is $(X_1 - 3) = -0.90\ (Y-10)$

(i) To estimate Y we use regression line of Y on X_1 which is

$$(Y-10) = -0.9\ (X_1 - 3)$$

putting X_1 =4.5 we get $(Y-10) = -0.9\ (4.5-3) \Rightarrow Y = 8.65$

UNIT 9 SAMPLING DISTRIBUTIONS

Structure

9.1 INTRODUCTION

In general, extracting information from all the elements or items of a large population may be time consuming and expensive if the population size is infinitely large. But even then sometimes there are so many problems attached to large size populations where it becomes necessary to draw inference about population parameters.For example, one may wish to estimate the average height of all the two thousand students in a college, a businessman may be interested to estimate the proportion of defective items in a production line, a manufacturer of car tyres may want to estimate the variations took place in the diameter of produced tyres, a pharmacist may want to estimate the difference of effect of two types of drugs, etc.In all such cases, there is always an unknown population involved whose characteristics are described through some parameters.

Due to large sizes of populations, in which we may be interested, for drawing inferences about the population parameters, generally we draw a sample and determine a function of sample values, which is called statistic.

Selection of a sample, that is, only a part of the population saves a lot of time, money and labour and results drawn on sample values are quicklyavailable forinterpretation and as good sometimes as obtained on the basis of the entire population. The process of generalising sample results to the population is called Statistical Inference. Sincethere might be a large number of samples of same size drawn from the population, the values of statistic generally vary from sample to sample and is associated with the probability of selection of the particular sample. Therefore, the sample statistic is a random variable following any probability distribution. Therefore, it may be a matter of interest for a statistician to know what distribution the statistic follows if the samples are assumed to be selected from a theoretical distribution like, a normal distribution with given mean and variance; a binomial distribution; a gamma distribution; a Poisson distribution and so on. In contrast to theoretical distributions, probability distribution of a statistic in popularly called a sampling distribution. In this unit we shall discuss the sampling distribution of sample mean; of sample median; of sample proportion; of difference between two sample means and sample proportions.

Due to this curiosity, Prof. R.A. Fisher, Prof. G. Snedecor and some other statisticians worked in this area and obtained exact sampling distributions which are followed by some of the important statistics. In present unit of this block, we shall also discuss some important sampling distributions such as χ^2 (read as chi-square), t and F. Generally, these sampling distributions are named on the name of the originator, for instance, Fisher's F-distribution is named like this after the name of its inventor Prof. R.A. Fisher.

9.2 OBJECTIVES

After studying this unit, you should be able to:

- explain the concept of sampling and sampling distribution;
- explain the concept of standard error and Central Limit Theorem;
- define the sampling distribution; and
- describe the sampling distribution of sample mean and difference of two sample means.

9.3 BASICS OF SAMPLING

Before discussing sampling distributions of different kinds of statistic, in this section we shall be discussing basic concepts and definitions of some of the important terms which areverymuch helpful in understanding the fundamentals of statistical inference and are frequently used in this unit.

Population, Sample and Sampling

Literally, "population" means a well-defined group (collection or bunch) of some objects (units or elements). Examples of populations and its units are : a city with some clear-cut territory having its dwellers as units (elements) or

having houses as its units; a hospital with indoor patients as units or its doctors as units; a library with its employees as units or books as its units; a river with its fishes as units; a school with enrolled students as units; a banyan tree with leaves as units; sky with stars as units and so on. Thus, population is the collection or group of individuals /items /units/ observations under study. The total number of elements / items / units / observations in a population is known as its size and generally denoted by N.

A finite subset of units of a population is called a "**sample"** and the number of units belonging to the sample is called the sample size. If n denotes the sample size, then necessarily we have the condition $n < N$ and then the sample is said to be a proper sample. However, in some situations, we have $n = N$, that is, all the units of the population are selected as sample. As discussed above, in almost all kinds of statistical studies, a sample is selected for inferential purposes because of the reasons stated above. However, even if the population is not large enough or it has a well-defined territory or it is not of destructive type; a sample is useful since besides it is always less costly, less time consuming, it helps us to get a handy result in very short time.

A sample of size n is not selected simultaneously, rather n units are selected one by one in n draws in order to constitute a sample. The process through which units of the population are selected in different draws to constitute a sample of specified size, is known as **"sampling"**. Thus, sampling is a draw-to-draw mechanism of selecting some units from a population. Sampling is mainly of two types, namely,

(a) Probability Sampling or Random Sampling, and

(b) Non-Probability Sampling.

The technique of random sampling is of fundamental importance in the application of statistics. The Estimation theory is based on the assumption of random sampling. It is a scientific method of selecting samples accordingly to some laws of chance in which each unit in the population has some definite pre assigned probability of being selected in the sample.Simple Random Sampling, Stratified Sampling, Systematic Random sampling, etc., are the examples of random sampling.

In non-probability sampling, the sample is selected with definite purpose and, hence, the choice of sampling units depends entirely on the discretion and judgment of the investigator. It is, therefore, not a scientific method and attracts severe criticisms in many circumstances. Purposive sampling, Judgement Sampling, etc., are the examples of non-probability sampling.

We know that if $n < N$, it is a sampling procedure. When $n = N$ the process is known as **"complete enumeration or census"**. An example of complete enumeration is the decennial census of India.

Need for Sampling in Statistical Inference

As we have discussed in Section 9.1, in many practical and real situations the population under consideration is either infinitely large in size or it is unbounded in the sense that its boundaries are not well-defined or it is of destructive nature. As for example, in ordert to determine average life of two hundred produced electric bulbs, it would be necessary to light these bulbs unless they all get fused, resulting destruction of the entire lot of production. Similarly, in order to estimate the proportion of smokers in a large city, information have to be gathered from each and every person dwelling in the city which would be a very difficult task in terms of manpower, money and time to be required. Sometimes, the population size is not known if its boundaries are not well-defined, like number of fish in a pond or lake. In all such cases, it is not feasible to gather information on all the units, rather it becomes almost impossible. Keeping in view the difficulty in contacting each and every unit of the population due to these reasons and also to save the time, money and manpower to be required for this, generally a part of the population, popularly known as a "**sample"** is selected in some pre-asigned manner which in turn is used for drawing the inferences on the population itself or on the parameters of the population. The results obtained from the sample are projected in such a way that they are valid for concluding about the entire population. Therefore, the sample works like a **"Vehicle"** to reach (drawing) at valid conclusions about the population. In fact, a sample helps us to reach to the **"whole"**(population) from a "**part"** (sample) in all types of statistical studies. Thus, statistical inference can be defined as:

"It is a process of concluding (projecting or infering) something desired about a given population on the basis of sample results".

Parameter and Statistic

A **parameter** is a function of population values which is used to represent certain characteristics of the population.For example, population total, population mean, population variance, population coefficient of variation, population proportion, population correlation coefficient, etc., are all parameters since their calculation involve all the population values.

A statistic is a function of sample values only and does not contain any unknown population parameter in it. For example, if $X_1, X_2, ..., X_n$ represent values of the variable X in a random sample of size n taken from a population then sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a statistic.

Estimator and Estimate

A statistic which is used to guessor to estimate (infer) an unknown population parameter then it is known as **estimator** and the value of the estimator based on observed value of the sample is known as **estimate** of

parameter. For example, suppose the parameter λ of the Poisson populationf(x, λ) is unknown. If we use sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, calculated on the basis of sample values $X_1, X_2, ..., X_n$ to estimate λ then $\bar{X}$ is an estimator and any particular value of $\bar{x}$ is called estimateof parameter λ.

9.4 SAMPLING DISTRIBUTION

As discussed under Section 9.1, a statistic is calculated on the basis of a sample of specific size, selected from the given population for drawing the inference about the population. But due to the fact that sample size is always smaller than the population size; (n < N), a number of samples of same size can theoretically be drawn from the population (theoretically ${}^{n}C_{N}$ possible samples). Thus, in fact, we have in all ${}^{n}C_{N}$ values of the statistic. Moreover, since in probability sampling, each sample is selected with some pre-defined probability of selection, it means that all these statistic values have some probability to appear. In this sense, we can say that a statistic $\hat{\theta}$ is a random variable defined as $\{\hat{\theta}_k, p(\hat{\theta}_k)\}$, where $\hat{\theta}_k$ and $p(\hat{\theta}_k)$ are respectively the value of the statistic $\hat{\theta}_k$ and the probability of its appearance due to the k^{th} sample; k = 1, 2,…, ${}^{n}C_{N}$.

Generally, in practice onlya single random sample is taken from a given population and its mean $\bar{X}$ is considered to be representative of the population mean μ. This sample mean may or may not represent the population mean. Since we cannot determine the proximity of sample mean and population mean on the basis of single random sample, we can use the concept of sampling distribution to bring the value of sample mean close to that of Population mean. Being a random variable, the statistic must possess a probability distribution which may be used to answer some questions regarding the nature of the statistic, for example, what is the probability that its value as obtained from the given sample is differing from the parameter value by amargin of 10? The probability distribution of a statistic is called sampling distribution of the statistic. Let us illustrate how the sampling distribution of a statistic can be obtained given a population. For instance, let we wish to estimate the population mean using the sample mean as an estimator.

Suppose that a baby-sitter has 5 children under her supervision. The ages of children are 2, 4, 6, 8 and 10 years. Supposing this group of children as a population of size 5, we get the population mean as

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{2+4+6+8+10}{5} = 6$$

Therefore, the variance of this population is given by:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2 = \frac{(2-6)^2 + (4-6)^2 + ... + (10-6)^2}{5} = 8$$

Now, let us take all the possible simple random samples of size 2 without replacement from this population. There are ${}^5C_2 = 10$ such possible samples which are listed below along with their respective means:

Table 9.1: Possible samples

Sample No.	Sample	Sample Mean
1	2, 4	3
2	2, 6	4
3	2, 8	5
4	2, 10	6
5	4, 6	5
6	4, 8	6
7	4, 10	7
8	6, 8	7
9	6, 10	8
10	8, 10	9

Now, suppose that our selected sample is either (2, 4) with mean 3 or it is the sample (8, 10) with mean 9. In both cases, the sample is not a good representative of the population since sample mean is far from the population mean 6. But if the selected samples are coincidently either (2, 10) or (4, 8), then these are good representatives of the population in the sense that both have sample means exactly equal to population mean. Thus, this example illustrates that a single random sample may or may not be representative for the decision maker to reach meaningful conclusion. However, the grand mean of the distribution of these ten sample means, is calculated and observed to be equal to the population mean as follows:

$$\bar{\bar{x}} = \frac{1}{10}\sum_{i=1}^{10}\bar{x}_i = \frac{3+4+5+6+5+6+7+8+9}{10} = 6.$$

Hence, the mean of the sample means can be considered to represent the population mean for analysis and decition making purposes.

Now let us put sample means along with their probabilities of occurrence as follows:

Table 9.2: Probability Distribution

Sample Mean	Frequency	Probability of Occurrence
3	1	0.1
4	1	0.1
5	2	0.2
6	2	0.2
7	2	0.2
8	1	0.1
9	1	0.1
Total	**10**	**1.0**

This distribution which shows the distribution of probabilities over all the possible values of the sample mean is refered to as sampling distribution of the sample mean. Symbolically, it can be denoted as $\{(\bar{x}, p(\bar{x})\}$. As explained above, the sampling distribution of a statistic can be defined as:

"The probability distribution of all possible values of a statistic that would be obtainedby drawing all possible samples of the same size from the population is called sampling distribution of that statistic."

The mean, variance and other measures of a sampling distribution can be obtained in a similar way as computed in a frequency distribution, taking probabilities as frequencies. Thus, mean of the above sampling distribution will be

$$\bar{\bar{x}} = \sum_{i=1}^{7} \bar{x}_i p(\bar{x}_i) = 3x0.1 + 4x0.1 + 5x0.2 + 6x0.2 + 7x0.1 + 8x0.1 + 9x0.1 = 6$$

This value is same as the population meanμ. The variance of the distribution is obtained as:

$$\sigma^2_{\bar{x}} = \frac{1}{7}\sum_{i=1}^{N}(\bar{x}_i - \bar{\bar{x}})^2 = \frac{(3-6)^2 + (4-6)^2 + ... + (9-6)^2}{7} = \frac{28}{7} = 4$$

9.4.1 Standard Error

The expression obtained above, obviously measures the variability of the sample mean around the actual mean, that is, how much samplestatistic may vary from sample to sample.

This is equivalent to population variance which measures the deviation of the population values around the population mean. If we consider the positive square root of this , it would be equivalent to standard deviation. In order to

differentiate with population standard deviation, it is called as "**standard error" (SE)** of that statistic. Thus, standard error ofa statistic can be defined as:

"The standard deviation of a sampling distribution is known as standard error".

The computation of the standard error is a tedious process. However, the formula for calculating standard error of sample mean on the basis of a sample size n is seen always to be

$$SE\left(\bar{X}\right) = \frac{\sigma}{\sqrt{n}},$$

where σstands for the standard deviation of the population.

Standard errors of some otherwell known statistics are given below:

1. The SE of sample proportion (p)is given by

$$SE(p) = \sqrt{\frac{PQ}{n}}$$

where, P is population proportion, n is sample size and Q = 1-P.

2. The SE of difference of two sample means is given by

$$SE\left(\bar{X} - \bar{Y}\right) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where, σ_1^2 and σ_2^2 are the population variances of two different populations and n_1 and n_2 are the sample sizes of two independent samples selected from the two populations respectively.

3. The SE of difference between two sample proportions is given by

$$SE(p_1 - p_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$$

where, P_1 and P_2 are population proportions in two different populations from which samples of sizes n_1 and n_2 are taken respectively; $Q_1 = 1-P_1$ and $Q_2 = 1-P_2$.

From all the above formulae, we can understand that standard error is inversely proportional to the sample size. Therefore, as sample size increases the standard error decreases.

The standard error is used to express the accuracy or precision of the estimate of population parameter because the reciprocal of the standard error is the measure of reliability or precision of the statistic.Standard error also determines the probable limits or confidence limitswithin which the

population parameter may be expected to lie with certain level of confidence.Standard error is also applicable in testing of hypothesis.

9.4.2 Central Limit Theorem

The central limit theorem is the most important theorem of Statistics. It was first introduced by De Movers in the early eighteenth century. The theorem states that regardless of the nature of the distribution of the population, the distribution of the sample mean approaches the Normal Probability distribution as the sample size increases. In general, the larger the sample size, the closer the proximity of the distribution of sample mean to the Normal distribution. However, in practice, sample sizes of 30 or larger are considered adequate for this purpose. It should be noted, however, that the sampling distribution of sample mean would be always normally distributed if the original population is normally distributed.

Therefore, according to the central limit theorem, if $X_1, X_2, ..., X_n$ is a random sample of size n taken from a population with mean μ and variance σ^2 thenthe sampling distribution of the sample mean tends to normal distribution with mean μ and variance σ^2/n as sample size tends to large $(n > 30)$ whatever be the form of parent population, that is,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

and the variate

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

follows normal distribution with mean 0 and variance unity, that is, the variate Z follows standard normal distribution.

9.5 SAMPLING DISTRIBUTION OF STATISTICS

Sample mean is one of the most commonly used statistic in any of the statistical studies in order to study the nature and charactristics of a given population. Examples are average income in a locality in a social survey, average life of produced items in a manufacturing process, average temperature in a day in meterological survey, etc. Oweing to these reasons, it is important to obtain the sampling distribution of sample mean.

9.5.1 Sampling Distribution of Sample Mean

In the previous section we have elaborated the meaning of sampling distribution of the sample mean and presented an example how it can be obtained. On the basis of these, the sampling distribution of sample mean can be defined as follows: “The probability distribution of sample mean that would be obtainedby drawing all possible samples of same size from the

population is called sampling distribution of sample mean or simply of mean."

If $X_1, X_2, ..., X_n$ is an independent random sample of size n taken from a normalpopulation with mean μ and variance σ^2 then it has been established that sampling distribution of sample mean $\bar{X}$ is also normal. The mean and variance of sampling distribution of $\bar{X}$ can be obtained as

$$\text{Mean of } \bar{X} = E(\bar{X}) = E\left[\frac{X_1 + X_2 + ... + X_n}{n}\right] \quad [\text{By defination of } \bar{X}]$$

$$= \frac{1}{n}[E(X_1) + E(X_2) + ... + E(X_n)]$$

Since $E(X_i)$ = mean of X_i for all i=1, 2,…,n; therefore, we have $E(X_i) = \mu$. Similarly, $Var(X_i) = \sigma^2$ for all i. We, therefore, have

$$E(\bar{X}) = \frac{1}{n}[\mu + \mu + \mu + ... + \mu(\text{n times})] = \frac{n\mu}{n} = \mu$$

and variance

$$Var(\bar{X}) = Var\left[\frac{1}{n}(X_1 + X_2 + ... + X_n)\right]$$

$$= \frac{1}{n^2}[Var(X_1) + Var(X_2) + ... + Var(X_n)]$$

$$= \frac{1}{n^2}[\sigma^2 + \sigma^2 + \sigma^2 + ... + \sigma^2(\text{n times})] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

We, therefore, conclude that If $X_i \sim N(\mu, \sigma^2)$ then

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

and

$$SE(\bar{X}) = SD(\bar{X}) = \sqrt{Var(\bar{X})} = \frac{\sigma}{\sqrt{n}}$$

Let us illustrate these results using some examples.

Example 1: Diameter of a steel ball bearing produced on a semi-automatic machine is known to be distributed normally with mean 12 cm and standard deviation 0.1 cm. If we take a random sample of 10 ball bearings then find mean and variance of sampling distribution of mean.

Solution: Here, we are given that

$\mu = 12$, $\sigma = 0.1$, $n = 10$

Since the sample is taken from the population of ball bearings in which diameter follows normal distribution N(12, 0.01), we have

$$E(\bar{X}) = \mu = 12$$

$$Var(\bar{X}) = \frac{\sigma^2}{n} = \frac{(0.1)^2}{10} = 0.001.$$

Thus, $\bar{X}$ ~N(12, 0.001).

Example 2: If ages of 5 employees of Account Section of a Company are 58, 66, 64, 62 and 50 years then construct the sampling distribution of average age of employees by taking all samples of size 2 with replacement.

Solution: Here, we are given that N= 5,n = 2

Hence the population mean:

$$\bar{\bar{X}} = \frac{54 + 56 + 60 + 64 + 66}{5} = 60$$

Therefore, the variance of this population is given by:

$$\sigma^2 = \frac{(54-60)^2 + (56-60)^2 + ... + (66-60)^2}{5} = \frac{104}{5} = 20.08$$

Now, all possible samples (without replacement) are 10 which are shown in the Table given below:

Table 9.3: Possible Samples

Sample Number	Sample Observation	Sample Mean
1	54, 56	55
2	54, 60	57
3	54, 64	59
4	54, 66	60
5	56, 60	58
6	56, 64	60
7	56, 66	61
8	60, 64	62
9	60, 66	63
10	64, 66	65

We, therefore, have the grand mean of the distribution as

$$\bar{\bar{x}} = \frac{1}{10}\sum_{i=1}^{10} \bar{x}_i = \frac{55 + 57 + 58 + ... + 62 + 63 + 65}{10} = 60$$

which is same as the population mean. Now, let us construct sampling distribution of the sample mean as given under:

Table 9.4: Probability Distribution

Sample Mean	Frequency	Probability of Occurrence
55	1	0.1
57	1	0.1
58	1	0.1
59	1	0.1
60	2	0.2
61	1	0.1
62	1	0.1
63	1	0.1
65	1	0.1
Total	**10**	**1.0**

Therefore, the mean of sample means can be obtained by the formula:

$$\bar{\bar{x}} = \sum_{i=1}^{7} \bar{x}_i p(\bar{x}_i) = (55x0.1) + (57x0.1) + (58x0.1) + (59x0.1) + (60x0.2)$$
$$+ 61x0.1 + 62x0.1 + 63x0.1 + 65x0.1 + = 60$$

This value is same as the population meanμ. The variance of the distribution is obtained as:

$$\sigma^2_{\bar{x}} = \frac{1}{9}\sum_{i=1}^{N}(\bar{x}_i - \bar{\bar{x}})^2 = \frac{(55-60)^2 + (57-60)^2 + ... + (65-60)^2}{9} = \frac{78}{9} = 8.66$$

CHECK YOUR PROGRESS 1

Note: i) Check your answers with those given at the end of the unit.

1) If lives of 5 televisions of certain company are 4, 6, 8, 10 and 12 years then construct the sampling distribution of average life of televisions by taking all samples of size 2.

2) The weight of certain type of a truck tyre is known to be distributed normally with mean 200 pounds and standard deviation 4 pounds. A random sample of 10 tyres is selected. What is the sampling distribution of sample mean? Also obtain the mean and variance of this distribution.

3) The mean life of CFL blubs produced by a company is 2550 hours. A random sample of 100 CFL bulbs is selected and the standard deviation is found to be 54 hours. Find the mean and variance of the sampling distribution of mean.

9.5.2 Sampling Distribution of Difference Between Two Sample Means

Instead of sample mean of a single population, sometimes one may be interested into two populations and, hence, in finding the sampling distribution of the difference of sample means which are obtained on the basis of the samples drawn from two populations. Such cases arise when we wish to compare average lives of elecrtric bulbs of two kinds or to compare the efficiency of two different drugs of same disease or to compare the average marks of two different sections in a school, etc.

Let the same characteristic measured on two populations, say population-I and population-II be represented by X and Y variables. Suppose population-I have mean μ_1 and variance σ_1^2 whereas population-II have mean μ_2 and variance σ_2^2. Then,let $\bar{X}$ be the sample mean based on a sample of size n_1 selected from population-I and $\bar{Y}$ be the sample mean based on a sample of size n_2 selected from the population-II. As before, it is obvious that, if necessary, one may select all possible samples of sizes n_1 and n_2 respectively from the two populations. Then considering all possible differences of means and sampling distribution of difference of population means can be obtained. The sampling distribution of difference of sample means can be defined as:

"The probability distribution of all values of the difference of two sample means would be obtained by drawing all possible samples from both the populations. Such a distribution is called sampling distribution of difference of two sample means."

If both the parent populations are normal, that is,

$$X \sim N(\mu_1, \sigma_1^2) \text{ and } Y \sim N(\mu_2, \sigma_2^2)$$

then as we discussed in previous section

$$\bar{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \text{ and } \bar{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

If two independent random variables X and Y are normally distributed then the difference $(X - Y)$ also normally distributed. Therefore, the sampling distribution of difference of two sample means $(\bar{X} - \bar{Y})$ also follows normal distribution with mean

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_1 - \mu_2$$

and variance

$$\text{Var}\left(\bar{X}-\bar{Y}\right)=\text{Var}\left(\bar{X}\right)+\text{Var}\left(\bar{Y}\right)=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$$

Therefore, standard error of difference of two sample means is given by

$$\text{SE}\left(\bar{X}-\bar{Y}\right)=\sqrt{\text{Var}\left(\bar{X}-\bar{Y}\right)}=\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\,.$$

Let us see an application of the sampling distribution of difference of two sample means with the help of an example.

Example 3: LED Bulbs manufactured by company A have mean lifetime of 2400 hours with standard deviation 200 hours, while LED Bulbs manufactured by company B have mean lifetime of 2200 hours with standard deviation of 100 hours. If random samples of 125 LED Bulbs of each company are tested, find. The mean and standard error of the sampling distribution of the difference of mean lifetime of LED Bulbs.

Solution: Here, we are given that

$\mu_1 = 2400$, $\sigma_1 = 200$, $\mu_2 = 2200$, $\sigma_2 = 100$ and $n_1 = n_2 = 125$

Let $\bar{X}$ and $\bar{Y}$ denote the mean lifetime of CFLs taken from companies A and B respectively. Since n_1 and n_2 are large ($n_1, n_2 > 30$) therefore, by central limit theorem, the sampling distribution of $(\bar{X}-\bar{Y})$ follows normal distribution with mean

$$E(\bar{X}-\bar{Y})=\mu_1-\mu_2=2400-2200=200$$

and variance

$$\text{Var}(\bar{X}-\bar{Y})=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}=\frac{(200)^2}{125}+\frac{(100)^2}{125}=400$$

Therefore, the standard error is given by

$$\text{SE}\left(\bar{X}-\bar{Y}\right)=\sqrt{\text{Var}\left(\bar{X}-\bar{Y}\right)}=\sqrt{400}=20\,.$$

Now, continuing our discussion, the sampling distribution of difference of two means, we consider another situation.

If population variances σ_1^2 and σ_2^2 are unknown then we estimate σ_1^2 and σ_2^2 by the values of the sample variances of the samples taken from the first and second population respectively. For large sample sizes n_1 and $n_2\,(>30)$, the sampling distribution of $(\bar{X}-\bar{Y})$ is very closely normally distributed with

mean $(\mu_1-\mu_2)$ and variance $\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)$.

If population variances σ_1^2 and σ_2^2 are unknown and $\sigma_1^2 = \sigma_2^2 = \sigma^2$ then σ^2 is estimated by pooled sample variance s_p^2 where,

$$s_p^2 = \frac{1}{n_1 + n_2 - 2}\left[n_1 s_1^2 + n_2 s_2^2\right]$$

and variate

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

follows t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom. Similar to sampling distribution of mean, for large sample sizes n_1 and n_2 (> 30) the sampling distribution of $(\bar{X} - \bar{Y})$ is very closely distributed as normal with mean $(\mu_1 - \mu_2)$ and variance $s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$.

CHECK YOUR PROGRESS 2

Note: i) Check your answers with those given at the end of the unit.

4) The average height of Male workers in a hospital is found to be 68 inches with a standard deviation of 2.3 inches whereas the average height of Female workers in a hospital is found to be 65 inches with a standard deviation of 2.5 inches. If a sample of 35 Male and 50 Female mean and standard error of the sampling distribution of the difference between workers are selected at random, find is the the sample means of height of Male workers and female workers.

9.5.3 Sampling Distribution of Sample Proportion

In Section 9.5.1, we have discussed the sampling distribution of sample mean if some quantitative characteristic is taken into consideration..But in many real word situations,when some qualitatrive characteristic is under consideration, sample proportion, instead of sample mean is computed on the basis of a sample and hence, sampling distribution of sample proportion is needed. Examples of cases in which sample proportions are needed for analysis are proportion of male births; proportion of defective items in manufacturing process; proportion of cancer cases in population, etc.

For sampling distribution of sample proportion, we need sample proportion p. Let a sample of size n is taken from the population, then p is given by

$$p = \frac{X}{n} \leq 1$$

where X is the number of observations /individuals / items / units in the sample which have the particular characteristic under study. For better

understanding of the process, we consider the following example in which size of the population is very small:

Suppose, there is a lot of 4 cartons A, B C and D of electric Tubes and each carton contains 20 Electric Tubes. The number of defective bulbs in each carton is given below:

Table 9.4: Number of Defective Bulbs per Carton

Carton	Number of Defectives Bulbs
A	2
B	4
C	1
D	3

The population proportion of defective tubes can be obtained as

$$P = \frac{2+4+1+3}{20+20+20} = \frac{10}{80} = \frac{1}{8}$$

Now, let us assume that we do not know the population proportion of defective tubes. So we decide to estimate population proportion of defective tubes on the basis of samples of size n = 2. There are ${}^{N}C_{n} = {}^{4}C_{2} = 6$ possible samples of size 2 with replacement. All possible samples and their respective proportion defectives are given in the following table:

Table 9.6: Calculation of Sample Proportion

Sample	Sample Carton	Sample Observation	Sample Proportion(p)
1	(A, B)	(2, 4)	6/40
2	(A, C)	(2, 1)	3/40
3	(A,D)	(2, 3)	5/40
4	(B, C)	(4, 1)	5/40
5	(B, D)	(4, 3)	7/40
6	(C, D)	(1, 3)	4/40

From the above table, we can see that value of the sample proportion is varying from sample to sample. So we consider all possible sample proportions and calculate their probability of occurrence. Since there are 6 possible samples therefore the probability of selecting a sample is 1/8. Then we arrange the possible sample proportions with their respective probability in Table 2.7:

Table 9.7: Sampling Distribution of Sample Proportion

S.No.	Sample Proportion(p)	Frequency	Probability
1	3/40	1	1/6
2	4/40	1	2/6
3	5/40	2	2/6
4	6/40	1	1/6
5	7/40	1	1/6
Total			1.00

This distribution is called the sampling distribution of sample proportion.

The mean of sampling distribution of sample proportion can be obtained as

$$\bar{p} = \frac{1}{K}\sum_{i=1}^{k} p_i f_i \quad \text{where, } K = \sum_{i=1}^{k} f_i$$

$$\frac{1}{6}\left(\frac{3}{40}\times 1 + \frac{4}{40} + 1 + \frac{5}{40}\times 2 + \frac{6}{40}\times 1 + \ldots + \frac{7}{40}\times 1\right) = \frac{30}{40}\times\frac{1}{6} = \frac{1}{8}$$

Thus, we have seen that mean of sample proportion is equal to the population proportion.

If a population whose elements are divided into two mutually exclusive groups– one containing the elements which possess a certain attribute and other containing elements which do not possess the attribute, then number of successes (elements possess a certain attribute) follows a binomial distribution with mean

$$E(X) = nP$$

and variance

$$Var(X) = nPQ \qquad \text{where} \quad Q = 1 - P$$

where, P is the probability or proportion of success in the population.

Now, we can easily find the mean and variance of the sampling distribution of sample proportion by using the above expression as

$$E(p) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}nP = P$$

and variance

$$Var(p) = Var\left(\frac{X}{n}\right) = \frac{1}{n^2}Var(X) \qquad \left[\because Var(aX) = a^2 Var(X)\right]$$

$$= \frac{1}{n^2}nPQ = \frac{PQ}{n}. \quad \left[\because Var(X) = nPQ\right]$$

Also standard error of sample proportion can be obtained as

$$SE(p) = \sqrt{Var(p)} = \sqrt{\frac{PQ}{n}}$$

If the sampling is done without replacement from a finite population then the mean and variance of sample proportion is given by

$$E(p) = P$$

and variance

$$Var(p) = \frac{N-n}{N-1}\frac{PQ}{n}$$

where, N is the population size and the factor (N-n) / (N-1) is called finite population correction.

If sample size is sufficiently large, such that np > 5 and nq > 5 then by central limit theorem, the sampling distribution of sample proportion p is approximately normally distributed with mean P and variance PQ/n where, Q= 1□ P.

Let us see an application of the sampling distribution proportion with the help of an example.

Example 4: A machine produces a large number of items of which 15% are found to be defective. If a random sample of 200 items is taken from the population and sample proportion is calculated then find mean and standard error of sampling distribution of proportion.

Solution: Here, we are given that

$$P = \frac{15}{100} = 0.15,\ n = 200$$

We know that when sample size is sufficiently large, such that np > 5 and nq > 5 then sample proportion p is approximately normally distributed with mean P and variance PQ/n where, Q = 1– P. But here the sample proportion is not given so we assume that the conditions of normality hold, that is, np > 5 and nq > 5. So mean of sampling distribution of sample proportion is given by

$$E(p) = P = 0.15$$

and variance

$$Var(p) = \frac{PQ}{n} = \frac{0.15 \times 0.85}{200} = 0.0006$$

Therefore, the standard error is given by

$$SE(p) = \sqrt{Var(p)} = \sqrt{0.0006} = 0.025$$

CHECK YOUR PROGRESS 3

Note: i) Check your answers with those given at the end of the unit.

5) A state introduced a policy to give loan to unemployed doctors to start own clinic. Out of 10000 unemployed doctors 7000 accept the policy and got the loan. A sample of 100 unemployed doctors is taken at the time of allotment of loan. Find the mean and standard error of the sampling distribution of proportion of doctors who accepted the policy and got the loan.

9.5.4 Sampling Distribution of Difference of Two Sample Proportions

Just like the sampling distribution of difference of two population means, which has been obtained in sub-Section 9.5.2, the sampling distribution of difference of two population proportions can be obtained which is required many times in statistical inference. We shall show how it can be obtained.

Suppose, there are two populations, say, population-I and population-II under study and the proportions of some attribute in populations I and II are respectively P_1 and P_2. For finding sampling distribution of difference of two sample proportions, let samples of sizes n_1 and n_2 be selected respectively from populations I and II and the sample proportions of the attribute in these samples are p_1 and p_2 respectively. Since sample proportions obviously vary from sample to sample, their difference will be a random variable having some probability distribution, which would be termed as the sampling distribution of difference of two sample proportions. On the basis of arguments made for one sampling proportion in the previous section, for the distribution of sampling proportion p due to central limit theorem, here also we can assume that if $n_1p_1 > 5$, $n_1q_1 > 5$, $n_2p_2 > 5$ and $n_2q_2 > 5$ then

$$p_1 \sim N\left(P_1, \frac{P_1Q_1}{n_1}\right) \text{ and } p_2 \sim N\left(P_2, \frac{P_2Q_2}{n_2}\right)$$

where, $Q_1 = 1 - P_1$ and $Q_2 = 1 - P_2$.

Also, by the property of normal distribution, the sampling distribution of the difference of sample proportions follows normal distribution with mean

$$E(p_1-p_2) = E(p_1)-E(p_2) = P_1-P_2$$

and variance

$$Var(p_1-p_2) = Var(p_1)+Var(p_2) = \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}$$

That is,

$$p_1 - p_2 \sim N\left(P_1 - P_2, \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}\right)$$

Thus, standard error is given by

$$SE(p_1 - p_2) = \sqrt{Var(p_1 - p_2)} = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$$

Thus, p_1-p_2 follows a normal distribution with above mentioned mean, variance and standard error. Let us see an application of the sampling distribution of difference of two sample proportions with the help of an example.

Example 5: In one population, 30% persons had hair colour Black and in second population 20% had the same hair colour. A random sample of 200 persons is taken from each population independently and calculate the sample proportion for both samples, then find the mean and variance of the sampling distribution of the difference between two sample proportions.

Solution: Here, we are given that

$P_1 = 0.30$, $P_2 = 0.20$, $n_1 = n_2 = 200$

Let p_1 and $p_2$be the sample proportions of blue-eye persons in the samples taken from both the populations respectively. We know that when n_1 and n_2 are sufficiently large, such that $n_1p_1 > 5$, $n_1q_1 > 5$, $n_2p_2 > 5$ and $n_2q_2 > 5$ then sampling distribution of the difference between two sample proportions is approximately normally distributed. But here the sample proportions are not given so we assume that the conditions of normality hold. So mean of sampling distribution of the difference between two sample proportions is given by

$$E(p_1 - p_2) = P_1 - P_2 = 0.30 - 0.20 = 0.10$$

and variance

$$Var(p_1 - p_2) = \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2} = \frac{0.30 \times 0.70}{200} + \frac{0.20 \times 0.80}{200} = 0.0019$$

Thus, standard error

$$\text{Standard Error} = (p_1 - p_2) = \sqrt{Var(p_1 - p_2)} = \sqrt{0.0019} = 0.04$$

CHECK YOUR PROGRESS 4

Note: i) Check your answers with those given at the end of the unit.

6) In city A, 25% persons were found to be smokers and in another city B, 20% persons were found smokers. If 250 persons of city A and 200 persons of city B are selected randomly, then find the mean and standard deviation of sampling distribution of the difference in sample proportions.

9.6 EXACT SAMPLING DISTRIBUTION

As we have discussed in Section 9.1, some of the well known statisticians, like Prof. R. A. Fisher, Prof. G. Snedecor, etc. worked on finding some of the statistic and determined the exact sampling distributions their properties and applications in differenct areas. These sampling distributions are named on the nmame of its originator for example, F- distribution is named as Fisher's F-distribution and t-distribution as student's t-distribution on the name of Prof. W.S. Gosset. Before describing the Exact Sampling distribution first we will discuss the term "Degree of Freedom" a very useful concept which is necessarily to be understood before discussing and understanding the concepts of Exact sampling distributions. The exact sampling distributions are described with the help of degrees of freedom.

Degrees of Freedom (df)

The term degree of freedom (df) is related to the independency of sample observations. In general, the number of degree of freedom is the total number of observations minus the number of independent constraints or restrictions imposed on the observations. For example, let $x_1, x_2,\ldots, x_n$ be n independent observations in a sample. Unless some condition is imposed on these x values, it would have n df. Now, let one condition $x_1+x_2+\ldots+ x_n = 100$ be imposed on this set, then it looses 1 df, that is, now df will be n – 1 since the last value x_n or any other value x_i will be dependent on all other remaing values and therefore, the number of independent values will be n-1. Further, let $x_1^2+x_2^2+\ldots+ x_n^2 = 4000$ be the another condition imposed, then now the df will be n – 2, etc.

For a sample of n observations, if there are k restrictions among observations (k < n), then the degrees of freedom will be (n–k).

9.6.1 Chi-square Distribution

The chi-square distribution was first discovered by Helmert in 1876 and later independently explained by Karl- Pearson in 1900. The chi-square distribution was discovered mainly as a measure of goodness of fit of any model on the given frequency distribution or probability distribution. .

If a random sample $X_1, X_2,\ldots, X_n$ of size n is drawn from a normal population having mean μ and variance σ^2 then the sample variance can be defined as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ or } \sum_{i=1}^{n}(x_i - \bar{x})^2 = (n-1)s^2 = \nu s^2$$

where, $\nu = n - 1$; the symbol ν read as 'nu'.

Then, the variate $\chi^2 = \frac{\nu s^2}{\sigma^2}$ which is the ratio of sample variance multiplied by its degrees of freedom and the population variance follows the χ^2-distribution with ν degrees of freedom.

The probability density functionof χ^2-distribution with ν df is given by

$$f\left(\chi^2\right) = \frac{1}{2^{\nu/2}\left|\frac{\nu}{2}\right.} e^{-\chi^2/2}\left(\chi^2\right)^{(\nu/2)-1}; \qquad 0 < \chi^2 < \infty \qquad \ldots (1)$$

where, ν = n −1.

It can be seen that χ^2-distribution is a sampling distribution of a statistic $\chi^2 = \frac{\nu s^2}{\sigma^2}$, the shape of probability distribution of which is shown by probability curve shown in Fig. 3.1 for n = 1, 4, 10 and 22.

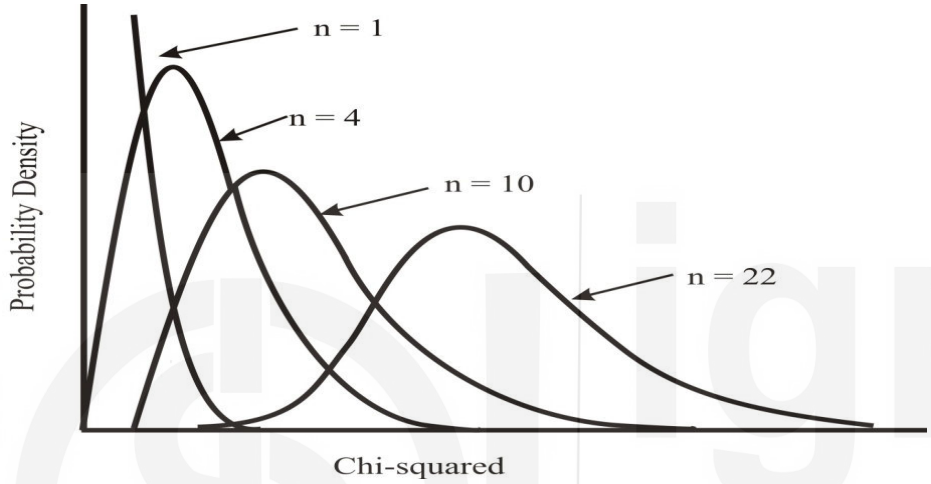


Fig. 9.1: Chi-square probability curves for n = 1, 4, 10 and 22

After looking the probability curves of χ^2-distribution for n = 1, 4, 10 and 22, one can understand that probability curve takes shape of inverse J for n = 1. The probability curve of chi-square distribution gets skewed more and more to the right as n becomes smaller and smaller. It becomes more and more symmetrical as n increases because as n tends to ∞, χ^2-distribution converges to a normal distribution. It is apparent from the figure that even when n= 22, it is tending to a symmetrical probability curve which is an essential property of a normal curve.

Example 8: What are the mean and variance of chi-square distribution with 5 degrees of freedom?

Solution: The mean and variance of chi-square distribution with n degrees of freedom are given by

Mean = n and Variance = 2n

In our case, n = 5, therefore,

Mean = 5 and Variance = 10.

Some of the salient features of χ^2-distribution are mentioned here without giving proofs of each:

1. The probability curve of the chi-square distribution lies in the first quadrant because the range of χ^2-variate is from 0 to ∞.
2. Chi-square distribution has only one parameter n, that is, the degrees of freedom.
3. Chi-square probability curve is highly positive skewed for smaller values of n but becomes a symmetrical curve for larger values of n.
4. Chi-square-distribution is a uni-modal distribution, that is, it has single mode.
5. The mean and variance of chi-square distribution with n df are n and 2n respectively.

The applications of chi-square distribution are very wide in Statistics. The chi-square distribution is used (i) to test the hypothesis that whether the population variance is same as the specified value or not in parametric test procedures, (ii) to test the goodness of fit, that is, to judge whether there is a discrepancy between theoretical and experimental observations or not and (iii) to test the independence of two attributes. The applications listed above shall be discussed in detail in Unit 4 of this block.

CHECK YOUR PROGRESS 5

Note: i) Check your answers with those given at the end of the unit.

7) What are the mean and variance of chi-square distribution with 10 degrees of freedom?

8) What are the mean and variance of chi-square distribution with pdf given below

$$f\left(\chi^2\right) = \frac{1}{96} e^{-\chi^2/2} \left(\chi^2\right)^3 ; \qquad 0 < \chi^2 < \infty$$

9) List the applications of chi-square distribution.

9.6.2 Student's t-Distribution

The t-distribution was discovered by W.S. Gosset in 1908. He was better known by the pseudonym **'Student'** and hence t-distribution is called **'Student's t-distribution'**.

If a random sample $X_1, X_2, \ldots, X_n$ of size n is drawn from a normal population having mean μ and variance σ^2 then we know that the sample mean $\bar{X}$ is distributed normally with mean μ and variance σ^2/n, that is, if $X_i \sim N\left(\mu, \sigma^2\right)$ then $\bar{X} \sim N\left(\mu, \sigma^2/n\right)$. Then as mentioned under sub-section 2.3.2, the variate

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is distributed normally with mean 0 and variance 1, that is, $Z \sim N(0, 1)$.

In general, the standard deviation σ is not known and in such a situation the only alternative left is to estimate it from a sample. The value of sample variance (S^2) is used to estimate it where,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2$$

But then in this case the variate $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ is not normally distributed whereas it follows t-distribution with (n−1) df, that is,

$$t = \frac{\bar{X} - \mu}{s\sqrt{n}} \sim t_{(n-1)} \qquad \ldots (2)$$

The t-variate is a widely used variable and its distribution is called student's t-distribution on the pseudonym name 'Student' of W.S. Gosset. The probability density function of variable t with (n-1) = ν degrees of freedom is given by

$$f(t) = \frac{1}{\sqrt{\nu}\,B\left(\frac{1}{2}, \frac{\nu}{2}\right)\left(1 + \frac{t^2}{\nu}\right)^{(\nu+1)/2}}; \qquad -\infty < t < \infty \qquad \ldots (3)$$

where, $B\left(\frac{1}{2}, \frac{\nu}{2}\right)$ is known as beta function.

The probability curve of t-distribution is bell shaped and symmetric about t = 0 line. The probability curves of t-distribution is shown in Fig. 9.2 at two different values of degrees of freedom n = 4 and 12.

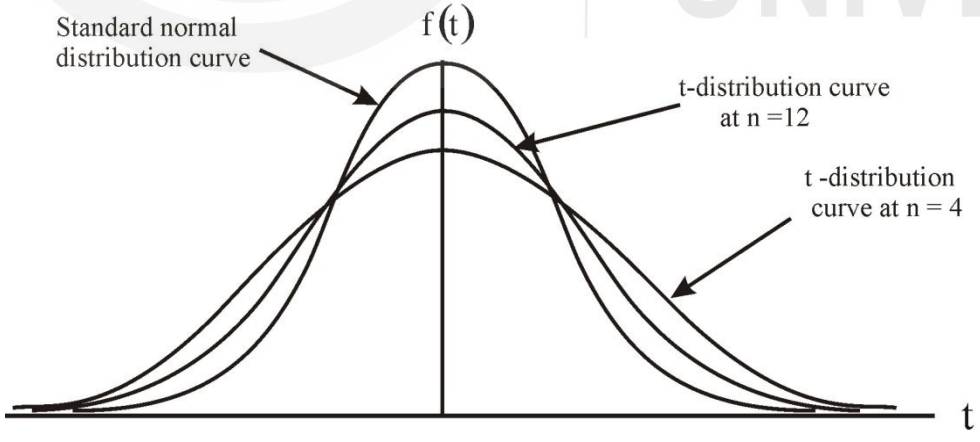


Fig. 9.2: Probability curves for t-distribution at n = 4, 12 along with standard normal curve

In the figure given above, we have drawn the probability curves of t-distribution at two different values of degrees of freedom along with probability curve of standard normal distribution. By looking at the figure, one can easily understand that the probability curve of t-distribution is similar

in shape to that of normal distribution and asymptotic to the horizontal-axis whereas it is flatter than standard normal curve. The probability curve of the t-distribution is tending to the normal curve as the value of n increases. Therefore, for sufficiently large value of sample size n, practically for(> 30), the t-distribution tends to the normal distribution.

The mean and variance of the t-distribution with n df are given by

$$\text{Mean} = 0 \text{ and Variance } = \frac{n}{n-2} \qquad n > 2$$

Now, we shall discuss some of the important properties of the t-distribution.The t-distribution has the following properties:

1. The t-distribution is a uni-modal distribution, that is, t-distribution has single mode.
2. The mean and variance of the t-distribution with n df are zero and $\frac{n}{n-2}$ respectively. The variance of the distribution exists only if n > 2.
3. The probability curve of t-distribution is similar in shape to the standard normal distribution and is symmetric about t = 0 line but flatter than normal curve.
4. The probability curve is bell shaped and asymptotic to the horizontal axis.

Above we have discussed some important properties of t-distribution without mentioning their proof. You may be interested to know the applications of the t-distribution. The t-distribution has wide number of applications in Statistics. The t-distribution is used (i) to test the hypothesis about the significance of a population mean, (ii) the the hypothesis of equality of two population means of two normal populations and (iii) to test the hypothesis whether the population correlation coefficient is zero or not.

Example 9: The life of light bulbs manufactured by the company A is known to be normally distributed. The CEO of the company claims that an average life time of the light bulbs is 300 days. A researcher randomly selects 25 bulbs for testing the life time and he observed the average life time of the sampled bulbs is 290 days with standard deviation of 50 days. Calculate value of t-variate.

Solution: Here, we are given that

$$\mu = 300, N = 25, \overline{X} = 290 \text{ and } s = 50$$

The value of t-variate can be calculated by the formula

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

Therefore, we have

$$t = \frac{290 - 300}{50/\sqrt{25}} = \frac{-10}{10} = -1$$

Note: i) Check your answers with those given at the end of the unit.

10) The scores on an IQ test of the students of a class are assumed to be normally distributed with a mean of 60. From the class, 15 students are randomly selected and an IQ test of the similar level is conducted. The average test score and standard deviation of test scores in the sample group are found to be 65 and 12 respectively. Calculate the value of t-statistic.

11) What are the mean and variance of t-distribution with 8 degrees of freedom?

12) Write any three applications of t-distribution.

9.6.3 F-Distribution

As we have mentioned in previous unit, F-distribution was introduced by Prof. R. A. Fisher and defined as the ratio of two independent chi-square variates when divided by their respective degrees of freedom. If we draw a random sample $X_1, X_2, ..., X_{n_1}$ of size n_1 from a normal population with mean μ_1 and variance σ_1^2 and another independent random sample $Y_1, Y_2, ..., Y_{n_2}$ of size n_2 from another normal population with mean μ_2 and variance σ_2^2 respectively then $\nu_1 s_1^2 / \sigma_1^2$ is distributed as chi-square variate with ν_1 df, that is,

$$\chi_1^2 = \frac{\nu_1 s_1^2}{\sigma_1^2} \sim \chi^2_{(\nu_1)} \quad \ldots (1)$$

where, $\nu_1 = n_1 - 1, \overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$ and $s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2$

Similarly, $\nu_2 s_2^2 / \sigma_2^2$ is distributed as chi-square variate with ν_2 df, that is,

$$\chi_2^2 = \frac{\nu_2 s_2^2}{\sigma_1^2} \sim \chi^2_{(\nu_2)} \quad \ldots (2)$$

where, $\nu_2 = n_2 - 1, \overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} X_i$ and $s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2$

Now, if we take the ratio of the above chi-square variates given in equations (1) and (2), then we get

$$\frac{\chi_1^2}{\chi_2^2} = \frac{\nu_1 s_1^2 / \sigma_1^2}{\nu_2 s_2^2 / \sigma_2^2} \Rightarrow \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2} = \frac{\chi_1^2 / \nu_1}{\chi_2^2 / \nu_2} \sim F_{(\nu_1, \nu_2)} \quad \ldots (3)$$

In the above expression F stands for the F-distribution. In the suffix, υ_1 and υ_2 are called the degrees of freedom of the F-distribution.

Now, if variances of both the populations are equal, that is, $\sigma_1^2 = \sigma_2^2$, then F-variate is written in the form of ratio of two sample variances which is as follows:

$$F = \frac{s_1^2}{s_2^2} \sim F_{(\nu_1, \nu_2)} \qquad \ldots (4)$$

The probability density function of F-variate is given as

$$f(F) = \frac{(\nu_1/\nu_2)^{\nu_1/2}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \frac{F^{(\nu_1/2)-1}}{\left(1+\frac{\nu_1}{\nu_2}F\right)^{(\nu_1+\nu_2)/2}}; \qquad 0 < F < \infty \qquad \ldots (5)$$

As shown in (5), F-variate varies from 0 to ∞, therefore, it is always positive so probability curve of F-distribution wholly lies in the first quadrant. ν_1 and ν_2 are the degrees of freedom and are called the parameters of F-distribution. Hence, the shape of probability curve of F-distribution depends on ν_1 and ν_2. Probability curves of F-distribution taking (ν_1, ν_2) as (5, 5), (5, 20) and (20, 5) are shown in Fig.2.3 below:

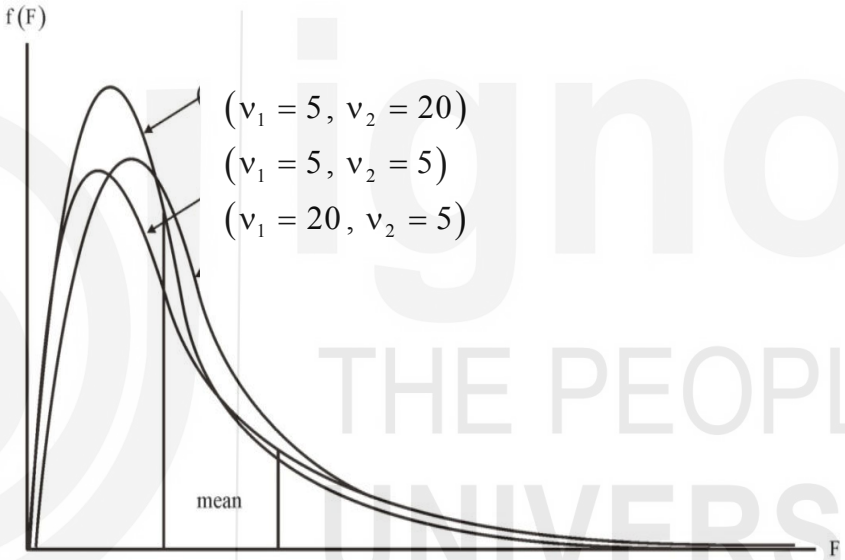


Fig. 9.3: Probability curves of F-distribution for (5, 5), (5, 20) and (20, 5) degrees of freedom.

As it appears from the figure, F-distribution is uni-modal curve. It can be seen from the figure that by increasing the first degrees of freedom from $\nu_1 = 5$ to $\nu_1 = 20$ the mean of the distribution (shown by vertical line) does not change but probability curve shifs from the tail to the centre of the distribution whereas increasing the second degrees of freedom from $\nu_2 = 5$ to $\nu_2 = 20$ the mean of the distribution (shown by vertical line) decrease and the probability curve shifts from the tail to the centre of the distribution. One can also get an idea about the skewness of the F-distribution. We observe from the probability curve that it is positively skewed curve and it becomes very highly positive skewed if ν_2 becomes small. Now, we shall discuss some of the important properties of F-distribution.

The F-distribution has the following important properties:

1. The probability curve of F-distribution is positively skewed curve. The curve becomes highly positive skewed when ν_2 is smaller than ν_1.
2. F-distribution curve extends on abscissa from 0 to ∞.
3. F-distribution is a uni-modal distribution, that is, it has single mode.
4. The square of t-variate with ν df follows F-distribution with 1 and ν degrees of freedom.
5. The mean of F-distribution with (ν_1,ν_2) df is $\frac{\nu_2}{\nu_2-2}$ for $\nu_2 > 2$.
6. The variance of F-distribution with (ν_1,ν_2) df is

$$\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)} \text{ for } \nu_2 > 4.$$

7. If we interchange the degrees of freedom ν_1 and ν_2 then there exists a very useful relation as

$$F_{(\nu_1,\nu_2),(1-\alpha)} = \frac{1}{F_{(\nu_2,\nu_1),\alpha}}$$

F-distribution has a lot of applications in Statistics. It is used to test the hypothesis of equality of variances of two normal populations, for the significance of multiple correlation coefficients, correlation ratio in the population. It is also used in one-way and two-way analysis of variance for testing the equality of several means at a time.

Example 10: A statistician selects 7 women randomly from the population of women, and 12 men from a population of men. The table given below shows the standard deviation of each sample and population:

Population	**Population Standard Deviation**	**Sample Standard Deviation**
Women	40	45
Men	80	75

Compute the value of F-variate.

Solution: Here, we are given that

$$n_1 = 7,\ n_2 = 12, \sigma_1 = 40,\ \sigma_2 = 80, s_1 = 45, s_2 = 75$$

The value of F-variate can be calculated by the formula given below

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

where, s_1^2 & s_2^2 are the sample variances.

Therefore, we have

$$F = \frac{(45)^2/(40)^2}{(75)^2/(80)^2} = \frac{1.27}{0.88} = 1.93$$

For the above calculation, the degrees of freedom ν_1 for women's data are7−1= 6 and the degrees of freedom ν_2 for men's data are 12 −1 =11.

CHECK YOUR PROGRESS 7

Note: i) Check your answers with those given at the end of the unit.

13) For the purpose of a survey 15 students are selected randomly from class A and 10 students are selected randomly from class B. At the stage of the analysis of the sample data, the following information is available:

Class	**Population Standard Deviation**	**Sample Standard Deviation**
Class A	65	60
Class B	45	50

Calculate the value of F-variate.

14) Write any five properties of F-distribution.

15) What are the mean and variance of F-distribution with $\nu_1 = 5$ and $\nu_2 = 12$ degrees of freedom?

16) Write four applications of F-distribution.

We now end this unit by giving a summary of what we have covered in it.

9.7 LET US SUM UP

In this unit, we have covered the following points:

1. The statistical procedure which is used for drawing conclusions about the population parameter on the basis of the sample data is called **"statistical inference"**.
2. A group of units or items under study is known as **"Population"** whereas a part or a fraction of it is known as **"sample"**.
3. A "**parameter"** is a function of population values which is used to represent certain characteristics of the population and any quantity calculated from sample values does not contain any unknown population parameter is known as "**statistic"**.
4. Any statistic used to estimate an unknown population parameter is known as "**estimator"** and the particular value of the estimator is known as "**estimate".**

5. The probability distribution of any statistic is called "**sampling distribution"**of that statistic.
6. The standard deviation of the sampling distribution of a statistic is known as **"standard error"**.
7. The most important theorem of Statistics is **"central limit theorem"** which states that the sampling distribution of the sample mean tends to normal distribution as sample size n tends to large (generally when n > 30).
8. The sampling distribution of sample mean and difference between two sample means.
9. The sampling distribution of sample proportion and difference of two sample proportions.
10. The properties, probability curve and applications of χ^2, t and F distributions, and
11. Mean and variance of χ^2, t and F distributions.

9.8 KEY WORDS

Standard Error of the Mean : A rough measure of the average amount by which sample means deviate from the population mean.

Sampling Distribution of the Mean : The probability distribution of means for all possible random samples of a given size from some population.

9.9 SUGGESTED FURTHER READING/ REFERENCES

Witte, R., & Witte, J. (2017). *Statistics*. Hoboken, NJ: John Wiley & Sons.

9.10 ANSWERS TO CHECK YOUR PROGRESS

1) Here, we are given that

N=5, n=2

Since we have to estimate the average life of Televisions on the basis of samples of size n = 2 therefore, all possible samples (with replacement) are ${}^{N}C_{N} = {}^{5}C_{2} = 10$ and for each sample we calculate the sample meanas shown in Table 2.8 given below:

Table 9.8: Calculation of Sample Mean

Sample Number	Sample Observation	Sample Mean ($\bar{X}$)
1	4, 6	5
2	4, 8	6
3	4,10	7
4	4, 12	8
5	6, 8	7
6	6, 10	8
7	6, 12	9
8	8, 10	9
9	8, 12	10
10	10, 12	11

Since the arrangement of all possible values of sample mean with their corresponding probabilities is called the sampling distribution of mean, thus,we arrange every possible value of sample mean with their respective probabilities in the following Table 9.9 given below:

Table 9.9: Sampling distribution of sample means

S.No.	Sample Mean ($\bar{X}$)	Frequency	Probability
1	5	1	1/10
2	6	1	1/10
3	7	2	2/10
4	8	2	2/10
5	9	2	2/10
6	10	1	1/10
7	11	1	1/10

Here, we are given that

$$\mu = 200, \sigma = 4, n = 10$$

2) Since parent population is normal so sampling distribution of sample means is also normal. Therefore, the mean of this

distribution is given by

$$E\left(\bar{X}\right) = \mu = 200$$

and variance

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} = \frac{(4)^2}{10} = \frac{16}{10} = 1.6$$

3) Here, we are given that

$\mu = 2550$, $n = 100$, $s = 54$

First of all, we find the sampling distribution of sample mean. Since sample size is large ($n = 100 > 30$) therefore, by the central limit theorem, the sampling distribution of sample mean follows normal distribution. Therefore, the mean of this distribution is given by

$$E(\bar{X}) = \mu = 2550$$

and variance

$$\text{Var}(\bar{X}) = \frac{s^2}{n} = \frac{(54)^2}{100} = \frac{2916}{100} = 29.16$$

4) Here, we are given that

$\mu_1 = 68$, $\sigma_1 = 2.3$, $n_1 = 35$

$\mu_2 = 65$, $\sigma_2 = 2.5$, $n_2 = 50$

To find the mean and standard error, first of all we find the sampling distribution of difference of two sample means. Let $\bar{X}$ and $\bar{Y}$ denote the mean height of male and female workers of hospital, respectively. Since n_1 and n_2 are large ($n_1, n_2 > 30$) therefore, by the central limit theorem, the sampling distribution of $(\bar{X} - \bar{Y})$ follows normal distribution with mean

$$E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2 = 68 - 65 = 3$$

and variance

$$\text{Var}(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{(2.3)^2}{35} + \frac{(2.5)^2}{50}$$

$$= 0.1511 + 0.1250 = 0.2761$$

Thus standard Error $= \sqrt{\text{Var}(\bar{X} - \bar{Y})}$

$$= \sqrt{0.2761} = 0.525$$

5) Here, we are given that

$$N = 10000,\ X = 7000 \Rightarrow P = \frac{X}{N} = \frac{7000}{10000} = 0.70 \text{ \& } n = 100$$

First of all, we find the sampling distribution of sample proportion. Here, the sample proportion is not given and n is large so we can assume that the conditions of normality hold. So the sampling distribution is approximately normally distributed with mean

$$E(p) = P = 0.70$$

and variance

$$Var(p) = \frac{PQ}{n} = \frac{0.70 \times 0.30}{100} = 0.0021 \quad [\because Q = 1 - P]$$

Thus, standard error is

$$\text{standard error} = (p) = \sqrt{Var(p)} \Rightarrow \sqrt{0.0021} = 0.0458$$

6) Here, we are given that

$$P_1 = \frac{25}{100} = 0.25,\ P_2 = \frac{20}{100} = 0.20,\ n_1 = 250,\ n_2 = 200$$

Let p_1 and $p_2$be the sample proportions of alcohol drinkers in two cities A and B respectively. Here the sample proportions are not given and n_1 and n_2 are large $(n_1, n_2 > 30)$ so we can assume that conditions of normality hold. So the sampling distribution of difference of proportions is approximately normally distributed with mean

$$E(p_1 - p_2) = P_1 - P_2 = 0.25 - 0.20 = 0.05$$

and variance

$$Var(p_1 - p_2) = \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2} = \frac{0.25 \times 0.75}{250} + \frac{0.20 \times 0.80}{200}$$

$$\Rightarrow \frac{0.75}{1000} + \frac{0.80}{1000} = \frac{1.55}{1000} = 0.00155$$

Hence, Standard Error $(p_1\text{-}p_2)$ =0.0394

7) We know that the mean and variance of chi-square distribution with n degrees of freedomare

Mean = n and Variance = 2n

In our case, n = 10, therefore,

Mean = 10 and Variance = 20

8) Here, we are given that

$$f(\chi^2) = \frac{1}{96} e^{-\chi^2/2} (\chi^2)^3 ; \qquad 0 < \chi^2 < \infty$$

We have the probability density function of χ^2 – distribution as:

$$f(\chi^2) = \frac{1}{2^{\nu/2}\sqrt{\frac{\nu}{2}}} e^{-\chi^2/2} (\chi^2)^{(\nu/2)-1} \quad ; \qquad 0 < \chi^2 < \alpha$$

Where $\nu = n = 1$

by comparision we have

$$\left(\frac{\nu}{2}\right) - 1 = 3 \Rightarrow \frac{\nu}{2} = 4 \Rightarrow \nu = 8$$

Thus, $\nu = 8 \Rightarrow n - 1 = 8 \Rightarrow n = 9$

Mean = n = 9 and Variance = 2n = 18.

9) Refer Sub-Section 9.6.1.

10) Here, we are given that

$\mu = 60, n = 15, \overline{X} = 65$ and $s = 12$

We know that the t-variate is

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

Therefore, we have

$$t = \frac{65 - 60}{12/\sqrt{15}} = \frac{5}{4.39} = 1.14$$

11) We know that the mean and variance of t-distribution with n degrees of freedomare

$$\text{Mean} = 0 \text{ and Variance } = \frac{n}{(n+2)}; \qquad n > 2$$

In our case, n = 8, therefore,

$$\text{Mean} = 0 \text{ and Variance } = \frac{8}{(80+2)} = \frac{8}{10} = 0.8$$

12) Refer Sub-Section 9.6.2

13) Here, we are given that

$n_1 = 15, n_2 = 10, \sigma_1 = 65, \sigma_2 = 45, s_1 = 60, s_2 = 50$

Thevalue of F-variate can be calculated as follows:

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

where, s_1^2 & s_2^2 are the values of sample variances.

Therefore, we have

$$F = \frac{(60)^2/(65)^2}{(50)^2/(45)^2} = \frac{0.85}{1.23} = 0.69$$

14) Refer Sub-Section 9.6.3.

15) We know that the mean and variance of F-distributionwith ν_1 and ν_2 degrees of freedom are

$$\text{Mean} = \frac{\nu_2}{\nu_2 - 2} \quad \text{for} \quad \nu_2 > 2$$

and

$$\text{Variance} = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)} \quad \text{for} \quad \nu_2 > 4.$$

In our case, $\nu_1 = 5$ and $\nu_2 = 12$, therefore,

$$\text{Mean} = \frac{\nu_2}{\nu_2 - 2} = \frac{12}{10} = 1.2$$

$$\text{Variance} = \frac{2(12)^2(5+12-2)}{5(12-2)^2(12-4)} = \frac{30 \times 144}{40 \times 100} = 10.8$$

16) Same as Sub-section 9.6.3.

UNIT 10 STATISTICAL ANALYSIS-I

Structure

10.1 INTRODUCTION

In the previous unit, we have discussed about the very important concept of 'Statistical Inference' that is known as sampling distribution. We also discussed some basic concepts like population sample, parameter, statistic estimator and estimates. We have discussed that the probability distribution of a sample statistic is known as sampling distribution of it. In previous unit we have also discussed the sampling distribution of the statistics mean and proportion as well as some of the exact sampling distributions, i.e., t, chi-square and F – distribution along with its properties and applications. In this unit we shall discuss the concept and meaning of testing of hypothesis. It provides us significant conclusions on the assumption(s) made for any

population parameter on the basis of sample based estimated values. The assumptions which are made for testing purpose are known as statistical hypotheses.

In testing a hypothesis, we take decision about trueness of statement made for a population parameter on the basis of observations. For example, a doctor may be interested to know whether the new medicine is really effective for controlling blood pressure, a manager may be interested to know whether one brand of electric bulb is better than the other, a psychologist may be willing to know whether the IQ of students studying in an Open University is up to the standard of the students of IIT. Such statistical decisions can be acceptable/unacceptable only if these are proved/disproved with the help of sample data taken from the concerned population.

Generally, the calculated value of the sample statistic differs from the assumed value of the population parameter since the statistic is based on a part of the population and not on the entire population. Now, a question arises whether this difference is actually significant or it appears only due to fluctuations of samples. A little difference sometimes may occur due to a main cause or due to sources of sampling errors.

It, therefore, seems that some theoretical backgrounds should be developed for the testing of correctness/incorrectness of the assumed statement (hypothesis) while the statements are to be verified on the basis of some sample values (statistic) which themselves may vary from sample to sample and, thus, may yield some kind of errors. Whatever procedure is applied for this purpose is generally known as testing of hypothesis.

10.2 OBJECTIVES

After studying this unit, you should be able to

- define a hypothesis, null hypothesis, alternative hypothesis, simple and composite hypothesis;
- define and explain the type-I and type-II errors;
- explore the concept of critical regions, level of significance and degree of freedom;
- describe the procedure of testing a hypothesis; and
- perform the test of hypothesis for large samples for population mean and difference between two population mean.

10.3 HYPOTHESIS

In our daily life, each one of us comes across with the problems of testing some kinds of assumptions which we may have in our mind. For example, a housewife while cooking rice, makes an assumption in her mind that if only few pieces of rice are checked and found well-cooked then the whole amount of rice may be thought of fully cooked. Thus, she verifies and tests her

believe on the basis of few observations and not with the whole amount of rice.; a manufacturer of some goods may think that advertisement of his product will increase the sale and, therefore, may decide the trueness of his thinking on the basis records of market evidences; a student, on the basis of records of passed-out students, may have a belief that few subjects might help him to pass out a bachelor degree course, etc. Each of these are examples where any statement (assumption or belief) about some system needs to be verified, that is, proved or disproved, on the basis of some data/information.

The above discussions and examples now help us to define a statistical hypothesis. It may be defined in a number of ways: (i) A statistical hypothesis is an assumption about a population or about its parameter. This assumption may or may not be true; (ii) A tentative theory about the natural world; a concept that is not yet verified but that if tried would explain certain facts or phenomenon; (iii) It is a supposition, a proposition or principle which is supposed or taken for granted, in order to draw a conclusion or inference for proof of the point in question.

A statistical hypothesis stands in contrast to a simple assertion. If one makes some statement and believes it to be true without its verification then it is an assertion and not a hypothesis. For example, the statement "It is cloudy weather outside" is an assertion, because the person believes it to be true and he wants other persons also to believe it; whereas the statement "I think it is cloudy weather outside" is a hypothesis since he/she involves in his/her statement the possibility to prove or disprove the statement by others. In fact, the intention is to determine at a later stage the truth or falsity or probability of his/her statement.

In many cases, it is necessary to assume some theoretical probability distributions for testing the hypothesis, particularly when there is no direct knowledge of the population from which the observations are taken. For example, a factory owner produces a product of weight exactly 10 g. The producer finds that the weight of product has been reduced a little at present. The owner is worried whether weight reduction is due to unavoidable reasons or due to poor quality of raw material? He may then frame the hypothesis of no change in weight and use a suitable theoretical probability distribution for the variations in weight.

10.3.1 Null and Alternative Hypothesis

A statistical hypothesis is an unproved, well-structured, assumed statement about the population. Null hypothesis is a preliminary part of it, and it is denoted by symbol H_0. It is relating to basic ignorant thought of "no difference, no change or no effect" between two or more occasions or situations or places or values. Null means "nothing" and null hypothesis is that nothing is happening, nothing is present and no changes are observed, etc., and in this sense, it is a hypothesis which might be falsified on the basis of the observed data.

For example, statements like, "effect of new injection is same as the old one (no difference)"; "the efficiency of new technology is same as the earlier technology (no change)"; "the two mathematics teachers are equally good (no one is better than the other)", etc. Mathematically, if θ_1 and θ_2 are values of similar parameters of two different theoretical populations, then our null hypothesis will be H_0: $\theta_1 = \theta_2$, that is, there is no difference in their values. Similarly, if θ_0 is the specified value of the parameter, we should frame our null hypothesis H_0: $\theta = \theta_0$, that is, there is no change in the given value. Thus, null hypothesis represents what we would believe by default, before seeing any evidence. It generally negates any change, any happening, etc. It is important to mention here that in all types of problems, it is only the null hypothesis which is either rejected (not accepted) or not rejected (accepted).

According to Prof. R.A. Fisher who coined the term "null hypothesis", a null hypothesis is tested for its possible rejection under the assumption that it is true. Formulation of null hypothesis is a vital step in testing statistical significance. It should be exact, that is, free of vagueness and ambiguity

Every statistical null hypothesis always occurs in conjunction with another hypothesis known as "alternative hypothesis" generally denoted by H_1. It is a well-structured assumed statement which complements the null hypothesis and is stated in such a way that they are mutually exclusive, that is, if one is true, the other must be false and vice versa. Therefore, when null hypothesis is rejected, it is equivalent to accept alternative hypothesis.

Some examples of H_0 with corresponding H_1 are shown below:

(i) If H_0: $\theta = \theta_0$ then H_1 may be either $\theta \neq \theta_0$ or $\theta > \theta_0$ or $\theta < \theta_0$

(ii) If H_0: $\theta_1 = \theta_2$ then H_1 may be either $\theta_1 \neq \theta_2$, $\theta_1 > \theta_2$ or $\theta_1 < \theta_2$

The acceptance of H_1: $\theta \neq \theta_0$ provides us an assurance of significant departure of the current parametric value to the past /standard value. Likewise, acceptance of H_1: $\theta > \theta_0$ (or H_1: $\theta < \theta_0$) provides us an assurance of departure of current value from the past /standard value θ_0 toward higher (or lower) level respectively.

10.3.2 Simple and Composite Hypothesis

Since a population is characterized by its parameters, therefore, when value of parameter(s) is/are known, the population is treated as known (or specified completely). A hypothesis which completely specifies parameter(s) of a theoretical population (probability distribution) is called a simple hypothesis. If a population possesses the only parameter θ then the statement H_0: $\theta = \theta_0$ is a simple hypothesis because acceptance of it specifies the population. In contrast, a hypothesis which does not specify the distribution completely is known as composite hypothesis.

As an example, let $x_1, x_2, \ldots, x_n$ be a random sample of size n taken from a normal population (normal probability distribution) with mean μ and variance σ^2. Then a hypothesis H_0: $\mu = 20$, given that $\sigma^2 = 5$ is a simple

hypothesis since then the normal distribution is completely specified. Acceptance of it would decide the value of the parameter μ. But a hypothesis like, H_0: $\mu > 250$ is a composite hypothesis as it does not exactly specify the value of μ; it may be any value greater than 250. If both null and alternative hypotheses are taken together, we may have following cases of combinations of simple and composite hypotheses:

Case I: Simple null versus Simple alternative Hypothesis, like,

$$H_0: \theta = \theta_0 \quad ; H_1: \theta = \theta_1$$

For instance,

H_0: weight of tomato is 130 gms (i.e. H_0: θ = 130gms)

H_1: weight of tomato is 200 gms (i.e. H_1: θ = 200gms)

The acceptance of H_0 or H_1, completely specify the population parameter (tomato weight).

Case II: Simple null versus Composite alternative Hypothesis, such as

$H_0: \theta = \theta_0$; $H_1: \theta > \theta_0$ or $H_1: \theta < \theta_0$ or $H_1: \theta \neq \theta_0$.

For example,

H_0: weight of tomato is 130 gm. (i.e. H_0: θ = 130gm).

H_1: weight of tomato is above to 130 gm. (i.e. H_1:$\theta > 130$gm).

The acceptance of H_0 specifies the population completely, but acceptance of H_1 does not, because a value greater to 130 will not be a unique value.

CHECK YOUR PROGRESS 1

Note: i) Use the space given below for your answers.
ii) Check your answers with those given at the end of the unit.

1. Define Null Hypothesis and Alternative Hypothesis.

……

……

……

……

2. Describe the simple and composite structure of Null hypothesis and Alternative hypothesis.

……

……

……

……

10.4 SOME BASIC CONCEPTS

A null hypothesis is always tested by ~~the~~ sample data. Let $x_1, x_2,\ldots,x_n$ be a random sample drawn from a population having unknown parameter θ. The collection of all possible values of $x_1, x_2, \ldots, x_n$ is a set called "sample space" S and a particular choice of $x_1, x_2, \ldots, x_n$ represents a point in that space.

10.4.1 Critical Region (Region of Rejection)

As per criterion of classical testing of hypotheses concepts, in fact, the entire sample space is partitioned into two disjoint sub-spaces, say ω *and* $S-\omega = \vec{\omega}$, such that if computed value of the test statistic falls in ω, we reject the null hypothesis against the alternative hypothesis and if it falls in $\overline{\omega}$ we do not reject the null hypothesis in favour of alternative hypothesis. In that sense, the region ω is called a "rejection region" or more popularly a "critical region". The $\overline{\omega}$ is known as "acceptance region".

As an example, a student appears in total 10 papers, two of each in English, Physics, Chemistry, Mathematics, and Computer Science. His scores in these papers are respectively $x_1, x_2, x_3\ldots x_9, x_{10}$ out of maximum marks 100 for each paper. It is a sample data with sample size n=10.

We define statistic $t_n = \sum_{i=1}^{n} x_i$ as a sum of all these scores. The range of t_n is $0 \le t_n \le 1000$. For obtaining the distinction award in a course program, a student needs to have total score more than 750 which is a rule. Suppose any one student is randomly selected from that course program and we frame hypotheses as:

H_0: selected student is not a distinction award holder

H_1: selected student is a distinction award holder.

Now, according to null hypotheses, we have two regions by partitioning the space [0- 1000]: (i) acceptance region [0-750] and (ii) rejection region [751-1000]) as shown in figure 3.1.

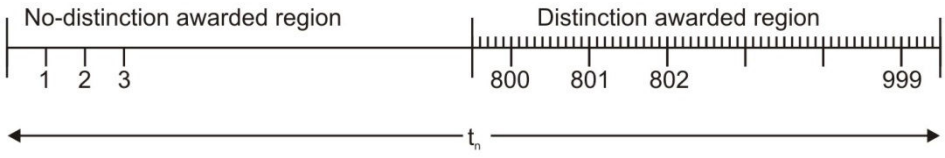


Figure 10.1: Acceptance and critical regions for Distinction Award

On the basis of sample of size n and statistic t_n the computed value may fall in distinction award region or not, depending how sample observations are.

For proving H_0, we use test statistic $t_{10} = \sum_{i=1}^{10} x_i$ (sum of scores of 10 papers) and conclude: reject H_0 if $t_{10} > 750$, accept H_0 if $t_{10} \le 750$. Clearly, it

indicates that a basic structure of the procedure of testing of hypothesis needs two regions.

The region of rejection has a pre-fixed area denoted by α, corresponding to a cut-off value in a probability distribution of test statistic. The prefixed probability area, that is, α is also called size of test or level of significance or probability of type I error.

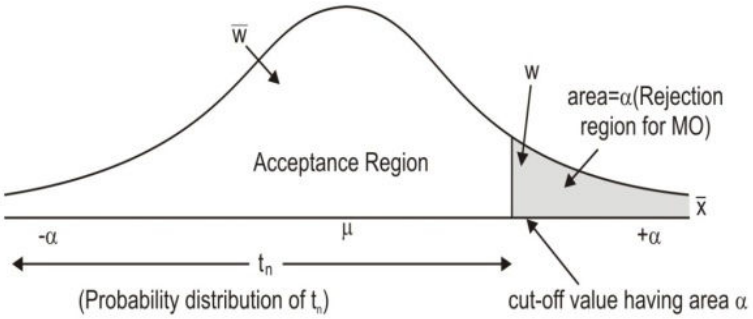


Figure 10.2: Acceptance and Critical regions for Statistic t_n.

Symbolically, if w be critical region and t_n be the value of statistic based on sample of size n, then

$$P[t_n \in \omega / H_0] = P[t_n \in \omega, \text{ when} H_0 \text{ is true}] = \alpha$$

$$P[t_n \in \bar{\omega} / H_1] = P[t_n \in \bar{\omega} \text{ when } H_1 \text{ true}] = \beta$$

10.4.2 Type-I and Type-II Error

In previous section, we have discussed a rule that if the value of test statistic falls in the critical region we reject the null hypothesis, otherwise we do not reject it. Since the test statistic itself is generated from the sample, which has been selected at random from the population, it might be very far from the actual value of the parameter concerned if the sample is not a good representative of the population or might be very close to the parameter value if the sample is a good representative. Therefore, in taking decision for accepting or rejecting null hypothesis on the basis of the value of the statistics, we may commit some kinds of errors.

For example, an engineer on the basis of a sample infers that a packet of screws is sub-standard whereas actually it is not or may infer the packet to be good whereas actually it is sub-standard. Similarly, a person may be declared a patient of cancer and hence kept under high potency medicines whereas he is not and vice versa. In both the cases, the error is because of inappropriate samples of screws and diagnostic tests respectively which have been taken. Accordingly, the errors which may arise in taking decisions are termed as Type – I and Type – II errors which are depicted in the following table:

Decision	**H_0 True**	**H_1 True**
Reject H_0	Type - I Error	Correct Decision
Accept H_0	Correct Decision	Type -II Error

Type-I Error:

From the table we see that the decision relating to rejection of H_0, when, in fact, it is true is called type-I error. The probability of committing type -I error is called "size of the test" or "level of significance" and is generally denoted by α Thus,

$$\alpha = P[\text{Rejecting } H_0 \text{ when } H_0 \text{ true}]$$

$$= P[\text{Rejecting } H_0 \mid H_0 \text{ true}]$$

$$= P[x \in \omega / H_0]$$ where x stands for the value of the statistic under study. Obviously then we have

$$1-\alpha = 1-P[\text{Rejecting } H_0 \mid H_0 \text{ true}]$$

$$= P[\text{Accepting } H_0 \mid H_0 \text{ true}] = P[\text{correct decision}].$$

$(1-\alpha)$ is probability of correct decision and it correlates to the concept of $100(1-\alpha)\%$ confidence interval used in the estimation. Theoretically, the test procedures are so constructed that the risk of rejecting H_0 when it is true is small.

Type -II error:

The decision relating to acceptance of H_0 when it is false, (that is H_1 true) is called type-II error. The probability of committing type-II error is denoted by β. Thus,

$$\beta = P[\text{Accepting } H_0 \text{ when } H_0 \text{ false}]$$

$$= P[\text{Accepting } H_0 \text{ when } H_1 \text{ true}]$$

$$= P[\text{Accepting } H_0 \mid H_1 \text{ true}] = P[x \in \bar{\omega} / H_1].$$

Hence $1-\beta = 1-P[\text{Accepting } H_0 \mid H_1 \text{ is true}]$

$$= P[\text{Rejecting } H_0 \mid H_1 \text{ is true}] = P[\text{correct decision}].$$

Therefore, $(1-\beta)$, that is, probability of not committing a type-II error is the probability of correct decision, which is called power of the test.

Being probabilities of type-I and type-II errors, both α and β should be as small as possible, but with a given set of data, both types of errors cannot be minimized simultaneously. Therefore, customarily, the test procedures are developed with an aim to minimize β with a fixed assumed value of α.

10.4.3 Level of Significance

We know that probability of Type-I error is known as level of significance of a test which is denoted by α. It is also called the size of a test or critical region. As discussed above, level of significance is a fixed assumed value which is decided before starting the test procedure. The most commonly used values of α are 0.05 (5%) and 0.01 (1%). By selecting $\alpha = 5\%$, we mean that if the null hypothesis is true, there is a 5% chance of rejecting it because of

random causes, or equivalently, the concluding statement about H_0" is true only with 95% assurance. Similar conclusions can be drawn for $\alpha = 1\%$.

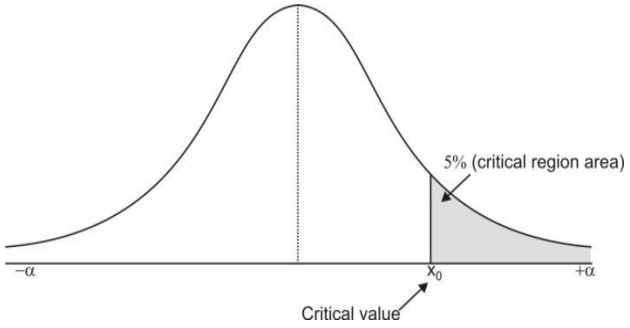


Figure 10.3: Critical region area at 5% level of significance.

Consider the figure 10.3, showing probability curve of any probability distribution, where entire area underlying the curve is obviously equal to 100% out of which let 5% area (shaded area) lies at the end of one side (right side) of the curve. Let this pre-fixed area constitute the critical region for a test, that is, here $\alpha = 5\%$. Now, as soon as we decide (pre – fix) the value of α for testing our hypothesis, the cut-off point on x axis (which is shown in the figure as x_0) can easily be obtained on the basis of the used probability distribution. This value (termed as "critical value") is used in decision making for the acceptance/rejection of H_0. The calculated statistic $t_n = f(x_1 \ldots x_n)$ may result in a value either less than or greater than the critical value x_0 So, we can make decisions rules, using t_n and x_0, as follows:

Reject H_0 (or equivalently accept H_1) if $t_n > x_0$; otherwise

accept H_0 (or equivalently accept H_1) if $t_n < x_0$.

This makes it clear that main focus in a testing problem is to decide the critical value x_0 on the basis of the probability distribution of the statistic t_n. This is important to mention here that for standard testing problems, that is, with known null and alternative hypotheses, the statistic to be used for testing purpose, is also pre-decided and its sampling distribution is also known. Some well-known distributions used for most of the testing problems are Z, χ^2, t and F distributions.

10.4.4 Degree of Freedom (df)

The term "degree of freedom" (df) is related with the concept of independency of sample observations. We have seen that the entire decision in testing of a hypothesis is based on the sample observations $x_1, x_2, x_3, \ldots, x_n$. If all these observations are independent among themselves then the conclusion drawn about H_0 is said to be having n degree of freedom.

Suppose for 50 observations $x_1, x_2, \ldots, x_{50}$, a restriction (condition) is imposed as $\sum_{i=1}^{50} x_i = 1000$ that is, there sum is fixed and it is 1000.

If it is like so, then actually, there are now only 49 values which are independent, because any value (say, the last one) can be obtained using the equation

$$x_{50} = \left(1000 - \sum_{i=1}^{49} x_i\right)$$

It implies that by putting one condition on the sample values, we lose one degree of freedom and in general if for a sample of size n, there are k restrictions (k<n), the degrees of freedom will be (n-k).

CHECK YOUR PROGRESS 2

Note: i) Use the space given below for your answers.

ii) Check your answers with those given at the end of the unit.

3. Define a Statistic.

……

……

……

……

4. Describe the critical region with an example.

……

……

……

……

5. Describe Type I and Type II errors.

……

……

……

……

6. Describe the level of significance with example.

……

……

……

……

7. Describe the degree of freedom with example.

……

……

……

……

10.5 TESTING OF HYPOTHESIS

So far, we have discussed the basic concepts of the terms which are used in determining the inference of the Population based on the sample results. The purpose of this type of inference is to determine whether a certain assumption or hypothesis about a parameter can be justified by statistical evidence and also to make valid decisions about the population parameters based on the analysis of samples. For example, a quality control manager must decide whether the machines producing a product are working properly on the basis of samples taken from the lot.

In order to make these statistical decisions, we make certain assumptions about the population parameters to be tested. Then a sample is taken to estimate the values of these population parameters. If the estimates favor the hypothesis, then we do not reject the hypothesis as being correct. Since we are testing for our assumptions or hypothesis being correct or not, this field of decision making is known as Testing of Hypothesis. Now we will discuss the step-by-step procedure of testing a hypothesis in details.

10.5.1 Procedure of Testing of Hypothesis

A hypothesis testing procedure should be designed in a number of steps. If x_1, x_2, …, x_n is a random sample drawn from a population having unknown parameter θ, then steps though which it should pass are:

Step I: The first step in any test is to state the relevant null and alternative hypotheses to be tested. For example, for testing the significance of a given value of θ say, θ_0, these may be formulated as H_0: $\theta = \theta_0$, H_1: $\theta \neq \theta_0$ [or for comparing two values θ_1 and θ_2 we may test in comparative sense, H_0: $\theta_1 = \theta_2$ against H_1: $\theta_1 \neq \theta_2$].

Step II: Establish a criterion for acceptance/rejection of null hypothesis, i.e., decides the level of significance α at which we want to test our hypothesis. Generally, it is taken as 1% or 5% ($\alpha = 0.05$ or 0.01).

Step III: Define a statistic t_n on sample observations x_1, x_2, x_3,…,x_n, and call it a test-statistic. For example, t_n= mean = $\bar{x} = \frac{1}{n}\sum x_i$, t_n = Max (x_1, x_2,…,x_n) = Highest sample observation, $t_n = \sum x_i^2$, etc. Depending upon the objective of the test, other test statistics may be proportion, difference between means, variance, correlation

coefficient, etc. In general, a test statistic is taken preferably in the standard format like Z(for large samples), t (for small samples), χ^2 and F as:

$$t_n = \frac{\text{Statistic - Parameter's Assumed Value}}{\text{Standard Error of Sample Statistc}}$$

Do numerical computation of t_n on the sample data.

Step IV: Obtain the probability distribution of t_n, under the null hypothesis from the assumptions. Generally, with the objectives and the defined test statistic, the distributions are in the standard form like Z, χ^2, t or F or any other well-known distribution in literature.

Step V: Using the distribution and value of α, obtain the critical value (or cut-off value) in that probability distribution.

Step VI: Take decision about H_0, in light of step I and II, as

(i) Reject H_0 if modulus of calculated value of t_n, i.e., $|t_n| \geq$ critical value (or tabulated value) at α level of significance.

(ii) Do not reject H_0 if modulus of calculated value of t_n, i.e., $|t_n| <$ critical value (or tabulated value) at α level of significance.

When we take H_0: $\theta = \theta_0$ and H_1: $\theta > \theta_0$ in step-I then decision in step VI is different and we reject H_0: $\theta = \theta_0$ if $t_n >$ critical value. No modulus sign is used for this case for t_n. Similarly, if H_1: $\theta < \theta_0$, we reject H_0: $\theta = \theta_0$ when $t_n <$ - critical value at α level of significance.

10.5.2 One-Tail and Two-Tail Tests

Figure 3 showed that the critical value lies on the right side of the curve, but always it is not the case. In any test, the critical region may lie either at one side (at the left tail or at the right tail of the curve) or at both sides, depending upon how the alternative hypothesis H_1 is formulated. Accordingly, we have one-tailed tests (right or left-tailed tests) or two-tailed test. If it is only at one side, the size is α, and if is on both the sides, then, the size will be $\alpha/2$ on both the sides. Accordingly, the test is said to be one-tailed test or two-tailed test respectively.

For example, consider a test for testing hypotheses:

(i) H_0: $\theta = \theta_0$, H_1: $\theta > \theta_0$ and (ii) H_0: $\theta = \theta_0$, H_1: $\theta < \theta_0$.

Obviously in (i), H_0 will be rejected (H_1 will be accepted) when $\theta > \theta_0$; this means that critical region lies on the upper side (right side) of the distribution, hence it would be a one-tailed test (right-tailed test). The same argument shows that (ii) would be a one-tailed test (left-tailed test). So, both (i) and (ii) are one-tailed tests.

Figure 10.4 show the situation of critical values and critical regions for right-tailed test.

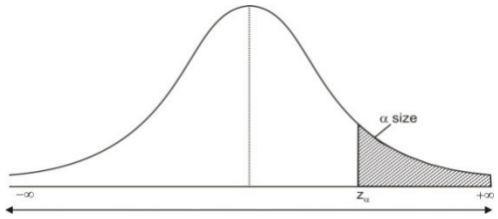


Figure 10.4: One-tail (right-tail) critical region of size α

In contrary, when the alternative hypothesis is like H_1: $\theta \neq \theta_0$ where θ may be lower than θ_0 or may be higher than θ_0, we have the situation of two-tail test. Here the critical region may be on the left tail or may be on the right tail according as $\theta < \theta_0$ or $> \theta_0$ and we have two critical values; one on left tail and another on right tail. Since the level of significance is α, it must be divided into two such that both sides have critical region equal to α/2 (see the Figure 3.5 below).

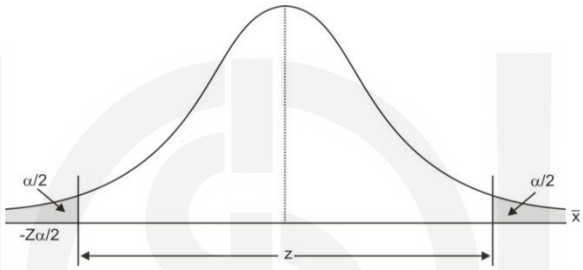


Figure 10.5: Two-tail critical regions of sizeα/2 at each end.

We summarize the three cases related to above discussion as follows:

Null hypothesis	Alterative hypothesis	Types of critical region / test
$H_0 : \theta = \theta_0$	$H_1 : \theta \neq \theta_0$	**Two tailed test having critical regions on both sides.**
$H_0 : \theta = \theta_0$	$H_1 : \theta > \theta_0$	**One tailed test having critical region on right hand side.**
$H_0 : \theta = \theta_0$	$H_1 : \theta < \theta_0$	**One tailed test having critical region on left hand side.**

Likewise, if the null hypothesis is formulated for testing the equality of two parameters as H_0: $\theta_1 = \theta_2$, we may have alternative hypothesis either H_1: $\theta_1 > \theta_2$ or H_1: $\theta_1 < \theta_2$. Then, in view of the discussions made above and from the above table, we can see that the test would be right-tailed test when H_1: $\theta_1 > \theta_2$ and left-tailed test when H_1: $\theta_1 < \theta_2$ with level of significance α in each

case. Further, if in conjunction with H_0: $\theta_1 = \theta_2$, we have H_1: $\theta_1 \neq \theta_2$, then it would be a two-tailed test with level of significance $\alpha/2$ on both sides.

CHECK YOUR PROGRESS 3

Note: i) Use the space given below for your answers.

ii) Check your answers with those given at the end of the unit.

8. Explain the design of testing of hypothesis.

……

……

……

……

9. Describe the One tail and Two tail tests with example.

……

……

……

……

10.6 LARGE SAMPLES TESTS

After discussing some of the preliminaries of testing of hypotheses, now we shall discuss some test procedures in which our sample is assumed to be of quite large size though finite.

In Statistics, it has been theoretically proved that if sample taken from any population is of large size, whatever be the parent distribution, the sampling distribution of the statistic based upon this sample can be approximated by a normal distribution. Therefore, one can apply normal distribution-based test procedures to obtain the cut off points at the pre-defined level of significance in such cases. Accordingly, such tests are called large sample tests.

Suppose $x_1, x_2, \ldots, x_n$ is a random sample of size n selected from a population having unknown parameter θ. Generally, for assuring n to be large enough, we assume $n \geq 30$. Let with this assumption, we require a test procedure for the significance of parameter θ. In this Section, we shall describe here a number of tests which are as follows:

10.6.1 Test for The Significance of Population Mean

Our aim here is to test whether the population mean, denoted by μ, assumes a specific value μ_0 or not. Thus, the null hypothesis would be H_0: $\mu = \mu_0$ and the alternative hypothesis H_1 may be any of these:

$$H_1 : \mu \neq \mu_0 \quad \text{or } H_1 : \mu > \mu_0 \quad \text{or } H_1 : \mu < \mu_0$$

If $x_1, x_2,\ldots,x_n$ be a random sample of size $n \geq 30$ taken from a population with mean μ and finite variance σ^2 then by central limit theorem the sample mean is asymptotically normally distributed with mean μ and variance σ^2/n irrespective of parent population be normal or non-normal.

Therefore, for mean $\bar{x} = (1/n)\,\Sigma x_i$, i=1,2,3,…. N

$$E(\bar{x}) = \mu \qquad Var(\bar{x}) = V(\bar{x}) = \frac{\sigma^2}{n}$$

The test statistic Z which is to be used for testing H_0:

$$Z = \frac{\bar{x} - E(\bar{x})}{\sqrt{V(\bar{x})}} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1) \text{ under } H_0 \text{ (assuming } \sigma^2 \text{ known)}$$

If σ is unknown, then we replace σ by its sample estimate $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ and then we compute the Z-statistic (since n large) as:

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim N(0, 1) \text{ under } H_0 \text{ (for large n)}$$

The test is one-tail if either H_1: $\mu > \mu_0$ or H_1: $\mu < \mu_0$ and two-tail if $H_1 : \mu_1 \neq \mu_2$ as the case may be. The tabulated value z_α is available in the Table 3.1 for one-tail and two-tail tests at different value of α as follows:

Table 10.1

Nature of Test	**Level of Significance and Critical (Tabulated) Value**		
	α=0.01	**α=0.05**	**α=0.10**
Right Tail Test	z_α=2.33	z_α=1.645	z_α=1.28
Right Tail Test	z_α= -2.33	z_α= -1.645	z_α= -1.28
Two Tail Test	z_α=2.58	z_α=1.96	z_α=1.645

For prefixed level of significance α, the calculated value Z, obtained by above formula is compared with the tabulated z_α as mentioned below:

One Tail Test: When H_0: $\mu = \mu_0$ and either H_1: $\mu > \mu_0$ or H_1: $\mu < \mu_0$

If $Z > z_\alpha$ and H_1: $\mu > \mu_0$, (right tail test) the null hypothesis is rejected and significant evidence of departure from standard value is confirmed. Otherwise, for $Z < z_\alpha$ the H_0 is not rejected with conclusion that no significant evidence exists against the null hypothesis. On the other hand, if $Z < -z_\alpha$ and H_1: $\mu < \mu_0$ (left tail test), we reject H_0, otherwise do not reject for $Z > -z_\alpha$.

Two Tail Test: When H_0: $\mu = \mu_0$ and H_1: $\mu \neq \mu_0$

If $|Z| < z_{\alpha/2}$ the null hypothesis will not be rejected and when $|Z| > z_{\alpha/2}$ the null hypothesis will be rejected.

Example 1: A sample of 100 male students is found to have a mean height of 72 inches. Can it reasonably be regarded as a sample from a large population with mean height 70 inches and standard deviation 4 inches?

Solution: We are given that

$$n = 100,\ \bar{x} = 72 \text{ inches},\ \mu = 70 \text{ inches and } \sigma = 4 \text{ inches}$$

We wish to test the null hypothesis:

$$H_0: \mu = 70 \text{ against } H_1: \mu \neq 70$$

It is two-tail test and the test statistic is Z, since n is more than 30. We have

$$Z = \frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} = \frac{72 - 70}{4/\sqrt{100}} = \frac{2}{0.4} = 5.0$$

Since calculated value of test statistic Z is greater than 3, we reject our null hypothesis H_0 at every level of significance (i.e. at 5%, 1% both). We conclude that sample is not from population with mean 70 and standard deviation 4.

Example 2: *A manufacturer of ball point pens claims that a certain pen he manufactures has a mean writing-life of 500 pages. A purchasing agent selects a sample of 100 pens and put them for the test. The mean writing-life of the sample found 490 pages with standard deviation 50 pages. Should the purchasing agent reject the manufacturer's claim at 1% level of significance?*

Solution: Given that

$$\mu = 500,\ n = 100,\ \bar{x} = 490 \text{ and } s = 50$$

We wish to test the null hypothesis

$$H_0: \mu = 500 \text{ against } H_1: \mu < 500$$

Since population standard deviation (σ) is unknown, so we take estimate of σ as sample standard deviation (s) in place of σ.

It is left-tail test and the test statistic Z is

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{490 - 500}{50/\sqrt{100}} = \frac{-10}{5} = -2.0$$

The tabulated value at 1% level of significance is $z_\alpha = z_{0.01} = -2.33$ for left tail test. Since calculated value of Z is greater than the tabulated value, so we do not reject the null hypothesis and conclude that the purchasing agent will be agree with the manufacture's claim at 1% level of significance.

Note: i) Check your answers with those given at the end of the unit.

10) A sample of 900 bolts has a mean length 3.4cm. Is the sample regarded to be taken from a large population of bolts with mean 3.25 cm. and standard deviation 2.61 cm. at 5% level of significance?

11) A big company uses thousands of CFL lights every year. The brand that the company has been using in the past has average life 1200 hours. A new brand is offered to the company at a price lower than they are paying for the old brand. Consequently, a sample of 100 new brand CLF light is tested and yields an average life of 1180 hours with standard deviation 90 hours. Should the company accept to the new brand?

10.6.2 Test for Equality of Two Population Means

Let there be two populations say population-I and population-II. The population-I has mean μ_1 and variance σ_1^2 whereas the population-II has mean μ_2 and variance σ_2^2.

For testing the hypothesis that two-population means are same, we draw independent random samples of size n_1 and n_2 respectively from population-I and population-II. Let x_{1i} (i =1, 2, …, n_1) and x_{2i} (i=1, 2,…, n_2) be the observations in samples selected from population-I and population-II respectively and $\bar{x}_1$ and $\bar{x}_2$ be the corresponding sample means. In order to test the null hypothesis, we assume that both sample sizes are large enough to ensure that sampling distribution of the test statistic to be used is asymptotically normal distribution. As we have discussed in previous unit that the sampling distribution of difference of two large sample means follows asymptotic normal distribution with mean $(\mu_1-\mu_2)$ and variance $(\sigma_1^2/n_1 + \sigma_2^2/n_2)$. The scheme of selecting the samples from both the populations is shown in the following **fig 10.6**:

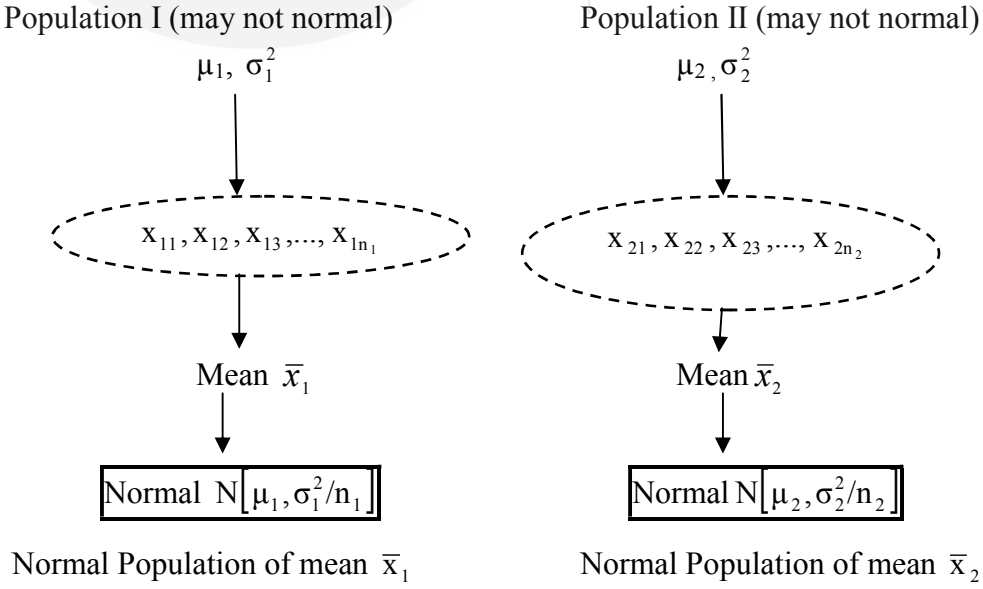


Fig 10.6: Two independent samples from two different populations having normal distribution of means

Thus, the Test statistic

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{\sqrt{V(\bar{x}_1 - \bar{x}_2)}} \sim N(0, 1) .$$

We assume that population variances σ_1^2, σ_2^2 are known to us. Thus we have

$$E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2) = \mu_1 - \mu_2$$

and

$$Var(\bar{x}_1 - \bar{x}_2) = Var(\bar{x}_1) + Var(\bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Our null hypothesis here is:

$$H_0: \mu_1 = \mu_2 \text{ (no difference in means)}$$

against the alternative hypotheses

$$H_1 : \mu_1 \neq \mu_2 \quad \text{or } H_1 : \mu_1 > \mu_2 \quad \text{or } H_1 : \mu_1 < \mu_2$$

Now from the above, we get the test statistic Z as

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

Since under H_0, $\mu_1 = \mu_2$, so $\mu_1 - \mu_2 = 0$, and hence

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1) \text{ under } H_0$$

For prefixed level of significance, α, the calculated Z, by the above formula, is compared with the tabulated z_α. The tabulated value of z_α may be determined from the Table 3.1 for either one-tail or two-tail, as the case may be.

For One Tailed Test: When H_0: $\mu_1 = \mu_2$ and H_1: $\mu_1 > \mu_2$ or H_1: $\mu_1 < \mu_2$

For H_1: $\mu_1 > \mu_2$ or H_1: $\mu_1 < \mu_2$ (right tail test), if $Z < z_\alpha$ the null hypothesis is not rejected and we conclude that, there is no significant evidence against the equality of means of two populations. So, both means will be treated as same at the α-level of significance otherwise if $Z > z_\alpha$ H_0 is rejected at α-level of significance and the difference between the two means $\mu_1$and μ_2 is established and concluded that one mean is greater than the other.

For H_1: $\mu_1 < \mu_2$ (left tail test), if $Z < -z_\alpha$ the null hypothesis is rejected at α level of significance, otherwise for $Z > -z_\alpha$, H_0 is not rejected.

For Two Tail Test: When H_0: $\theta_1 = \theta_2$ and H_1: $\theta_1 \neq \theta_2$

For H_1: $\theta_1 \neq \theta_2$ and if $|Z| < z_{\alpha/2}$, the null hypothesis is not rejected and for $|Z| > z_{\alpha/2}$ the null hypothesis is rejected at α-level of significance.

Case-I: When Variances are Equal and Known

If $\sigma_1^2 = \sigma_2^2 = \sigma^2$ i.e. populations have **same variances, i.e.,** σ^2 (known) then for testing $H_0 : \mu_1 = \mu_2$, the test statistic Z reduces to

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N[0, 1]$$

Case-II: When Variances are Equal and Unknown

If σ_1^2, σ_2^2 **are unknown** these two could be estimated by their respective unbiased estimators s_1^2, s_2^2 respectively, where

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 \quad ; \quad s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2$$

When $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (unknown), then the estimate of σ^2 is:

$$\hat{\sigma}^2 = \left[\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}\right]$$

and
$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}^2\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$$

Case-III: **When Variances are Unequal and Unknown**

If $\sigma_1^2 \neq \sigma_2^2$ **and** σ_1, σ_2 **(both are unknown), then if sample sizes** n_1, n_2 are large enough, we use s_1^2 and s_2^2 respectively for σ_1^2 and σ_2^2. Then the Z statistic will be:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}} \sim N[0, 1]$$

Example 3: In two samples of women from Punjab and Tamil Nadu, the mean height of 1000 and 2000 women are 67.5 and 68.0 inches respectively. Can the samples be regarded as drawn from the same population having standard deviation 2.5 inches? In other words, can the mean height of Punjab and Tamil Nadu women be regarded as same?

Solution: Given $n_1 = 1000$, $n_2 = 2000$, $\bar{x}_1 = 67.5$, $\bar{x}_2 = 68.0$, $\sigma = 2.5$ inches

We wish to test the null hypothesis that the samples are drawn from the same population, that is, mean height of Punjab and Tamil Nadu women are same.

H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 \neq \mu_2$

It is two tail test and test statistic Z is

$$|Z| = \frac{|\bar{x}_1 - \bar{x}_2|}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{|67.5 - 68.0|}{2.5\sqrt{\left(\frac{1}{1000} + \frac{1}{2000}\right)}} = 5.16$$

The calculated $|Z| = 5.16 >$ tabulated $z_{\alpha/2} = 1.96$ at5% level of significance which implies rejection of hypothesis H_0.

We conclude that women of Punjab and Tamil Nadu differ significantly in terms of their mean height.

Example 4: IGNOU conducts both face-to-face and distance-mode classes for particular courses indented both to be identical. One sample of 50 students yields examination results mean and SD as:

$$\bar{x}_1 = 80.4 \quad , \quad s_1 = 12.8$$

and other sample of 100 distance-mode students yields mean and SD examination results as

$$\bar{x}_2 = 78.3 \quad , \quad s_2 = 20.5$$

Are the two means of education methods statistically equal at 5% level?

Solution: Given that

$$n_1 = 50, \quad \bar{x}_1 = 80.4, \quad s_1 = 12.8;$$

$$n_2 = 100, \quad \bar{x}_2 = 78.3, \quad s_2 = 20.5$$

We wish to test the null hypothesis that the two means of teaching methods are statistically equal, that is, no difference in both type of teaching methods, or

$$H_0: \mu_1 = \mu_2 \qquad \text{against} \qquad H_1 : \mu_1 \neq \mu_2$$

Since population standard deviations σ_1 and σ_2 are unknown, so we use estimates of σ_1 and σ_2 as sample standard deviations s_1 and s_2 respectively.

It is two tail test and the test statistic Z is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{80.4 - 78.3}{\sqrt{\frac{(12.8)^2}{50} + \frac{(20.5)^2}{100}}} = \frac{2.1}{\sqrt{3.28 + 4.20}} = 0.77$$

The calculated $|Z| = 0.77 <$ tabulated $z_{\alpha/2} = 1.96$ so we accept H_0 at 5% level of significance. We conclude that both methods of education, i.e., face-to-face and distance-mode are statistically equal. No significant difference observed in sample data.

Note: i) Check your answers with those given at the end of the unit.

12) Two brands of electric bulbs are quoted at the same price. A buyer tested a random sample of 200 bulbs of each brand and found the following information:

	Mean life (hrs.)	**S.D. (hrs.)**
Brand A	1300	41
Brand B	1280	46

Is there significant difference in the quality(life) of two brands of electric bulb at 1% level of significance?

13) A marketing company has undertaken an advertisement campaign for a popular breakfast food and claimed that it is an improved product. A survey is conducted to find out the monthly demand of 100 consumers **before and after** the campaign. We have following information.

	Mean	**S.D.**
Before	120	25
After	138	36

Analyze the above data and test whether the campaign was successful?

10.6.3 Test for the Significance of Population Proportion

Sometime we come across the populations divided into two groups or categories and instead of testing the significance of population mean, we may be interested to test the significance of population proportion, for example, proportion of car-owners, smokers, cancer patients, etc., in the concerned population.

In general, let N be the size of a finite population and 'A' be an attribute. Suppose there are M people (or items) among N (M < N) who possess attribute A (or in favor of A).

Let in a random sample of size n < N is drawn from the same population N which contains x people possess attribute A, remaining (n−x) do not possess it where $(0 \le x \le n)$. The scheme is shown in **fig. 10.7.**

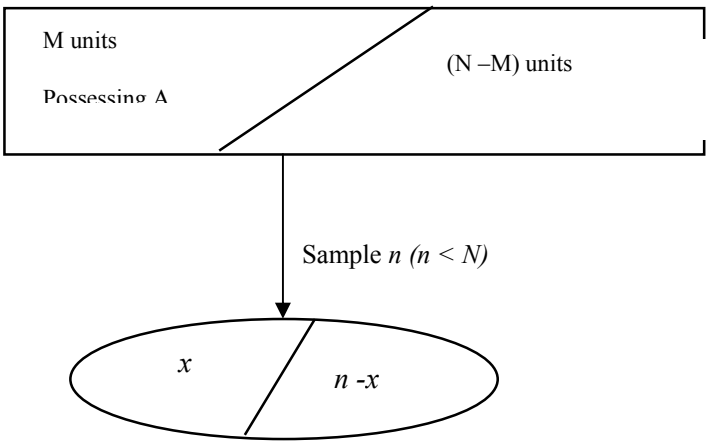


Fig. 10.7: Sample from population possessing attribute A (having x units in favor of A)

Let us denote population and sample proportions, respectively by

$$P = \frac{M}{N} \qquad \text{and} \qquad p = \frac{x}{n}$$

It is shown in the previous unit that E(p) = P. Suppose our problem here is to test the significance of Population proportion, that is, whether in the population the proportion is equal to a given value (say P_0) or not, for instance we want to know whether the proportion of cancer patients is 18%; proportion of literate persons is 27%, etc. Obviously, then the null hypothesis will be H_0: $P = P_0$ and it can be tested, as usual, by selecting a sample of prefixed size from the population.

From the theory of sampling distribution, we observe that for sufficiently large N, the distribution of p is asymptotically normal with mean P and variance $\frac{PQ}{n}$. Hence a test statistic for testing H_0: $P = P_0$ and alternative hypothesis either $H_1 : P \neq P_0$ or H_1: $P > P_0$ or H_1: $P < P_0$ the Z statistic can be written as

$$Z = \frac{p - E(p)}{\sqrt{V(p)}} = \frac{p - P}{\sqrt{\frac{PQ}{n}}} \sim N[0, 1]$$

Two-Tail Test: If $|Z| \geq Z_{\alpha/2}$ we reject H_0 against $H_1 : P \neq P_0$ at α level of significance in a two-tail test and conclude that sample data does not support to the claimed population proportion P_0. When $|Z| < Z_{\alpha/2}$ we do not reject H_0 inferring that claimed or past population proportion still stands as revealed by the sample data.

One-Tail Test: When to test H_0: $P = P_0$ against H_1: $P > P_0$, (right-tail test) we reject H_0 if $Z > Z_\alpha$. It is equivalent to conclude that in the population

proportion possessing A has been significantly improved compare to earlier P_0. In contrary to this, when H_0: $P = P_0$ against H_1: $P < P_0$ (left-tail test) is tested, we reject H_0 if $Z < -Z_\alpha$ at α level of significance. It infers there might be significant reduction in the proportion in the present scenario.

Example 5: A machine produces a large number of items out of which 30% are found to be defective. In a random sample of 100 items, 40 are found to be defective. Is there any evidence of more deterioration of quality at 5% level of significance.

Solution: The company manager knows that his machine produces 30% defective items. The "defects" is an attribute which some objects of population possess, so P_0=0.30, and he is worry whether this has increased presently?

It is given that sample proportion p = 40/100 = 0.40

The null hypothesis would be

H_0: $P = P_0 = 0.30$ against H_1: $P > 0.30$

It is right-tail test and the test statistic is obtained as

$$Z = \frac{p - P_0}{\sqrt{P_0 Q_0 / n}} = \frac{0.40 - 0.30}{\sqrt{(0.30 \times 0.70)/100}} = \frac{0.10}{0.0458} = 2.18$$

The calculated Z= 2.18 is greater than the tabulated value $z_\alpha = 1.64$ at α = 0.05, Hence, we reject null hypothesis H_0 and conclude that there is significant deterioration of quality of items.

Example 6: A dice is thrown 900 times and draw of a 2 or a 5 is observed 450 times. Show that the die cannot be regarded as as unbiased at 5% level of significance.

Solution: If getting a 2 or 5 is a success, then in usual notions, we have from the questions

n = 9000, x = number of successes = 450, p = 450/900 = 1/2

Here, H_0: Dice is unbiased, i.e., Probability of getting a 2 or 5, $P = P_0 = 1/3$, and hence

$$H_0 : P_0 = \frac{1}{3} \quad \text{against} \quad H_1 : P_0 \neq \frac{1}{3}$$

It is two tail test and test statistic Z is

$$Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}} = \frac{(1/2) - (1/3)}{\sqrt{\frac{(1/3) \times (2/3)}{900}}} = \frac{(1/6)}{\sqrt{\frac{2}{8100}}} = 10.6$$

Calculated $|Z| = 10.6 >$ tabulated $z_{\alpha/2} = 1.96$, so H_0 is rejected and we infer that the dice is significantly biased at 5% level of significance.

CHECK YOUR PROGRESS 6

Note: i) Check your answers with those given at the end of the unit.

14) In a sample of 100 M.Sc. Economics 1st year students of IGNOU, it was seen that 54 came from Science background and the rest are from other. Can we assume that students from both backgrounds (i.e. Science and others) have equally performed in the examination at 1% level of significance?

15) Out of 20 patients who are given a particular injection, 18 survived. Test the hypothesis that the survival rate is 80% or more at 5% level of significance?

10.6.4 Test for Equality of Two Population Proportions

Suppose there are two populations, each having persons (or items) possessing attribute A. We are interested to test whether proportions possessing A in both the population are same?

For example, percentage of failed students in two different Sections in a school are same or populations of smokers in two different cities are same or one city exhibit more number smokers than that in another city. We may use notations P_1 and P_2, respectively to denote proportions of units in two populations possessing any attribute A. Obviously, the Null and alternative hypothesis will be:

H_0: Failed percentage in both sections are same (proportion of Smokers in both the cities are same), i.e., $P_1 = P_2$

H_1: $P_1 \neq P_2$

Let samples of sizes n_1 and n_2 be the samples selected independently from the Population-I and Population-II (scheme of selection of samples is depicted in figure 10.7).

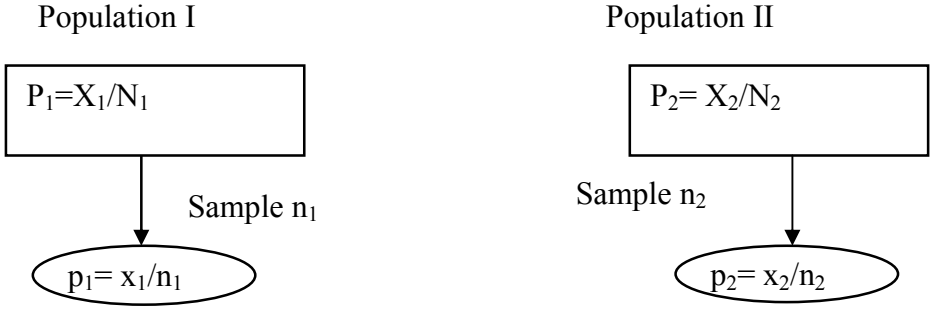


Fig 10.6.1: Two independent samples from different populations possessing attribute A

Let n_1 has x_1 units and n_2 has x_2 units possessing the attribute A (say A = smoking-habit) in samples. Therefore, from the theory of sampling

distribution of $(p_1 - p_2)$, following are the expected values, variances and standard errors of statistic respectively:

$$p_1 = (x_1 / n_1) \; ; \qquad p_2 = (x_2 / n_2)$$

$$E(p_1) = P_1 \; ; \qquad E(p_2) = P_2 \qquad E(p_1 - p_2) = P_1 - P_2$$

$$V(p_1) = \frac{P_1Q_1}{n_1} \; V(p_2) = \frac{P_2Q_2}{n_2} ; \; V(p_1 - p_2) = \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}$$

$$SE(p_1) = \sqrt{\frac{P_1Q_1}{n_1}} ; SE(p_2) = \sqrt{\frac{P_2Q_2}{n_2}} ; SE(p_1 - p_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$$

Under the assumptions of large population and sample sizes due to central limit theorem, the term (p_1-p_2) is distributed asymptotically normal with mean $E(p_1-p_2)$ and variance $V(p_1-p_2)$ and the Z-statistic is:

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\sqrt{V(p_1 - p_2)}} \sim N[0, 1]$$

Under H_0: $P_1 = P_2 = P$ (given value), and H_1: $P_1 \neq P_2$, the test statistic Z is

$$Z = \frac{(p_1 - p_2)}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N[0, 1] \qquad \text{where } Q = 1-P.$$

For H_1: $P_1 \neq P_2$ and if $|Z| < z_{\alpha/2}$, the null hypothesis is not rejected and for $|Z| > z_{\alpha/2}$ the null hypothesis is rejected at α-level of significance.

If P is unknown, its unbiased sample-based estimate $\hat{p}$ is used to replace P in Z as

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \text{ where } \hat{p} = \left(\frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}\right); \quad \hat{q} = 1 - \hat{p}$$

In such case for alterative hypothesis H_1: $P_1 > P_2$ the test statistic Z is used (instead of $|Z|$) and H_0: $P_1 = P_2$ is rejected when $Z \geq z_\alpha$ (right-tail test). Otherwise, if H_1: $P_1 < P_2$ the H_0: $P_1 = P_2$ is rejected when $Z < -z_\alpha$ (left-tail test).

Example 7: In a random sample of 100 persons from town A, 55 are found to be high consumers of wheat. In another sample of 90 from town B, 45 are found to be high consumers of wheat. Do these data reveal a significant difference between town A and town B, so far as the proportions of high wheat consumers are concerns (at α= 0.05).

Solution: Given $x_1 = 55, \quad x_2 = 45, \quad n_1 = 100, \quad n_2 = 90$

Let us set up the hypothesis that the proportions of consumers of wheat in the two towns say P_1 and P_2, are same.

i.e., H_0: = $P_1 = P_2 = P$ against $H_1 : P_1 \neq P_2$

From the data the corresponding sample proportions are obtained as:

$$p_1 = \frac{x_1}{n_1} = \frac{55}{100} = 0.55 \qquad \text{and} \qquad p_2 = \frac{x_2}{n_2} = \frac{45}{90} = 0.50$$

Since, here population proportions P_1 and P_2 are unknown, we obtain the estimate of combined proportion (p) of high wheat consumers in the two towns as:

$$\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{55 + 45}{100 + 90} = \frac{10}{19} \qquad \hat{q} = 1 - \frac{10}{19} = \frac{9}{19}.$$

As per assumption of H_1, it is two tail test. The value of test statistic Z as

$$Z = \frac{p_1 - p_2}{S.E.(p_1 - p_2)} = \frac{p_1 - p_2}{\sqrt{\left[\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]}}$$

$$= \frac{0.55 - 0.50}{\sqrt{\left[\frac{10}{19}.\frac{9}{19}\left(\frac{1}{100} + \frac{1}{90}\right)\right]}} = \frac{0.05}{0.075} = 0.666$$

The calculated $|Z| = 0.666$ is smaller than the tabulated $z_{\alpha/2} = z_{0.025} = 1.96$ at 5% level of significance, so H_0 is not rejected against H_1 and accordingly we conclude that towns A and B do not differ significantly so far as the high wheat consumption is concerned.

Example 8: A machine produced 80 defective articles in a batch of 400. After overhauling it produced 45 defectives in a batch of 300. Has the machine improved due to overhauling? (Take $\alpha = 0.01$).

Solution: Here the population is produced articles and let attribute A stands for defective articles. Let P_1 and P_2, respectively proportions of defective articles before and after overhauling. Since null hypothesis always negate the things, we assume that the proportions of defective articles are same before and after overhauling, that is, H_0: $P_1 = P_2 = P$ (says) against H_1: $P_1 > P_2$ (which means that proportion of defective articles would be smaller than earlier before overhauling).

The sample defective proportions before and after overhauling are:

$$p_1 = \frac{x_1}{n_1} = \frac{80}{400} = 0.20 \qquad \text{and} \qquad p_2 = \frac{x_2}{n_2} = \frac{45}{300} = 0.15$$

Since P unknown, the pooled estimate of proportion is

$$\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{80 + 45}{400 + 300} = \frac{5}{28} \text{ and } \hat{q} = 1 - \hat{p} = 1 - \frac{5}{28} = \frac{23}{28}.$$

It is right-tail test. The calculated value of test statistic Z is obtained as:

$$Z = \frac{p_1 - p_2}{S.E.(p_1 - p_2)} = \frac{p_1 - p_2}{\sqrt{\left[\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]}}$$

$$= \frac{0.20 - 0.15}{\sqrt{\left[\frac{5}{28} \times \frac{23}{28}\left(\frac{1}{400} + \frac{1}{300}\right)\right]}} = \frac{0.050}{0.027} = 1.852$$

The tabulated (or critical) value at 1% level of significance is $z_\alpha = 2.33$ for one right-tail test. Hence, the calculated $Z = 1.852 <$ tabulated $z_\alpha = 2.33$ implying that H_0 is not rejected. Thus, we conclude that overhauling did not improve the situation.

CHECK YOUR PROGRESS 7

Note: i) Check your answers with those given at the end of the unit.

16) The proportion of literates between groups of people of two districts A and B are tested. Out of the 100 persons selected at random from each district, 50 from district A and 40 from district B are found literates. Test whether the observed difference in sample proportion of literates is statistically significant at 1% level of significance?

10.7 LET US SUM UP

In this unit we have discussed:

1. Statistical hypothesis, Null hypothesis, Alternative hypothesis , Simple & Composite hypothesis;
2. Type-I and Type-II errors, Critical region, One tailed and two tailed test;
3. Procedure of testing of hypothesis;
4. Level of significance and degree of freedom;
5. Design of a test of significance;
6. Large sample tests for Population Mean and difference between two Population Means; and
7. Large sample tests for Population Proportion and difference between two Population Proportions.

10.8 KEY WORDS

Null Hypothesis: A statistical hypothesis that usually asserts that nothing special is happening with respect to some characteristic of the underlying population.

One-tailed (or directional) Test: The rejection region is located in just one tail of the sampling distribution.

Two-tailed (or nondirectional) Test: Rejection regions are located in both tails of the sampling distribution.

10.9 SUGGESTED FURTHER READING/ REFERENCES

Witte, R., & Witte, J. (2017). *Statistics*. Hoboken, NJ: John Wiley & Sons.

10.10 ANSWERS TO CHECK YOUR PROGRESS

1) Please refer to section 10.3.1 and 10.3.2.

2) Please refer to section 10.3.3.

3) Please refer to section 10.4.

4) Please refer to section 10.4.

5) Please refer to section 10.5.

6) Please refer to section 10.6.1.

7) Please refer to section 10.6.2.

8) Please refer to section 10.7.

9) Please refer to section 10.8.

10) We first set up our null hypothesis

$$H_0: \mu = 3.25 \text{ cm.} \quad \text{against} \quad H_1: \mu \neq 3.25$$

It is two tail test using one sample.

Given, $\bar{x} = 3.4$ cm, $n = 900$ cm, $\mu = 3.25$ cm and $\sigma = 2.61$ cm

$$Z = \frac{3.40 - 3.25}{2.61/\sqrt{900}} = 1.73$$

Since $|Z|<1.96$, we conclude that H_0 is not rejected at 5% level of significance.

11) We given that

$$\mu_0 = 1200,\ n = 100,\ \bar{x} = 1180,\ s = 90$$

To test the null hypothesis regarding mean

H_0: $\mu = 1200$ against H_1: $\mu < 1200$

The sample is showing lower mean value which is a doubt. It is one tail left test. Then the test statistic Z is

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1180 - 1200}{90/\sqrt{100}} = \frac{-20}{9.0} = -2.22$$

Since calculated $Z > z_\alpha = -2.33$, so we do not reject H_0 at 5% level of significance.

12) Given that

$$n_1 = 200, \quad \bar{x}_1 = 1300, \quad \sigma_1 = 41$$

$$n_2 = 200, \quad \bar{x}_2 = 1280, \quad \sigma_2 = 46$$

We wish to test the null hypothesis

H_0: $\mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ (i.e., mean life of both brands differ)

It is two tail test and test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{1300 - 1280}{\sqrt{\frac{(41)^2}{200} + \frac{(46)^2}{200}}} = \frac{20}{4.36} = 4.59$$

Calculated $|Z| = 4.59 >$ tabulated $z_{\alpha/2} = 2.58$, we reject H_0 at 1% level of significance.

13) Given that

$$n_1 = 100, \quad \bar{x}_1 = 120, \quad s_1 = 25$$

$$n_2 = 100, \quad \bar{x}_2 = 138, \quad s_2 = 36$$

We wish to test the null hypothesis

H_0: $\mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$

It is one tail (left) test and the test statistic Z is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{120 - 138}{\sqrt{\frac{(25)^2}{100} + \frac{(36)^2}{100}}} = \frac{-18}{4.38} = -4.11$$

Since calculated $Z = -4.11 <$ tabulated $- z_\alpha = -2.33$, so we reject H_0 at 5% level of significance.

14) Here, n = 100 and x = Number of students from science background

$$p = \text{proportion of students from Science background} = \frac{54}{100} = 0.54$$

We wish to test the hypothesis: H_0: $P = P_0 = 0.5$ against H_1: $P_0 \neq 0.5$

It is two tail test and test statistics Z is

$$Z = \frac{p - P_0}{\sqrt{P_0 Q_0 / n}} = \frac{0.54 - 0.50}{\sqrt{0.5 \times 0.5 / 100}} = \frac{0.04}{0.05} = 0.80$$

The calculated value $|Z| = 0.80$ is smaller than the tabulated value z = 2.58 at 1% level of significance then we do not reject null hypothesis.

15) Here, the null hypothesis is

$H_0 = P = P_0$ against $H_1 : P > P_0$

$$P_0 = \frac{80}{100} = 0.80 \;\; ; \; Q_0 = \frac{20}{100} = 0.20 \qquad p = \frac{x}{n} = \frac{18}{20} = 0.9$$

Also, by using one tail right test, test statistic Z is

$$Z = \frac{p - P_0}{\sqrt{P_0 Q_0 / n}} = \frac{0.1}{\sqrt{(0.8)(0.2)/20}} = \frac{0.1}{0.089} = 1.12$$

The calculated Z= 1.12is smaller than the tabulated $z_\alpha = 1.645$ at 5% level of significance and we do not reject the null hypothesis.

16) Let p_1 and p_2 stand for proportions of literates in districts A and B respectively. We have to test H_0: $P_1 = P_2 = P$ against $H_1 : P_1 \neq P_2$.

Given

$$p_1 = \frac{x_1}{n_1} = \frac{50}{100} = 0.50 \qquad \text{and} \qquad p_2 = \frac{40}{100} = 0.40$$

Also $$\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{50 + 40}{100 + 100} = 0.45 \qquad \hat{q} = 1 - \hat{p} = 0.55$$

It is two tail test and, we calculate the Z statistic

$$Z = \frac{p_1 - p_2}{S.E.(p_1 - p_2)} = \frac{p_1 - p_2}{\sqrt{\left[\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]}}$$

$$= \frac{0.50 - 0.40}{\sqrt{\left[0.45 \times 0.55\left(\frac{1}{100} + \frac{1}{100}\right)\right]}} = \frac{0.10}{0.049} = 2.04$$

The calculated value $|Z| = 2.04 < 2.58$ (critical value in two tail test at 1% level of significance), then the null hypothesis is not rejected.

UNIT 11 STATISTICAL ANALYSIS-II

Structure

11.1 INTRODUCTION

The entire large sample theory, as applied to testing of hypothesis in the previous unit, is based on the assumption that sample size being large enough (generally more than 30) the sampling distribution of the test statistic can always be approximated by the normal distribution owing to the results of central limit theorem.

You might have observed there that due to this reason, the test statistics used everywhere was Z, where Z stands for the normal distribution with mean zero and variance one. However, if the sample size n is small (practically less than 30), the sampling distribution of many test statistics are far from the normal, and as such, normal distribution is not the basis of those tests. For small samples, therefore, the assumption of normality of the distributions of test statistic is not valid and exact sampling distribution of the test statistic has to be known. Fortunately, in the literature of sampling distributions, many such distributions are obtained which can be applied for small sample problems of testing of hypotheses. In this Unit, we will study various tests of significance based on these statistics.

The small sample tests for testing the significance of population mean and the test for the equality of two population means are very frequent in real life problems. We shall discuss here the relevant testing procedures, namely, t- tests, derived on the basis of t- distribution.

In testing of hypothesis or even in the theory of estimation, we generally assume that the random variable under study follows a particular probability distribution, such as, normal, binomial, Poisson distribution, etc., but such assumptions need to be verified as to whether our assumptions are true or not. We have already discussed and have also seen that the genesis of the problem of testing the hypotheses lies in this fact.

Although, in many testing problems, we frequently deal with certain variables (that is, quantitative characteristics), but in many real-world situations of business and other areas, the collected data are qualitative by nature (that is, attributes), classified into different categories or groups according to one or more attributes. Such types of data are known as "categorical data". For example, the number of persons in a sample may be categorized into different categories of their age, income, job, etc. Then a question arises "how the inference problems, particularly, the problems of testing of hypothesis, arising out of categorical data can be tackled?" In this unit, we shall see that chi- square test, one of the sampling distributions, is helpful for such problems. We shall discuss here two most widely used tests of chi-square.

Sometimes, it is required to test whether the variances of two different populations are equal or not; similar to the situation of testing the equality of population means discussed in the previous unit. We shall discuss this problem and will see how the F-distribution can be applied for this purpose.

11.2 OBJECTIVES

After studying this unit, you should be able to:

- define the meaning of small sample procedures of testing the hypothesis;
- describe the small sample test procedures for single population mean and difference of two population means on the basis of t-distribution;
- describe the concept and procedure of paired t-test;
- explain the chi-square test for the goodness of fit and independence of attributes; and
- describe the test for the equality of two variances using F-distribution.

11.3 PROCEDURE FOR SMALL SAMPLE TESTS

Throughout in this unit unless stated otherwise, we shall assume that the random sample $x_1, x_2,\ldots,x_n$ is a sample of size n (where $n < 30$) selected from a population having parameter θ (which is unknown and to be tested). For developing a test procedure for the parameter θ based upon the small sample, we follow some steps, which are more or less similar as already mentioned in

the previous unit under Section 3.4. For readiness of the material, we briefly present here the steps:

Step I: Formulation of null hypothesis H_0, to be tested against the alternative hypothesis H_1. For example, $H_0: \theta = \theta_0$ against alternative $H_1: \theta \neq \theta_0$ or $H_0: \theta = \theta_0$ against $H_1: \theta > \theta_0$ (or $\theta < \theta_0$), etc.

Step II: Decision about the level of significance, α. Generally, we take α = 5% (or 1%) as the case may be desired.

Step III: Depending upon the nature of the null and alternative hypotheses, selection of an appropriate test statistic for testing H_0, and computation of the test statistic on the basis of the random sample of size n.

Step IV: Depending upon the assumption of the null hypothesis, selection of an appropriate sampling distribution which the test statistic has to follow. The sampling distribution to be used for deciding the critical region (either one-tailed or two-tailed, as the case may be) is always known as soon as we choose our test statistic and formulate our null and alternative hypotheses.

Step V: Using the distribution and value of α, obtain the critical value (or cut-off value) in that sampling distribution.

Step VI: Take the decision for rejecting or not rejecting H_0, in the light of step I and II, in the same way as mentioned in Step VI of Section 3.4, which is as follows:

(i) Case-I: When we take $H_1: \theta \neq \theta_0$ then we reject H_0 if modulus of calculated value of test statistic is greater than or equal to the critical value (or tabulated value) at α level of significance. We do not reject H_0 if modulus of calculated value of test statistic is less than the critical value (or tabulated value) at α level of significance.

(ii) Case-II: When we take $H_1: \theta > \theta_0$ or $H_1: \theta < \theta_0$ then step VI is different and we reject $H_0: \theta = \theta_0$ if test statistic is greater than the critical value. No modulus sign is used for this case for the test statistic. Similarly, if $H_1: \theta < \theta_0$, we reject $H_0: \theta = \theta_0$ when the test statistic is less than the negative of the critical value at α level of significance.

CHECK YOUR PROGRESS 1

Note: i) Use the space given below for your answers.

ii) Check your answers with those given at the end of the unit.

1. Describe the procedure of testing of hypothesis for the small sample.

…….

……

……

……

11.3.1 Test for Population Mean

In some cases, we attempt to test the hypothesis regarding the significance of a population mean when the sample size is small (n<30) and population variance σ^2 is unknown.

For example, for Boarding schools, one can assume that the average score in mathematics paper is 55 (that is, μ_0= 55; μ_0 being the assumed population mean) or average weight of an Indian army person is 60 kg (that is, μ_0 = 60). Our aim here would be to test whether the assumed mean in these populations is same as given or not, where populations are respectively of school children and of army persons.

To test the significance of given population mean, as usual, we assume that a random sample $x_1, x_2,\ldots, x_n$ of size n from the given population (assuming normal) has been drawn. For testing purpose, we set up the hypotheses H_0 (null) and H_1(alternative) as

$$H_0 : \mu = \mu_0 \text{ against } H_1 : \mu > \mu_0 \text{ or } H_1 : \mu < \mu_0 \text{ or } H_1 : \mu \neq \mu_0$$

Let X be the statistic computed on the basis of the random sample $x_1, x_2,\ldots, x_n$; E(X) and V(X) respectively be the mean and variance of X, then under the null hypothesis, the test- statistic t is given by

$$t = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{(n-1)\text{d.f.}}$$

where, E(X) is sample mean $\bar{x} = \frac{\sum x_i}{n}$ and $s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$ is an unbiased estimate of the unknown population variance σ^2. It is well known that statistic t follows a t- distribution with (n-1) degrees of freedom (df).

Substituting the values of $\bar{x}$, μ_0, s and n in the expression of t, we find the calculated value of t. Now we search the critical value (or cut-off value or tabulated value) of the test statistic t from the t-table at (n-1) df at α level of significance. Comparing these two values we take the decision as:

(A) One -Tail Test:

Case I: When $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$ (right-tail test), H_0 is rejected if calculated value of t > tabulated t at (α, (n-1) df).

Case II: When the null hypothesis is H_0: $\mu = \mu_0$ against H_1: $\mu < \mu_0$ (left-tail test), H_0 is rejected if calculated t< −tabulated t at (α, (n-1) df).

(B) Two-Tail Test: When $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$, H_0 is rejected if calculated value of $|t|$ > tabulated t at (α/2, (n-1) df).

Example 1: A hand wash manufacturing company was selling a particular type of brand-hand wash through a large number of retail shops. Before an advertising campaign, the mean sale per weak per shop was 60 dozen. After

the heavy campaign to judge the impact, a random sample of 16 shops was taken and the mean sale was found to be 64 dozen with standard deviation 6. Can you consider the advertisement campaign effective?

Solution: Given that

$$\bar{x} = 64,\ \mu_0 = 60,\ s = 6 \text{ and } n = 16$$

For testing H_0: $\mu = \mu_0$ (=60) against H_1:$\mu > 60$ dozens (right- tail test) the test statistic t is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{64 - 60}{6/\sqrt{16}} = \frac{4}{1.5} = 2.67$$

The critical value (or tabulated value) of t for (n-1) = 15 df at 5% level of significance [(one- tail right- hand test)] is $t_{15,0.05} = 2.602$ (obtained from the t- table). Since the calculated value is greater than tabulated value, we reject H_0 at 5% level of significance, that is, $H_1$is accepted and we conclude that advertising campaign is effective for increasing the bath-soap sale in the market.

Example 2: The mean Share price of Real estate companies is Rs.68. After a month, these Share prices of all companies have been changed. A sample of 10 Pharma companies was taken and their share prices were noted as: 70, 76, 75, 69, 70, 72, 68, 65, 75, 72. Test, whether mean share price is still the same?

Solution: Given sample size n = 10 (< 30), so it is a small sample case and we wish to test the null hypothesis that the population mean is still 70.0 besides all changes, that is,

H_0: $\mu = \mu_0 = 68$ against H_1:$\mu \neq \mu_0 = 68$

The test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

For calculating the value of t, we prepare the following table:

S. No.	Sample value (x)	Deviations d= (x -a), with a=70	d^2
1	70	0	0
2	76	6	36
3	75	5	25
4	69	-1	1
5	70	0	0
6	72	2	4

7	68	-2	4
8	65	-5	25
9	75	4	16
10	72	2	4
Total		11	115

The assumed value is 70, so we get

$$\bar{x} = 70 + \frac{\sum d}{n} = 70 + \frac{11}{10} = 70.1$$

$$s^2 = \frac{1}{n-1}\left[\left\{\sum d^2 - \frac{\left(\sum d\right)^2}{n}\right\}\right] = \frac{1}{10-1}\left[115 - \frac{(11)^2}{10}\right] = \frac{1}{9}\left[115 - \frac{121}{10}\right] = 11.43$$

$$\Rightarrow s = \sqrt{11.43} = 3.38$$

It is a two-tail test and the test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{70.1 - 68}{3.38/\sqrt{10}} = \frac{2.10}{1.07} = 1.962$$

The tabulated value of t for n -1 = 9 df and at 5% level of significance, that is, $t_{9,0.025}$ is 2.262.

Since the calculated value of t is smaller than the tabulated value, we do not reject H_0 at 5% level of significance and, therefore, conclude that mean prices of the Shares in the given month is still 68 besides many ups and downs.

CHECK YOUR PROGRESS 2

Note: i) Check your answers with those given at the end of the unit.

2) Describe the procedure of small sample test for single population mean.

3) An automobile tyre manufacturer claims that the average life of a particular category of his tyre is 18000 kms when used under normal driving conditions. A random sample of 16 tyres was tested. The mean and S.D. of life in the sample were 20,000 and 60,000 kms., respectively. Assuming the life of the tyre in kms to be normally distributed, decide whether the manufacture's claims are true at 1% level of significance.

4) A random sample of weights in kg of10 students of IGNOU are given as 48, 50, 62, 75, 80, 60, 70, 56, 52, 78. Can we say that mean of the distribution of weights of all students of IGNOU from which the above sample was taken is equal to 60 kg.

11.3.2 Tests for Equality of Two Population Means

Suppose we have to compare means of two groups when sample sizes taken from these two groups are small. For example, performance of English-medium and Hindi-medium school children in mathematics paper are to be compared or performance of two brand-promotion campaigns of a product in two different markets are to be compared.

Let us assume that two independent random samples $x_1, x_2, ..., x_{n_1}$ and $y_1, y_2, ..., y_{n_2}$ of sizes n_1 and n_2 respectively are drawn from two normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. Further, suppose the variances of the both populations are unknown but are equal, i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (say) where σ^2 is an unknown value.

In this situation, we want to test the null hypothesis

$H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ or $H_1 : \mu_1 > \mu_2$ or $H_1 : \mu_1 < \mu_2$

We calculate the statistic t as (assuming H_0 true):

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \sim t_{(n_1+n_2-2)}$$

Here, $\bar{x}$ and $\bar{y}$ are the means of first and second sample respectively and

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \text{ where, } s_1^2 = \frac{\sum(x-\bar{x})^2}{n_1-1} \text{ and } s_2^2 = \frac{\sum(y-\bar{y})^2}{n_2-1}$$

After substituting values of $\bar{x}$, $\bar{y}$, s, n_1 and n_2 in the expression of t, we obtain the calculated value of t. Then we look for critical value (or cut-off value or tabulated value) of test statistic t from the t-table. On comparing these two values, we take the decision as below:

[A] One-Tail Test:

Case I: When $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$ (right-tail test), H_0 is rejected if calculated value of t > tabulated t at (α, ($n_1 + n_2$-2) df).

Case II: When $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$ (left-tail test), H_0 is rejected when if calculated value of t < −tabulated t at (α, ($n_1 + n_2$-2) df).

[B] Two-Tail Test: When $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$, H_0 is rejected if calculated value of $|t|$ > tabulated t at (α/2, ($n_1 + n_2$-2) df).

Example 3: In a random sample of 10 pigs fed by diet A, the increase in weight in pounds in a certain period were 12, 8, 14, 16, 13, 12, 8, 14, 10, 9 lbs. In another random sample of 10 pigs fed by diet B, the increase in weights in the same period were14, 13, 12, 15, 16, 14, 18, 17, 21, 15 lbs. Test

whether diets A and B differ significantly regarding their effect on increase in weight.

Solution: Obviously here the hypothesis of no difference in mean weights of pigs under two diets will be tested, that is,

$$H_0: \mu_1 = \mu_2 \text{ against } H_1: \mu_1 \neq \mu_2.$$

In order to find the value of t-statistic, we prepare the following table:

Diet A			**Diet B**		
x	**d_1 = (x-a); with a=12**	**d_1^2**	**y**	**d_2 = (y-b); with b=16**	**d_2^2**
12	0	0	14	-2	4
8	-4	16	13	-3	9
14	2	4	12	-4	16
16	4	16	15	-1	1
13	1	1	16	0	0
12	0	0	14	-2	4
8	-4	16	18	2	4
14	2	4	17	1	1
10	-2	4	21	5	25
9	-3	9	15	-1	1
	$\sum d_1 = -4$	$\sum d_1^2 = 70$		$\sum d_2 = -5$	$\sum d_2^2 = 65$

We calculate the test statistic t:

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

Here, a=12, b= 16 are assumed means

$$\bar{x} = a + \frac{\sum d_1}{n_1} = 12 + \frac{-4}{10} = 11.6, \qquad \bar{y} = b + \frac{\sum d_2}{n_2} = 16 + \frac{-5}{10} = 15.5$$

$$s^2 = \frac{1}{n_1 + n_2 - 2}\left[\left\{\sum d_1^2 - \frac{\left(\sum d_1\right)^2}{n_1}\right\} + \left\{\sum d_2^2 - \frac{\left(\sum d_2\right)^2}{n_2}\right\}\right]$$

$$= \frac{1}{10+10-2}\left[\left\{70 - \frac{(-4)^2}{10}\right\} + \left\{65 - \frac{(-5)^2}{10}\right\}\right] = \frac{68.4 + 62.5}{18} = 7.27$$

$$s = \sqrt{7.27} = 2.70$$

Putting these values in test statistic t, we get

$$t = \frac{11.6 - 15.5}{2.70\sqrt{\left[\frac{1}{10} + \frac{1}{10}\right]}} = \frac{-3.90}{2.70 \times 0.45} = \frac{-3.90}{1.21} = -3.23$$

Also, calculated $|t| = 3.23$ and tabulated $t_{n_1+n_2-2,\alpha/2} = t_{18,0.025} = 2.101$

The calculated value of t is greater than the tabulated value so H_0 is rejected at 5% level of significance and we conclude that diets A and B differ significantly in terms of increase in weights of pigs. Definitely any one diet performs better to other significantly.

Example 4: The means of two random samples of sizes 10 and 8 are 210.40 and 208.42 respectively. The sum of squares of the deviations from means is 26.94 and 24.50, respectively. Can samples be considered to have been drawn from the normal populations having equal mean.

Solution: Given that

$$n_1 = 10, \quad n_2 = 8, \quad \bar{x} = 210.40, \quad \bar{y} = 208.92,$$

$$\sum_{i=1}^{n_1}(x_i - \bar{x})^2 = 26.94, \qquad \sum_{i=1}^{n_2}(y_i - \bar{y})^2 = 24.50,$$

So

$$s^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2\right] = \frac{1}{16} \times 51.44 = 3.215$$

$$s = \sqrt{3.215} = 1.79$$

We wish to test the null hypothesis that both the samples are drawn from the normal populations having the same mean. Let μ_1 and μ_2 be means of the first and second populations respectively, then the null hypothesis will be

H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 \neq \mu_2$.

It is a two- tail test and we calculate the test-statistic t as

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} = \frac{210.40 - 208.92}{1.79\sqrt{\left[\frac{1}{10} + \frac{1}{8}\right]}} = \frac{1.48}{1.79 \times 0.47} = \frac{1.48}{0.85} = 1.74$$

Also, tabulated value $t_{n_1+n_2-2,\alpha/2} = t_{16,0.25} = 2.12$

The calculated value of t is smaller than the tabulated value so we do not reject H_0 at 5% level of significance. We, therefore, conclude that both samples are from the normal populations having same means.

CHECK YOUR PROGRESS 3

Note: i) Use the space given below for your answers.
ii) Check your answers with those given at the end of the unit.

5) Describe the procedure of small sample test for difference between two population means.

6) Two different types drugs A and B were tried on some patients for increasing their weights. The 6 persons were given drug A and other 7 persons were given drug B. The increase in weight (in ponds) is given below:

Drug A	5	8	7	10	9	6	-
Drug B	9	10	15	12	14	8	12

Do the two drugs differ significantly with regard to their mean weight increment?

7) To test the effect of fertilizer on wheat production, 26 plots of land with equal areas were chosen. Half of these plots were treated with fertilizer and the other half were untreated. Other conditions were the same. The mean yield of wheat on the untreated plots was 4.6 quintals with a standard deviation of 0.5 quintals, while the mean yield of the treated plots was 5.0 quintals with standard deviations of 0.3 quintal. Can we conclude that there is significant improvement in wheat production because of the effect of fertilizer at 1% level of significance?

11.3.3 Paired t-Test

In the sub-section 11.3.2, for testing the difference of two population means, we assumed that the populations are different from which the samples are drawn independently. Sometimes, we may come across with the problems of testing the equality of means within the same population when two samples at different times or occasions are taken on same units.

For instance, if we wish to test the effectiveness of a new diet marketed for reducing weight, the weights of a small group of individuals (say, 15 individuals) may be recorded before and after administering the diet which will give rise to two samples of same size. Here the observations may be grouped into pairs for each individual. The hypothesis to be tested is again same as considered in sub-section 4.2.2, that is, there is no effect of administering the diet, or equivalently, that the average weights before and after administering the diet are same. The sample values taken before and after administering the diet (denoted by x_i and y_i respectively for i = 1, 2, …, n) may be put in the following manner for n individuals:

Sample I: $x_1, \quad x_2, \quad x_3, \quad \ldots \quad x_n$

Sample II: $y_1, \quad y_2, \quad y_3, \quad \ldots \quad y_n$.

Whenever we face such type of experiments for testing the equality of two population means, we apply t-test but due to paired observations, the test is called a paired t- test. The procedure of testing under paired t-test is described below:

Using the notations of sub-section 4.2.2, we wish to test

$$H_0: \mu_1 = \mu_2 \quad \text{or} \quad \overline{D} = (\mu_1 - \mu_2) = 0$$

for which the test statistic t is

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \sim t_{(n-1)\text{d.f.}}$$

where $d_i = (x_i - y_i) \ \forall \ i = 1, 2, ..., n$ and similarly, $\bar{d} = \frac{1}{n}\sum d_i = (\bar{x} - \bar{y})$

$$s_d^2 = \frac{1}{n-1}\left[\sum (d_i - \bar{d})^2\right] = \frac{1}{n-1}\left[\sum d_i^2 - \frac{(\sum d_i)^2}{n}\right]$$

We need to calculate only Σd_i , Σd_i^2 for obtaining the value of t. Then we obtain the critical value (or cut-off value or tabulated value) of test statistic t from the t-table.

The methods of taking decision about the rejection of H_0 or not rejecting H_0, on the basis of calculated and tabulated values of t-statistic, are exactly similar as mentioned in sub-section 4.2.2 for H_1: $\mu_1 > \mu_2$; $\mu_1 < \mu_2$ and $\mu_1 \neq \mu_2$ except that here degrees of freedom (df) is (n – 1) instead of ($n_1 + n_2$-2).

Note: Sometimes data are recorded in the form of increments/decrements only, like; weight of child increased by 2 kg, marks increased by 3%, temperature reduced to 3^0C ,etc. Then actually for n objects, this set of data show [d_1 , d_2 ,d_3 , .., d_n]. Accordingly, our hypothesis will be

$$H_0 : \overline{D} = 0 \quad \text{against} \quad H_1 : \overline{D} > 0 \quad \text{or} \quad H_1 : \overline{D} < 0$$

The sample with larger mean may be treated as the first sample and hence is denoted by x. Therefore, this has to be checked before attempting the paired-t test that whether values (x_i, yi) are given or increments/ decrements are given in the data set.

Note: In good practice, if we talk about increments then we take $d_i = (y_i - x_i)$, and alternative hypothesis $H_1 : \overline{D} > 0$ (one sided right-tail test). For decrements we take d_i = (xi-y_i) and same alternative $H_1 : \overline{D} > 0$

Example 5: A group of 12 children was tested to find out how many digits they would repeat from memory after hearing them once. They were given practice session for this test. Next week they were retested. The results obtained were as follows:

Child no.	1	2	3	4	5	6	7	8	9	10	11	12
Recall Before	6	4	5	7	6	4	3	7	8	4	6	5
Recall After	6	6	4	7	6	5	5	9	9	7	8	7

Can the memory practice session improve the performance of children?

Solution: We set up the null hypothesis that the children have no improvement after the practice session, that is, no effect due to practice or both populations before and after practice have same mean,

So, we have H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 < \mu_2$ (one sided left- tail test)

The μ_1 and μ_2 stand for the average before and after the memory practice session of test respectively. In this example, same 12 children are tested, hence, for testing H_0, we use paired t-test. For computing t value, we do the following calculations:

Child No.	**Digit Recall**		**d = (x-y)**	**d^2**
	Before (x)	**After (y)**		
1	6	6	0	0
2	4	6	-2	4
3	5	7	-2	4
4	7	7	0	0
5	6	8	-2	4
6	4	5	-1	1
7	3	5	-2	4
8	7	9	-2	4
9	8	9	-1	1
10	4	7	-3	9
11	6	8	-2	4
12	5	7	-2	4
			$\sum d = -19$	$\sum d^2 = 35$

$$\bar{d} = \frac{\sum d}{n} = \frac{-19}{12} = -1.58$$

$$s_d^2 = \frac{1}{n-1}\sum\left(d_i - \bar{d}\right)^2 = \frac{1}{n-1}\left[\sum d_i^2 - \frac{\left(\sum d_i\right)^2}{n}\right]$$

$$= \frac{1}{11}\left[35 - \frac{(-19)^2}{12}\right] = \frac{1}{11} \times 4.92 = 0.45$$

$$\Rightarrow \quad s_d = \sqrt{0.45} = 0.67$$

Substituting the values, we have test statistic t as

$$t = \frac{-1.58}{0.67/\sqrt{12}} = \frac{-1.58}{0.1934} = -8.17$$

The critical (tabulated) value of t for 11 df at 5% level of significance is $t_{11,0.05}$ = -1.796 (one-sided left-test). Since calculated value of t is smaller than the tabulated value, so H_0 is rejected at 5% level of significance and we conclude that the children have improvement after the memory practice session.

CHECK YOUR PROGRESS 4

Note: i) Check your answers with those given at the end of the unit.

8) Describe the procedure of paired t-test for small sample.

9) To verify whether a course "Post Graduate Certificate in Climate Change (PGCCC)" improved performance, a similar test was given to 10 participants both before and after the course. The original marks out of 100 (before course) recorded in a alphabetical order of the participants are 42, 46, 50, 36, 44, 60, 62, 43, 70 and 53. After the course the marks are in the same order 45, 46, 60, 42, 60, 72, 63, 43, 80 and 65. Test whether the course PGCCC was useful?

11.4 CHI-SQUARE TESTS

In the previous units, we discussed parametric tests in which we first assumed the distribution of the parent population and then performed a test about some parameter(s) of the population(s) such as mean, variance, proportion, etc. Generally, the parametric tests are based on the assumption that the parent population is a normal population and, thus, involves one or more parameters for its complete specification. In this section, we shall discuss about some χ^2 (chi-square) tests. Some chi square tests are generally known as non-parametric tests, that is, we do not presume that the given set of data have come from any specific distribution (particularly from a normal distribution). Thus, there is no question of involving a parameter in the test.

11.4.1 Test for Goodness of Fit

The chi-square (χ^2) test for goodness of fit was given by Karl Pearson in 1900, which is the oldest non-parametric method of testing of hypothesis. In this test, given a set of observed frequencies for values of a discrete variable, we test whether the data follows a specific probability distribution or not.

Under the assumption of a particular distribution for the given data (as for example, binomial, Poisson, etc.), which is considered to be the null hypothesis to be tested; expected frequencies for the given variable values are obtained which would be expected to occur if the data follows the assumed distribution. This test is known as "goodness of fit test" since the aim is to observe how close are the given values of the variable to the frequencies yielded by the assumed probability distribution, or, in other words, we judge the fitness of the data to the assumed theoretical distribution. This is a non-parametric test since we do not assume any specific probability distribution beforehand for the data.

Assumptions for the test:

1. Sample observations are independent.
2. The measurement scale is at least nominal.
3. The observations may be classified into non-overlapping categories.

Let a random sample of size N be drawn from a population with unknown distribution of k characteristics and the data categorized into k groups or classes. Also, let $O_1, O_2, \ldots, O_k$ be the observed frequencies (as given in the data set) and $E_1, E_2, \ldots, E_k$ be the corresponding expected frequencies as are expected if the data follows the assumed distribution. We generally put a linear constraint of equality on the observed and expected frequencies which is $\sum O_i = \sum E_i$.

To perform the chi- square goodness of fit test the steps are as follows:

Step 1: First of all, we form the null and alternative hypothesis. Null hypothesis is that the given data set follows the given probability distribution or the pattern specified in the question. Alternative hypothesis contradicts this assumption.

Step 2: For computing expected frequencies corresponding to given values of the variable, for the assumed distribution (mentioned in the null hypothesis), we compute those required sample statistic(s) using the sampled data which are necessary to obtain an estimate of the corresponding population parameter(s) and assume that it is equal to the theoretical value of the corresponding parameter. This, in fact, provides us an estimated value of the parameter which is generally not known in most of the cases. However, if value(s) of parameters are given, they can be used directly for the purpose and there is no need to estimate it on the basis of the given sample data. For instance, let mean (and/or variance) of the theoretical distribution is needed for calculating the expected frequencies, we compute mean (and/or variance) of the sample for this purpose.

Step 3: In the next step, we find the probability that an observation falls (belongs) to a particular category (to a particular value of the variable) using the assumed probability distribution.

Step 4: The calculated probabilities are then used to find the corresponding expected frequencies using the result

$$E_i = Np_i ; \qquad \text{for all } i = 1, 2, \ldots, k$$

Step 5: For testing the null hypothesis, compute the test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(k-r-1)}$$

The χ^2 statistic follows chi-square distribution with (k –r –1) degrees of freedom where, r is the number of parameters estimated from the sample.

Step 6: Now obtain the tabulated value of χ^2 at the specified level of significance and (k-r-1) degrees of freedom from the table and compare it with the calculated value and take the decision as follows:

If computed value of test statistic is greater than tabulated value with $(k-r-1)$ degrees of freedom and at α level of significance, we may reject the null hypothesis, otherwise we shall not reject it.

Note: Table 5 in Appendix provides the critical value of chi-square distribution at α level of significance.

Example 6: The following data are collected during a test to determine consumer preference among five leading brands of soaps:

Brand Preferred	A	B	C	D	E
Number of Customers	194	205	204	196	202

Test the hypothesis that the preference is uniform over the five brands at the 5% level of significance.

Solution: Here, we are interest to test the null hypothesis that the preference of customers over five brands is uniform, that is, all brands are equally preferred by customers. Therefore, we have

H_0: The preference of customers over the five brands of bath soap is uniform

H_1: The preference of customers over the five brands of bath soap is not uniform

Here, equivalently we can say that a uniform distribution is assumed for the data., that is, the proportions of customers for each brand is same. In other words, $p_1 = p_2 = p_3 = p_4 = p_5 = p = \frac{1}{5}$

where p_i (i = 1, 2,…, 5) is the proportion of customer for the i^{th} brand.

For testing the null hypothesis, the test statistic χ^2 as mentioned above.

The theoretical or expected number of customers or frequency for each brand is obtained by multiplying the appropriate probability by total number of customers, that is, sample size N. Therefore,

$$E_1 = E_2 = E_3 = E_4 = E_5 = Np = 1000 \times \frac{1}{5} = 200$$

For calculating the value of test statistic, we prepare the following table:

Brand	Observed Frequency (O)	Expected Frequency (E)	(O−E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
A	194	200	-6	36	0.18
B	205	200	5	25	0.13
C	204	200	4	16	0.08
D	196	200	−4	16	0.08
E	202	200	1	1	0.01
Total	1000	1000			0.48

Therefore, the calculated value of the test statistic is 0.48.

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 0.48$$

The tabulated value of chi-square with $k-1=5-1=4$ degrees of freedom and 5% level of significance is 7.78. Since computed value of test statistic is less than tabulated value χ^2_α at 5% level of significance, therefore, we do not reject the null hypothesis at and conclude that the preference of customers over the five brands of bath soap is uniform.

CHECK YOUR PROGRESS 5

Note: i) Check your answers with those given at the end of the unit.

10) Describe the chi-square goodness of fit test.

11) The following table gives the numbers of aircraft accidents that occurred during the various days of the week:

Days	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Number of Accidents	14	15	8	20	11	9	14

Test whether the accidents are uniformly distributed over the week.

11.4.2 Test for Independence of Two Attributes

In the analysis of research data, the investigator often needs to decide whether two variables or attributes are related (associated). For example, a sociologist may wish to know whether level of formal education is associated

with income, a quality-control engineer may be interested to know whether the number of defective items produced on the various production lines is independent of the day of the week, a doctor may wish to know whether a drug is effective in curing colds, etc. If there is no association between two variables, we say that they are independent. In other words, we can say that two variables are independent if the distribution of one is not depending on the distribution of other. Similarly, in case we are given two attributes instead of two variables and we want to test the independence of two attributes, we may use chi-square test for independence. This test will indicate only whether or not any association exists between attributes. The null hypothesis in such a test is always that the attributes are independent against the alternative hypothesis that they are not independent. Symbolically, we write

H_0: The two attributes are independent;

H_1: The attributes are not independent.

To conduct the test, a sample is drawn from the population and the observed frequencies are cross-classified according to the two criteria so that each observation belongs to one and only one level of each criterion. The cross-classification can be conveniently displayed by mean of a table called a contingency table. Therefore, a contingency table is an arrangement of data into a two-way classification. One of the classifications is entered in rows and the other in columns.

Assumptions:

1. The sample of N observation is random sample.
2. Each observation in the sample may be classified according to two criteria so that each observation belongs to one and only one level of each criterion.

Let the N observations are classified into the following type of contingency table where we have p levels of the attribute A and q levels of the attribute B. Here O_{ij} (i = 1, 2,…, p and j = 1, 2,…, q) represents the observed frequency of the cell (i, j).

Table

B / A	B_1	B_2	… B_j …	B_q	Total of rows
A_1	O_{11}	O_{12}	… O_{1j} …	O_{1q}	R_1
A_2	O_{21}	O_{22}	… O_{2j} …	O_{2q}	R_2
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
A_i	O_{i1}	O_{i2}	… O_{ij} …	O_{iq}	R_i
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
A_p	O_{p1}	O_{p2}	… O_{pj} …	O_{pq}	R_p
Total of columns	C_1	C_2	… C_j …	C_q	N

To calculate the expected frequencies corresponding to each O_{ij}, we use the multiplication law of probability. According to this law if two events are independent then the probability of their joint occurrence is equal to the product of their individual probabilities. Therefore, if two criteria of classification of observations in the contingency table are independent, the joint probability for a cell is equal to the product of the two corresponding marginal probabilities. Since, row totals are showing the marginal probabilities $P(A_i)$; i = 1, 2,…,p and column totals are showing probabilities $P(B_j)$; j = 1, 2,…, q; the probability in the cell (i., j) is equal to $P(A_i)P(B_j)$. Clearly, we know that $P(A_i) = R_i/N$ and $P(B_j) = C_j/N$.

Thus, we have

$$P[A_i B_j] = \frac{R_i}{N} \cdot \frac{C_j}{N}$$

To obtain expected frequency E_{ij} for each cell, we multiply this estimated probability by the total sample size. Thus,

$$E_{ij} = N \cdot \frac{R_i}{N} \cdot \frac{C_j}{N}$$

which reduces to

$$E_{ij} = \frac{R_i \times C_j}{N} = \frac{\text{Sum of } i^{th} \text{ row} \times \text{Sum of } j^{th} \text{ column}}{\text{Total sample size}}$$

Thus, it is evident that calculation of expected frequencies is quite simple.

Once each E_{ij} is calculated, for testing the null hypothesis H_0, we use the same χ^2- statistics, which is

$$\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(p-1)(q-1)}$$

Here the test statistic follows ~~as~~ a chi-square distribution with (p−1) (q−1) degrees of freedom.

The decision about rejecting or not rejecting the null hypothesis of independence of the two attributes against the alternative hypothesis is exactly same as mentioned in the sub-section 11.4.1 above.

Example 7: Calculate the expected frequencies for the following data presuming the two attributes condition of child and condition of home and check whether these are independent.

Condition of Child	Condition of Home	
	Clean	Dirty
Clear	75	45
Fairly clean	85	15
Dirty	40	40

Solution: Here, we want to test the null hypothesis that condition of home and condition of the child are independent, that is,

H_0: Condition of child is independent of condition of home against

H_1: Condition of child depends upon the condition of home.

For the calculation of χ^2 statistic, we do the following computations for expected frequencies:

Condition of Child	Condition of Home		Total
	Clean	Dirty	
Clear	75	45	120
Fairly clean	85	15	100
Dirty	40	40	80
Total	200	100	300

Under H_0, E_{ij}'s are obtained as

$$E_{11} = \frac{R_1 \times C_1}{N} = \frac{120 \times 200}{300} = 80;\ E_{12} = \frac{R_1 \times C_2}{N} = \frac{120 \times 100}{300} = 40;$$

$$E_{21} = \frac{R_2 \times C_1}{N} = \frac{100 \times 200}{300} = 66.67 \approx 67;\ E_{22} = \frac{R_2 \times C_2}{N} = \frac{100 \times 100}{300} = 33.33 \approx 33;$$

$$E_{31} = \frac{R_3 \times C_1}{N} = \frac{80 \times 200}{300} = 53.33 \approx 53;\ E_{32} = \frac{R_3 \times C_2}{N} = \frac{80 \times 100}{300} = 26.67 \approx 27;$$

For calculating the value of χ^2 test statistic, we prepare the following table:

Observed Frequency (O_{ij})	**Expected Frequency (E_{ij})**	**(O – E)**	**$(O-E)^2$**	**$\frac{(O-E)^2}{E}$**
75	80	−5	25	0.3125
45	40	5	25	0.6250
85	67	18	324	4.8350
15	33	−18	324	9.8180
40	53	−14	169	3.1880
40	27	14	169	6.2590
Total				25.0375

Therefore,

$$\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 25.0375$$

The degrees of freedom will be (p –1)(q –1) = (3 – 1)(2 – 1) = 2.
The tabulated value of chi square at 2 df and 5% level of significance is 5.99.

Since calculated value of test statistic is greater than tabulated value, we reject our null hypothesis and conclude that condition of home does affect the condition of child.

CHECK YOUR PROGRESS 6

Note: i) Check your answers with those given at the end of the unit.

12) Discuss chi-square test for testing independence of attributes.

13) 1500 families were selected at random in a city to test the belief that high income families usually send their children to public schools and low-income families often send their children to government schools. The following results were obtained in the study conducted.

Income	Public School	Government School	Total
Low	300	600	900
High	400	200	600
Total	700	800	1500

Use chi-square test at 1% level to state whether the two attributes are independent.

11.5 F-TEST

11.5.1 TEST FOR EQUALITY OF TWO VARIANCES

While applying t-distribution for testing H_0: $\mu_1 = \mu_2$, the basic assumption was: the population variances of the two populations are equal. But, in fact, it is a very rare case in practice. Therefore, sometimes it is necessary to test the equality of population variances first. The F distribution is used for ascertaining the equality of two population variances. We shall describe here the F-statistic for this purpose which follows the F distribution.

Let $x_1, x_2, ..., x_{n_1}$ be a sample of size n_1 from a normal population with mean μ_1 and variance σ_1^2.Similarly, let $y_1, y_2, ..., y_{n_2}$ be a sample of size n_2 from another normal population with mean μ_2 and variance σ_2^2. We will show how the two samples are used for defining F-statistic. According to the objective of the test, the null hypothesis will be

$$H_0 : \sigma_1^2 = \sigma_2^2$$

against the alternative hypothesis

$$H_1 : \sigma_1^2 \neq \sigma_2^2 \quad \text{or} \quad H_1 : \sigma_1^2 > \sigma_2^2 \quad \text{or} \quad H_1 : \sigma_1^2 < \sigma_2^2$$

To test H_0, we calculate the test statistic

$$F = \frac{s_1^2}{s_2^2} \sim F_{(n_1-1, n_2-1)}; \text{ where}$$

$$s_1^2 = \frac{1}{n_1 - 1}\sum (x - \bar{x})^2 \quad \text{and} \quad s_2^2 = \frac{1}{n_2 - 1}\sum (y - \bar{y})^2.$$

The sampling distribution of F-statistic so defined, follows a F-distribution with $\nu_1 = (n_1 - 1)$ and $\nu_2 = (n_2 - 1)$ degrees of freedom. As a norm, larger variance is taken in the numerator of F-ratio and the degree of freedom corresponding to this variance is taken as ν_1 .The sample with larger s^2 is treated as the first sample s_1^2 with population variance σ_1^2.

Using the calculated of F-statistic, we take the decision about H_0 as:

[A] For One-Tail Test: (i) When $H_0 : \sigma_1^2 = \sigma_2^2$ with $H_1 : \sigma_1^2 > \sigma_2^2$

H_0 is rejected when calculated value of $F > F_{(n_1-1),(n_2-1)}(\alpha)$, the tabulated value, at α% level of significance and ν_1 and ν_2 df, otherwise, H_0 is not rejected.

(ii) When $H_0 : \sigma_1^2 = \sigma_2^2$ with $H_1 : \sigma_1^2 < \sigma_2^2$

H_0 is rejected when calculated value of $F < F_{(n_1-1),(n_2-1)}(1-\alpha)$, not rejected otherwise.

[B] For Two -Tail Test: When $H_0 : \sigma_1^2 = \sigma_2^2$ and $H_1 : \sigma_1^2 \neq \sigma_2^2$

H_0 is rejected when calculated value of $F > F_{(n_1-1),(n_2-1)}(\alpha/2)$ or $F < F_{(n_1-1),(n_2-1)}(1-\alpha/2)$, otherwise not rejected.

Example 8: Two random samples gave the following results:

	Size	**Mean**	**Sum of Squares of deviation from the Mean**
Sample I	9	59	26
Sample II	11	60	32

Test whether both samples are from the same normal population?

Solution: Since we have to test whether both the samples are from same normal population, therefore, we will test two things separately:

(i) The equality of two population means,

(ii) The equality of two population variances.

That is $H_0 : \mu_1 = \mu_2$ first and then test $H_0 : \sigma_1^2 = \sigma_2^2$

The equality of two means will be tested using t-test whereas equality of two variances will be tested using F-test. But since t-test is based on the prior assumption that both population variances are same, therefore, first we apply F-test and later the t-test (when F-test accepts equality of variance hypothesis).

Given that $n_1 = 9, \quad \bar{x} = 59, \quad \sum(x - \bar{x})^2 = 26$

$n_2 = 11, \quad \bar{y} = 60, \quad \sum(y - \bar{y})^2 = 32$

we get

$$s_1^2 = \frac{1}{n_1 - 1}\sum(x - \bar{x})^2 = \frac{1}{9-1} \times 26 = 3.25 \qquad s_2^2 = \frac{1}{n_2 - 1}\sum(y - \bar{y})^2$$

$$= \frac{1}{11-1} \times 36 = 3.60$$

For $H_0 : \sigma_1^2 = \sigma_2^2$, $H_1 : \sigma_1^2 > \sigma_2^2$ (right-tail test), the test statistic is: $F = \frac{s_1^2}{s_2^2}$

Since $s_2^2 > s_1^2$ therefore we take reverse $F = \frac{s_2^2}{s_1^2} \sim F_{(n_2-1, n_1-1)} = (3.60/3.25) = 1.1$

The tabulated values at 1% and 5% level of significance are $F_{10, 80(.01)} = 3.34$ and $F_{10, 8(0.05)} = 5.81$ (one sided left-tail test) respectively. Since calculated F is smaller than tabulated F at both level of significance; H_0 is not rejected and we conclude that the variances of two populations are same. Now, applying t-test, the equality of two population means can be tested as discussed earlier in 4.2.2.

Example 9: The following data relate to the number of items produced in a shift by two workers A and B for some days:

A	26	37	40	35	30	30	40	26	30	35	45
B	19	22	24	27	24	18	20	19	25		

Can it be inferred that A is more stable (or consistent) worker compared to B?

Solution: We wish to test the null hypothesis that worker A and B are equally efficient in terms of stability of item production. Let σ_1^2 and σ_2^2 be the variances for the two workers A and B respectively. Then our hypothesis will be

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{against} \quad H_1 : \sigma_1^2 < \sigma_2^2$$

For calculating s_1^2 and s_2^2, we first find two means which are found to be 34 and 22 respectively. Then we prepare following table:

Items produced by A (variable x)	$(x-\bar{x})$ $=(x-34)$	$(x-\bar{x})^2$	**Items produced by B (variable y)**	$(y-\bar{y})$ $=(y-22)$	$(y-\bar{y})^2$
26	-8	64	19	−3	9
37	3	9	22	0	0
40	6	36	24	2	4
35	1	1	27	5	25
30	-4	16	24	2	4
30	-4	16	18	−4	16
40	6	36	20	−2	4
26	-8	64	19	−3	9
30	-4	16	25	3	9
35	1	1			
45	11	121			
Total 374	**0**	**380**	**198**	**0**	**80**

The test statistic is

$$F = \frac{s_1^2}{s_2^2}$$

where s_1^2 and s_2^2 are as same as given above.

Using the values obtained in the table, we have

$$s_1^2 = \frac{1}{n_1 - 1}\sum (x - \bar{x})^2 = \frac{1}{10} \times 380 = 38 \quad s_2^2 = \frac{1}{n_2 - 1}\sum (y - \bar{y})^2 = \frac{1}{8} \times 80 = 10$$

Therefore, for one sided right tailed test

$$F = \frac{38}{10} = 3.8$$

The tabulated value of $F_{10, 8\,(0.01)} = 3.34$ at 1% level of significance. Since the calculated value of F is greater than the tabulated value, we reject the null hypothesis and conclude that the worker A is more stable than worker B.

CHECK YOUR PROGRESS 7

Note: i) Check your answers with those given at the end of the unit.

14) Two sources raw materials are under consideration by a bulb manufacturing company. Both sources seem to have similar

characteristics but the company is not sure about their respective uniformity. A sample of 12 lots from source A yields a variance 125 and a sample of 10 lots from source B yields a variance of 112. Is it likely the variance of source A significantly differ to the variance of source B at significance level $\alpha = 0.01$?

15) Two samples are drawn from two different normal populations.

Sample I	60	65	71	74	76	82	85	87	-	-
Sample II	61	66	67	85	78	63	85	86	88	91

Test the equality of two population variances at 1% level of significance.

11.6 LET US SUM UP

In this unit, we discussed some testing procedures for different kinds of hypotheses based on the assumption that the sample(s) collected were small in sizes so that the assumption of normality of the data were not fulfilled. The tests described are as follows:

1. Small sample tests for testing the hypotheses using t-tests for (i) significance of population mean and (ii) equality of two population means;
2. Paired t-test when the two samples are selected from the same population on same number of units with the aim to test the equality of means;
3. Chi-square tests for (i) testing the goodness of fit of a theoretical probability distribution on a given set of data and (ii) testing the independence of two attributes with different levels of classification;
4. Test based on F distribution for testing the equality of two population variances.

11.7 KEY WORDS

Degrees of freedom (df): The number of values free to vary, given one or more mathematical restrictions.

11.8 SUGGESTED FURTHER READING/ REFERENCES

Witte, R., & Witte, J. (2017). *Statistics*. Hoboken, NJ: John Wiley & Sons.

11.9 ANSWERS TO CHECK YOUR PROGRESS

1) Please refer to section 11.3

2) Please refer to section 11.4

3) We are given that

$\mu_0 = 18{,}000,\ \ n = 16,\ \ \bar{x} = 20{,}000,\ \ \ s = 6{,}000$

and we wish to test the null hypothesis

H_0: $\mu = 18{,}000$ against H_1: $\mu > 18{,}000$

The test statistic is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{(20000 - 18000)}{6000/\sqrt{16}} = \frac{2000}{1500} = 1.33$$

The critical (tabulated) value of t for (n-1) = 15d.f. at 1% level of significance [(one tail (right)] is $t_{15,0.01} = 2.602$. Since the calculated value is smaller than tabulated value, we do not reject H_0 at 1% level of significance.

4) Given that past value of mean weight is 60 kg. which is like an old information. We wish to test the null hypothesis that

$$H_0 : \mu = \mu_0 = 60 \ \text{ against } \ H_1 : \mu \neq \mu_0 = 60$$

The test statistic is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{(n-1)}$$

From the data one can calculate: Mean = (630/10) = 63

$$s^2 = \frac{1}{n-1}\sum (x - \bar{x})^2 = \frac{1252}{9} = 139.11 \Rightarrow s = \sqrt{139.11} = 11.79$$

The test statistic is

$$t = \frac{(63 - 60)}{11.79/\sqrt{10}} = \frac{3}{37.26} = 0.80$$

The tabulated value of t for n – 1 = 10–1 = 9 degree of freedom for 5% level or $t_{9,0.025}$ is 2.262 (two tail test). Since the calculate value of t is smaller than the tabulated value, we do not reject H_0 at 5% level of significance.

5) Please refer to section 11.5

6) Setting the hypothesis that there is no difference between diets A and B, i.e., H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 \neq \mu_2$,

we calculate the test statistic:

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

From the data we may have:

$$\bar{x} = 8 + \frac{\sum d_1}{n_1} = 8 + \frac{-3}{6} = 7.5 \quad \bar{y} = b + \frac{\sum d_2}{n_2} = 12 + \frac{-4}{7} = 11.43$$

,

$$s^2 = \frac{1}{n_1 + n_2 - 2}\left[\left\{\sum d_1^2 - \frac{\left(\sum d_1\right)^2}{n_1}\right\} + \left\{\sum d_2^2 - \frac{\left(\sum d_2\right)^2}{n_2}\right\}\right]$$

$$= \frac{1}{6+7-2}\left[\left\{19 - \frac{(-3)^2}{6}\right\} + \left\{39 - \frac{(-4)^2}{7}\right\}\right] = \frac{17.5 + 36.71}{11} = 4.93$$

$$s = \sqrt{4.93} = 2.22$$

Putting the values in test statistic

$$t = \frac{7.5 - 11.43}{2.22\sqrt{\left(\frac{1}{6} + \frac{1}{7}\right)}} = \frac{-3.93}{2.22 \times 0.56} = \frac{-3.93}{1.24} = -3.17$$

Calculated $|t| = 3.17$ and tabulated $t_{11,0.025} = 2.201$ (two tail at 5 %)

The calculated value is greater than the tabulated value so H_0 is rejected at 5% level of significance (two tail test).

7) Given that:

$$n_1 = 13, \quad \bar{x} = 4.6, \quad s_1 = 0.5,$$

$$n_2 = 13, \quad \bar{y} = 5.0, \quad s_2 = 0.3$$

$$s^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2\right] = \frac{1}{n_1 + n_2 - 2}\left[n_1 s_1^2 + n_2 s_2^2\right]$$

$$= \frac{1}{13+13-2}\left[13 \times (0.5)^2 + 13 \times (0.3)^2\right] = \frac{1}{24} \times 4.42 = 0.18$$

$$\Rightarrow s = \sqrt{0.18} = 0.43$$

We want to test the null hypothesis that the there is no significant improvement in wheat production due to fertilizer, that is

H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 \neq \mu_2$

We calculate the test statistic

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} = \frac{4.6 - 5.0}{0.43\sqrt{\left[\frac{1}{13} + \frac{1}{13}\right]}} = \frac{-0.4}{0.17} = -2.35$$

Tabulated value $t_{24,\ 0.005}$=-2.792 at α=1% . The calculated value is greater than the tabulated value, so we do not reject H_0.

8) Please refer to section 11.6

9) We set up the null hypothesis

H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 < \mu_2$ (one-tail left)

So, to test the given H_0, we use paired t-test for which the test statistic is:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \sim t_{(n-1)\text{d.f.}}$$

From the data one can calculate:

$$\bar{d} = \frac{\sum d}{n} = \frac{-70}{10} = -7.0$$

$$s_d^2 = \frac{1}{n-1}\left\{\sum d^2 - \frac{\left(\sum d\right)^2}{n}\right\} = \frac{1}{9}\left[790 - \frac{(-70)^2}{10}\right] = \frac{1}{9}\times 300 = 33.33$$

$$\Rightarrow \quad s_d = \sqrt{33.33} = 5.77$$

Putting the values, we have test statistic

$$t = \frac{-7.0}{5.77/\sqrt{10}} = -3.83$$

The critical (tabulated) value of t for 9 df at 5% level of significance is $t_{9,0.05}$ = -1.833 (one tail left test). Since calculated value of t is smaller than the tabulated value, so H_0 is rejected at 5% level of significance.

10) Please refer to section 11.4.

11) Here, we are interest to test that the null hypothesis

H_0: The accidents are uniformly distributed over the week

against the alternative hypothesis

H_1: The accidents are not uniformly distributed over the week

For testing the null hypothesis, the test statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(k-1)}$$

Since the uniform distribution is one in which all outcomes considered to have an equal or uniform probability. Therefore, the probability that the accident occurs in each day is same. Thus,

$$p_1 = p_2 = p_3 = p_4 = p_5 = p = \frac{1}{7}$$

The theoretical or expected frequency for each day is obtained by multiplying the appropriate probability by the total number of accidents, that is, sample size N. Therefore,

$$E_1 = E_2 = E_3 = E_4 = E_5 = Np = 91 \times \frac{1}{7} = 13$$

Therefore, the test statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 7.6923$$

The tabulated value of chi-square with $k-1=7-1=6$ degrees of freedom and at 1% level of significance is 16.81. Since computed value of test statistic is less than tabulated value at 1% level of significance therefore, we do not reject the null hypothesis.

12) Please refer to section 11.5.

13) Here, we want to test the null hypothesis

H_0: Two attributes are independent

against the alternative hypothesis

H_1: Two attributes are not independent

For testing the null hypothesis, that test statistic is

$$\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(p-1)(q-1)}$$

Now, under H_0, the expected frequencies can be obtained as:

$$E_{ij} = \frac{R_i \times C_j}{N} = \frac{\text{Sum of } i^{th} \text{ row} \times \text{Sum of } j^{th} \text{ column}}{\text{Total sample size}}$$

Therefore,

$$E_{11} = \frac{R_1 \times C_1}{N} = \frac{900 \times 735}{1500} = 441,\ E_{12} = \frac{R_1 \times C_2}{N} = \frac{900 \times 765}{1500} = 459,$$

$$E_{21} = \frac{R_2 \times C_1}{N} = \frac{600 \times 735}{1500} = 294,\ E_{22} = \frac{R_2 \times C_2}{N} = \frac{600 \times 765}{1500} = 306$$

Therefore,

$$\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 220.98$$

The degrees of freedom will be (p –1)(q –1) = (2 – 1)(2 – 1) =1. The tabulated value of chi square 1df and α=1% is 6.63. Since calculated value of χ^2 is greater than tabulated value at 1% level of significance therefore, we reject our null hypothesis.

14) Given that $n_1 = 12, \quad s_1^2 = 125; \quad n_2 = 10, \quad s_2^2 = 112$

We wish to test the null hypothesis that the **variances of source A and B are same,** that is

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{against} \quad H_1 : \sigma_1^2 > \sigma_2^2$$

which means one source is larger to other. The test statistic is

$$F = \frac{s_1^2}{s_2^2}$$

Therefore,

$$F = \frac{125}{112} = 1.116$$

The tabulated value of $F_{11,\,9,\,(0.01)} = 3.10$(one tail right) at 1% level of significance. Since the calculated value of F is smaller than the tabulated value of F, therefore, we do not reject the null hypothesis.

15) We wish to test the null hypothesis

$$H_0 : \sigma_2^2 = \sigma_1^2 \quad \text{against} \quad H_1 : \sigma_2^2 > \sigma_1^2$$

The test statistic is

$$F = \frac{s_1^2}{s_2^2}$$

From the data we have obtained:

$$\bar{x} = \frac{\sum x}{n_1} = \frac{600}{8} = 75, \qquad \bar{y} = \frac{\sum y}{n_2} = \frac{770}{10} = 77$$

$$s_1^2 = \frac{1}{n_1 - 1}\sum (x - \bar{x})^2 = \frac{1}{7} \times 636 = 90.86$$

$$s_2^2 = \frac{1}{n_2 - 1}\sum (y - \bar{y})^2 = \frac{1}{9} \times 1200 = 133.33$$

Since $s_2^2 > s_1^2$ therefore we take reverse of the F- statistic

$$F = \frac{s_2^2}{s_1^2} = \frac{133.33}{90.86} = 1.47$$

Now the tabulated value of $F_{9,7\,(0.05)} = 5.21$ (one tail right at 5% level).

Since the calculated value of F is smaller than the tabulated value of F, therefore, we do not reject the null hypothesis.

UNIT 12 ANALYSIS OF VARIANCE TESTS

Structure

12.1 INTRODUCTION

In Unit 4 of this block, we tested the equality of means of two independent groups by using t-test. Sometimes situations may arise where testing of more than two means is required. As an example, in crop-cutting experiments it may be required to test whether under similar conditions the average yield of some crop in a number of fields is same or not. For obvious reasons, in such cases, t-test cannot be applied.Generally, for such situations, the technique of Analysis of Variance (ANOVA) is used, in which the testing of equality of several means is doe by dividing the population variability into different components. The usual F-test is used to test the equality of means of several groups.

As its name suggests, the analysis of variance focuses on variability. It involves the calculation of several measures of variability, when the total variability of the population is divided into many components, like, variability within the smaller sub-groups, variability between the smaller sub-groups, etc. In other words, ANOVA is a technique which split up the total variation of data which may be attributed to various “sources” or “causes” of variation. There may be variation between variables and also within different levels of variables. In this way, Analysis of Variance is used to test the

homogeneity of several population means by comparing the variances between the sample and within the sample.

In this unit, we shall discuss the one-way as well as two-way Analysis of Variance. One-way Analysis of Variance is a technique where only one independent variable at different levels is considered which affects the response variable whereas in two-way Analysis of Variance technique, we will consider two variables at different levels which affect the response variables.

12.2 OBJECTIVES

After studying this unit, you should be able to:

- understand the Analysis of Variance technique;
- describe various types of assumptions underlying the Analysis of Variance technique and applications of it;
- define various types of linear models used in Analysis of Variance technique;
- understand how to test the hypothesis under One-way Analysis of Variance; and
- explain the method of performing Two-way ANOVA Test.

12.3 ANALYSIS OF VARIANCE (ANOVA)

The statistical technique known as "Analysis of Variance"(abbreviated as ANOVA), was propounded by Professor R.A. Fisher in 1920's, in which he developed the method of testing equality of means of several sub-populations by dividing the total variability into different components. Variation is inherent in nature, so analysis of variance means examining the variation present in data or parts of data. In other words, analysis of variance means to find out the cause of variation in the data. The total variation in any set of numerical data of an experiment is due to number of causes which may beAssignable causes and Chance causes.

The variation in the data due to assignable causes can be detected, measured and controlled whereas the variation due to chance causes is not in the control of human being and cannot be traced or find out separately.

According to Professor R.A. Fisher, Analysis of Variance (ANOVA) is the "separation of variance ascribable to one group of causes from the variance ascribable to other group". So, by this technique, the total variation present in the data is divided into two components of variation; one due to assignable causes (Between the Groups Variability) and the other due to chance causes (Within Group Variability). Analysis of variance technique can be classified as Parametric ANOVA and Non-Parametric ANOVA.The topic of this unit is related to parametric ANOVA.

Parametric ANOVA can be classified as simply ANOVA if only one response variable is considered. If more than one response variables are under consideration than it is called multivariate analysis of variance (MANOVA).

If we consider, only one independent variable which affects the response/dependent variable then it is called One-way ANOVA. If the independent variables/explanatory variables are more than one, say n, then it is called n-way ANOVA. If n is equal to two than the ANOVA is called Two-way classified ANOVA.

12.3.1 Significance of Analysis of Variance

One obvious question may arise that why do we call it Analysis of Variance, even though we are testing for the equality of means? Why do not we simply call it the Analysis of Means? How do we propose test for means by analysis the variances? As a matter of fact, in order to determine if means of several populations are equal, we do consider the measure of variance, σ^2.

The estimate of population variance σ^2 is computed by two different estimates of it each one by a different method. One approach is to compute an estimator of σ^2 in such a manner that even if the population means are not equal it will have no effect on the value of this estimator. This means that the differences in the values of the population means does not affect the value of σ^2 as calculated by a given method. This estimator of σ^2 is the average of the variances found within each of the samples. For example, if we take 10 samples of size n, then each sample will have a mean and a variance. Then the mean of these 10 variances would be considered as an unbiased estimator of σ^2, the population variance, and its value remains appropriate irrespective of whether the population means are equal or not. This is really done by pooling all the sample variances to estimate a common population variance, which is the average of all sample variances. This common variance is known as variance within sample or σ^2_{within}

The second approach to calculate the estimate of σ^2 is based upon the Central Limit Theorem and is valid only under the null hypothesis assumption that all the population means are equal. This means that in fact, if there are no differences among the population means, then the computed value of σ^2 by the second approach should not differ significantly from the computed value of σ^2 by the first approach. Hence, if these two values of σ^2 are approximately the same, then we can decide to accept the null hypothesis of equality of means.

The second approach results in the following computation:

Based upon the Central Limit Theorem, we have previously found that the standard error of the sample means is calculated by:

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

or, the variance would be:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \qquad \text{or,} \qquad \sigma^2 = n\sigma_{\bar{x}}^2$$

Thus, by knowing the square of the standard error of the mean $(\sigma_{\bar{x}})^2$, we could multiply it by n and obtain a precise estimate of σ^2. This approach of estimating σ^2, is known as $\sigma^2_{between}$. Now, if all population means are equal then, $\sigma^2_{between}$ value should be approximately the same as $\sigma^2_{between}$ value. A significant difference between these two values would lead us to conclude that this difference is the result of difference between the population means.

But, how do we know that any difference between these two values is significant or not? How do we know whether this difference, if any, is simply due to random sampling error or due to actual differences among the population means?

R. A. Fisher developed a Fisher test or F-test to answer the above question. He determined that the difference between the $\sigma^2_{between}$ and σ^2_{within} could be expressed as a ratio to be designated as the F-value, so that

$$F = \frac{\sigma^2_{between}}{\sigma^2_{within}}$$

In the above, case, if the population means are exactly the same, then $\sigma^2_{between}$ will be equal to the σ^2_{within} and the value of F will be equal to 1.

However, because of sampling errors and other variations, some disparity between these two values will be there, even when the null hypothesis is true, meaning that all population means are equal. The extent of disparity between the two variances and consequently, the value of F, will influence our decision on whether to accept or reject the null hypothesis. It is logical to conclude that if the population means are not equal then their sample means will also vary greatly from one another, resulting in a large value of $\sigma^2_{between}$, and hence a larger value of F (σ^2_{within} is based only on sample variances and not on sample means and hence is not affected by differences in sample means). Accordingly, larger the value of F statistic, the more likely the decision is to reject the null hypothesis. But how large the value of F be so as to reject the null hypothesis? The answer is that the computed value of F must be larger than the critical value of F, given in the table for a given level of significance and calculated number of degrees of freedom. (The F distribution is a family of curves, so that there are different curves for different degrees of freedom.)

12.3.2 Degrees of Freedom

We have talked about the F-distribution being a family of curves, each curve reflecting the degrees of freedom relative to both $\sigma^2_{between}$ and σ^2_{within}. This

means that the degrees of freedom are associated both with the numerator as well as with the denominator of the F-ratio.

Since the variance between samples $\sigma^2_{between}$ comes from many samples and if there are k number of samples, then the degrees of freedom, associated with the numerator would be (k – 1).

The denominator is the mean variance of the variances of k samples and since each variance in each sample is associated with the size of the sample (n), then the degrees of freedom associated with each sample would be (n – 1). Hence, the total degrees of freedom would be the sum of degrees of freedom of k sample or df = k(n – 1), when each sample is of size n.

12.3.3 Uses of ANOVA

The following are some of the uses of ANOVA:

1. **To test the homogeneity of several means (say, k groups) that is, H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$**

   If H_0 is rejected then we can say that there is a significant difference between these k groups or there is a significant effect of these k independent variables.

2. **To test the relationship between two variables**

   This test provides evidence that dependent variable Y and independent variable X are related in their movements. If Y do not relate with Xthen we expect H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$, which is the null hypothesis for testing the absence of relationship.

3. **Test for Linearity of Regression**

   If in 2, some relationship is established, the next step is to find the appropriate regression function. At the first stage we try to find out whether the linear regression fits the observed data, that is the null hypothesis will be

   $$H_0: \mu_i = \alpha + \beta X_i$$

   with the sample model $Y_{ij} = \mu_i + e_{ij}$, when α and β are the parameters of the model.

4. **Test for Polynomial Regression**

   If a linear relationship is not established, we proceed to test the hypothesis

   $$H_0: \mu_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_k X_i^k$$

   That is, the relationship between X and Y can be explained by a polynomial of degree k.

5. **Some Other Uses of ANOVA**

   - Test of Homogeneity of a Group of Regression Coefficients.
   - Test for Equality of Regression Equations from p Groups.
   - Test for Multiple Linear Regression Model.

CHECK YOUR PROGRESS 1

Note: i) Use the space given below for your answers.

ii) Check your answers with those given at the end of the unit.

1. Describe the Analysis of Variance and differentiate between One-way and Two-way ANOVA.

……

……

……

……

2. Explain briefly the uses of Analysis of Variance.

……

……

……

……

12.4 ONE-WAY ANALYSIS OF VARIANCE (ANOVA)

One-factor analysis of variance (or one-way analysis of variance) is a special case of ANOVA.Whereas two sample t-test is used to decide whether two groups (two levels) of a factor have the same mean; one-way ANOVA generalizes this problem to k levels (greater than two) of a factor.

In the following, subscript i refers to the i^{th} level of the factor and subscript j refers to the j^{th} observations within a level of factor. For example, y_{23} refers to third observation in the second level of a factor.

The observations on different levels of a factor can be exhibited as shown below:

Level of a factor	**Observations**			**Totals**	**Means**
1	y_{11}	y_{12}.....	y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}.....	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
.			.	.	.
.			.	.	.
.			.	.	.
k	y_{k1}	y_{k2}.....	y_{kn}	$y_{k.}$	$\bar{y}_{k.}$

A linear mathematical model for one-way classified data is used which can be written as

$$y_{ij} = \mu_i + e_{ij} \quad \text{where} i = 1, 2, \ldots, k; j = 1, 2, \ldots, n.$$

Here, y_{ij} is continuous dependent or response variable, whereas μ_i is discrete independent variable, also called an explanatory variable. Total number of observations is

$$n_1 + n_2 + \ldots + n_k = \sum_{i=1}^{k} n_i = N$$

In fact, this model decomposes the responses into a mean for each level of a factor and error term, that is,

Response = A mean for each level of a factor + error term

The analysis of variance provides estimates for each level mean. These estimated level means are the predicted values of the model and the difference between the response variable and the estimated/predicted level means are the residuals.

That is,

$$y_{ij} = \mu_i + e_{ij} \text{ implying that } \quad e_{ij} = y_{ij} - \mu_i$$

The above model can also be re-written as $y_{ij} = \mu + (\mu_i - \mu) + e_{ij}$

or $$y_{ij} = \mu + \alpha_i + e_{ij}, \quad \forall\ i = 1, 2, ..., k \text{ and } j = 1, 2, ..., n_i$$

where, $\alpha_i = \mu_i - \mu$.

12.4.1 Assumptions

The following are the basic assumptionsof one-way ANOVA:

1. Dependent variable measured on interval scale;
2. k samples are independently and randomly drawn from the population;
3. Population can be assumed reasonably to have a normal distribution;
4. k samples have approximately equal variance;
5. Various effects are additive in nature; and
6. e_{ij} are independently and identically distributed (i.i.d.) normal variables with mean zero and variance σ_e^2.

12.4.2 Procedure of One-Way ANOVA Test

Now, when we discuss the step-by-step computation procedure for one way analysis of variance for k independent samples, the first step of the procedure is to make the null and alternative assumptions.

Null Hypothesis

We want to test the equality of the population means, that is, to test the homogeneity of effect of different levels of a factor. Hence, the null hypothesis is given by

$$H_0: \quad \mu_1 = \mu_2 = \ldots = \mu_k$$

against the alternative hypothesis

$$H_1: \quad \mu_1 \neq \mu_2 \neq \ldots \neq \mu_k \quad \text{(or some } \mu_i\text{'s are not equal)}$$

Determining Level of Significance:

The next is to decide a criterion for acceptance or rejection of null hypothesis i.e., cut-off value and level of significance α, at which we have to test our

null hypothesis. As discussed in Unit 3 and 4, level of significance is a fixed assumed value which is decided before starting the test procedure. The most commonly used values of α are 0.05 (5%) and 0.01 (1%). **Let us take it as α = 5% (or 1%) if not given** anywhere or follow whatever given for α.

Computation of Ratio of Variations F:

Since here F ratio contains only two elements, which are the variances between the samples and within the samples respectively, as discussed before, let us recapitulate the calculation of these variances.

If all the means of samples were exactly equal and all samples were exactly representative of their respective populations so that all the sample means were exactly equal to each other and to the population mean, then there will be no variance. However, this can never be the case. We always have variation both between samples and within samples, even if we take these samples randomly and from the same population. This variation is known as the total variation.

The total variation designated by $\sum\left(X-\overline{\overline{X}}\right)^2$, where X represents individual observations for all samples and $\overline{\overline{X}}$ is the grand mean of all sample means and equals μ, the population mean, is also known as the total sum of squares or SST, and is simply the sum of squared differences between each observation and the overall mean. This total variation represents the contribution of two elements. These elements are:

a) **Variance between samples.** The variance between samples may be due to the effect of different treatments, meaning that the population means may be affected by the factor under consideration, thus making the population means actually different, and some variance may be due to the inter-sample variability. This variance is also known as the sum of squares between samples. Let this sum of squares be designed as SSB.
Then SSB is calculated by the following steps:

Step-I: Take k samples of size n each and calculate the mean of each sample, designated as $\overline{X}_1, \overline{X}_2, \overline{X}_3, ..., \overline{X}_k$.

Step-II: Calculate the grand mean $\overline{\overline{X}}$ of the distribution of these sample means, so that

$$\overline{\overline{X}} = \frac{\sum_{i=1}^{k} \overline{X}_i}{k}$$

Step-III: Take the difference between the means of the various samples and the grand mean, which can be denoted as

$$(\overline{X}_1 - \overline{\overline{X}}), (X_2 - \overline{\overline{X}}), (\overline{X}_3 - \overline{\overline{X}}), ..., (\overline{X}_k - \overline{\overline{X}})$$

Step-IV: Square these deviations or differences individually, multiply each of these squared deviations by its respective sample size and sum up all these products, so as to find

$$\sum_{i=1}^{k} n_i \left(\overline{X}_i - \overline{\overline{X}}\right)^2, \text{ where } n_i = \text{size of the } i^{th} \text{ sample.}$$

This will be the value of the SSB.

However, sometimes individual observations in each k samples are not available, but means of the samples are given, then the step 1 can be skipped and we start the computation procedure from step 2. Divide SSB by the degrees of freedom, which are (k – 1), where k is the number of samples and this would give us the value of $\sigma^2_{between}$, so that:

$$\sigma^2_{between} = \frac{SSB}{(k-1)}$$

This is also known as mean square between samples or MSB.

b) **Variance within samples.** Even though each observation in a given sample comes from the same population and is subjected to the same treatment, some chance variation can still occur. The variance may be due to sampling errors or other natural causes. The variance or sum of squares is calculated through the following steps:

Step-I: Calculate the mean value of each sample, that is, $\overline{X}_1, \overline{X}_2, \overline{X}_3, ..., \overline{X}_k$.

Step-II: Take one sample at a time and take the deviation of each item in the sample from its mean. Do this for all the samples, so that we would have a difference between each value in each sample and their respective means for all values in all samples.

Step-III: Square these differences and take a total sum of all these squared differences (or deviations). This sum is also known as SSW or sum of squares within samples.

Step-IV: Divide this SSW by the corresponding degrees of freedom. The degrees of freedom are obtained by subtracting the total number of samples from the total number of items.

Thus if N is the total number of items or observations, and k is the number of samples, then, df = (N – k) which are the degrees of freedom within samples (If all samples are of equal size n, then df = k(n – 1), since (n – 1) are the degrees of freedom for each sample and there are k samples).

This figure SSW/df is also known as σ^2_{within}, or MSW (mean of sum of squares within samples).

Now the value of F can be computed as:

$$F = \frac{\sigma^2_{between}}{\sigma^2_{within}} = \frac{SSB/df}{SSW/df} = \frac{SSB/(k-1)}{SSW/(N-k)} = \frac{MSB}{MSW}$$

Construction of ANOVA Table

After the various calculations for SSB, SSW and the degrees of freedom have been, made, these figures can be presented in a simple table called Analysis of Variance table or simply ANOVA table, as follows;

ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	Variance Ratio F
Treatment	SSB	(k -1)	$MSB = \frac{SSB}{(k-1)}$	$F = \frac{MSB}{MSW}$
Within	SSW	(N – k)	$MSW = \frac{SSW}{(N-k)}$	
Total	SST			

Then, the variance ratio is

$$F = \frac{MSB}{MSW}$$

This calculated value of F is then compared with the critical value of F, obtained from the F-table at respective d.f. and a decision is made about the validity of null hypothesis. The critical value of F from the table for α level of significance and degrees of freedom as follows:

df(numerator) = (k – 1)

df (denominator) = (N – k)

This value of F from the table is compared with calculated value of F and if calculated value of F is less than the critical value of F, we cannot reject the null hypothesis.

Now after discussing the procedure of One-way ANOVA test let us practice to solve some examples.

Example 1: To test whether all professors teach the same material in different sections of the introductory statistics class or not, 4 sections of the same course were selected and a common test was administered to 5 students selected at random from each section. The scores for each student from each section were noted and are given below. We want to test for any differences in learning, as reflected in the average scores for each section.

Student #	Section 1 Scores (X_1)	Section 2 Scores (X_2)	Section 3 Scores (X_3)	Section 4 Scores (X_4)
1	8	12	10	12
2	10	12	13	15
3	12	10	11	13
4	10	8	12	10
5	5	13	14	10
Totals	$\sum X_1 = 45$	$\sum X_2 = 55$	$\sum X_3 = 60$	$\sum X_4 = 60$
Means	$\overline{X}_1 = 9$	$\overline{X}_2 = 11$	$\overline{X}_3 = 12$	$\overline{X}_4 = 12$

Solution: The method is as follows:

1) State the null hypothesis. We are assuming that there is no significant difference among the average scores of students from these 4 sections and hence all professors are teaching the same material with the same effectiveness, i.e.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$.

H_1 : All means are not equal or at least two means differ from each other

2) Establish a level of significance. Let $\alpha = 0.05$.

3) Calculate the variance between the samples, as follows:

   a) The mean of each sample is,

   $\overline{X}_1 = 9,\ \overline{X}_2 = 11,\ \overline{X}_3 = 12,\ \overline{X}_4 = 12$

   b) The grand mean or $\overline{\overline{X}}$ is:

   $$\overline{\overline{X}} = \frac{\sum \overline{X}}{n} = \frac{9+11+12+12}{4} = 11$$

   c) Calculate the value of SSB.

   $$SSB = \sum n\left(\overline{X} - \overline{\overline{X}}\right)^2$$
   $$= 5(9-11)^2 + 5(11-11)^2 + 5\ (12-11)^2 + 5\ (12-11)^2$$
   $$= 20 + 0 + 5 + 5 = 30$$

   d) The variance between samples $\sigma^2_{between}$ or MSB is given by:

   $$MSB = \frac{SSB}{df} = \frac{(30)}{(k-1)} = \frac{(30)}{3} = 10$$

4) Calculate the variance within samples, as follows:

   The find the sum of squares within samples (SSW) we square each deviation between the individual value of each sample and its mean, for all samples and then sum these squares deviations, as follows:

$$\sum\left(X_i - \overline{X}_i\right)^2 = \sum\left(X_1 - \overline{X}_1\right)^2 + \sum\left(X_2 - \overline{X}_2\right)^2 + \sum\left(X_3 - \overline{X}_3\right)^2 + \sum\left(X_4 - \overline{X}_4\right)^2$$

We have the mean of each sample as

$\overline{X}_1 = 9,\ \overline{X}_2 = 11,\ \overline{X}_3 = 12,\ \overline{X}_4 = 12$

Thus

$$\sum\left(X_i - \overline{X}_i\right)^2 = (8-9)^2 + (10-9)^2 + (12-9)^2 + (10-9)^2 + (5-9)^2$$
$$+ (12-11)^2 + (12-11)^2 + (10-11)^2 + (8-11)^2 + (13-11)^2$$
$$+ (10-12)^2 + (13-12)^2 + (11-12)^2 + (12-12)^2 + (14-12)^2$$
$$+ (12-12)^2 + (15-12)^2 + (13-12)^2 + (10-12)^2 + (10-12)^2$$
$$= 1 + 1 + 9 + 1 + 16 + 1 + 1 + 1 + 9 + 4 + 4 + 1 + 1 + 0 + 4$$
$$+ 0 + 9 + 1 + 4 + 4$$
$$= 28 + 16 + 10 + 18 = 72$$

Then SSW = 28 + 16 + 10 + 18 = 72

Now, the variance within samples, σ^2_{within}, or MSW is given by

$$MSW = \frac{SSW}{df} = \frac{SSW}{(N-K)} = \frac{72}{20-4} = \frac{72}{16} = 4.5$$

Then the F-ratio $= \frac{MSB}{MSW} = \frac{10}{4.5} = 2.22$.

We can construct an ANOVA table for the problem solved above as follows:

ANOVA Table

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**Variance Ratio F**
Treatment	SSB = 30	(k – 1) = 3	$MSB = \frac{SSB}{(k-1)}$ $MSB = \frac{30}{3} = 10$	$\frac{MSB}{MSW}$ = 2.22
Within (or error)	SSW = 72	(N – k) = 16	$MSW = \frac{SSW}{(N-k)}$ $MSW = \frac{72}{16} = 4.5$	
Total	SST = 102			

Now, we check for the critical value of F from the table for α = 0.05 and degrees of freedom as follows:

df(numerator) = (k – 1) = (4 – 1) = 3

df (denominator) = (N – k) = (20 – 4) = 16

This value of F from the table is given as 3.24. Now, since our calculated value of F = 2.22 is less than the critical value of F = 3.24, we cannot reject the null hypothesis.

Example 2: A department store chain is considering building a new store at one of the three locations. An important factor in making such a decision is the household income in these areas. If the average income per household is similar then they can pick any one of these locations. A random survey of various households in each location is undertaken and their annual combined income is recorded. This data is tabulated as follows:

	Annual Household Income ($ 1, 000s)		
	Area (1)	**Area (2)**	**Area (3)**
	70	100	60
	72	110	65
	75	108	57
	80	112	84
	83	113	84
	-	120	70
	-	100	-
Total	380	763	420

Test if the average income per household in all these localities can be considered as the same at $\alpha = 0.01$.

Solution: If μ_i denotes the mean of the i^{th} area (i= 1, 2, 3) then the null hypothesis is:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

against H_1 : At least two means are not equal.

The null hypothesis can be tested by computing the F-ratio for the data given and then comparing it with the critical ratio of F from the table.

As before, let us first calculate the values of SSB and SSW.

Here: $\overline{X}_1 = \frac{380}{5} = 76 \quad \overline{X}_2 = \frac{763}{7} = 109 \qquad \overline{X}_3 = \frac{420}{6} = 70$

so that, $\overline{\overline{X}} = \frac{76 + 109 + 70}{3} = \frac{255}{3} = 85.$

Then, $SSB = 5(76 - 85)^2 + 7(109 - 85)^2 + 6(70 - 85)^2$

$$= 405 + 4032 + 1350 = 5787$$

$$SSW = \sum_{i=1}^{n_1} (X_{i1} - \overline{X}_1)^2 + \sum_{i=-}^{n_2} (X_{i2} - X_2)_2 + \sum_{i=1}^{n_{k1}} (X_{ik} - \overline{X}_k)^2$$

where $n_i = 5\ n_1 = 7 \qquad n_3 = 6 \quad k = 3$

$$SSW = \sum_{i=1}^{5} (X_{i1} - \overline{X}_1)^2 + \sum_{i=1}^{7} (X_{i2} - \overline{X}_2)^2 + \sum_{i=1}^{6} (X_{i3} - \overline{X}_3)^2$$

$$= (70 - 76)^2 + (72 - 76)^2 + (75 - 76)^2 + (80 - 76)^2 + (83 - 76)^2$$

$$+(100-109)^2+(110-109)^2+(108-109)^2+(112-109)^2$$

$$+(113-109)^2+(120-109)^2+(100-109)^2$$

$$+(60-70)^2+(65-70)^2+(57-70)^2+(84-70)^2+(84-70)^2+(70-70)^2$$

$$= 36 + 16 + 1 + 16 + 49 + 81 + 1 + 1 + 9 + 6 + 121 + 81$$

$$+100 + 25 + 169 + 196 + 196 + 0$$

$$= 118 + 310 + 686$$

Then, SSW = 118 + 310 + 686 = 1114.

Now, $$MSB = \frac{SSB}{(k-1)} = \frac{5787}{2} = 2893.5$$

$$MSW = \frac{SSW}{(N-k)} = \frac{1114}{15} = 74.26$$

Then, $$F = \frac{MSB}{MSW} = \frac{2893.5}{74.26} = 38.96$$

We can construct an ANOVA table for the problem solved above as follows:

ANOVA Table

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**F**
Treatment	SSB =5787	(k – 1) = 2	$MSB = \frac{SSB}{(k-1)}$ $\frac{5787}{2} = 2893.5$	$F = \frac{MSB}{MSW}$ = 38.96
Within (or error)	SSW =1114	(N – k) = 15	$MSW = \frac{SSW}{(N-k)}$ $\frac{1114}{15} = 74.26$	
Total	SST =6901			

The critical value of F from table for α = 0.01 and df 2 and 15 respectively is 6.36. Since our calculated value of F is higher than the table value of F, we cannot accept the null hypothesis.

CHECK YOUR PROGRESS 2

Note: i) Check your answers with those given at the end of the unit.

3) There are three sections of an introductory course in Statistics. Each section is being taught by a different professor. There are some complaints that at least one of the professors does not cover the

necessary material. To make sure that all the students receive the same level of material in a similar manner, the chairperson of the department has prepared a common test to be given to students of the three sections. A random sample of seven students is selected from each class and their test scores out of a total of 20 points are tabulated as follows;

Students	Section (1)	Section (2)	Section (3)
1	20	12	16
2	18	11	15
3	18	10	18
4	16	14	16
5	14	15	16
6	18	12	17
7	15	10	14
Total	119	84	112

Do you think that at 95% confidence level, there is significant difference in the average test scores of students taught by the different professors?

4) Able Insurance Company wants to test whether three of its salesmen, A, B, and C, in a territory make similar number of appointments with prospective customers during a given period of time. A record of previous four months showed the following results for the number of appointments made by each salesman for each month.

	Salesman		
Month	**(A)**	**(B)**	**(C)**
1	8	6	14
2	9	8	12
3	11	10	18
4	12	4	8
Totals	40	28	52

Do you think that at 95% confidence level, there is significant difference in the average number of appointments made by the salesmen per month?

12.5 TWO-WAY ANALYSIS OF VARIANCE (ANOVA)

In the previous Section we considered the case where only one predictor/ independent variable/explanatory was categorized at different levels. In this section, let us consider the case with two categorical predictors, each categorized at different levels and a continuous response variable. Then it is called two-way classification and the analysis is called Two-way ANOVA.

In such an ANOVA, generally we have an experiment in which we simultaneously study the effect of two factors in the same experiment. For each factor, there will be a number of classes/groups or levels. In the fixed effect model, there will be only fixed levels of the two factors. We shall first consider the case of one observation per cell. Let the factors be A and B and the respective levels be $A_1, A_2, \ldots, A_r$ and $B_1, B_2, \ldots, B_{qs}$. Let y_{ij} be the observation/response/dependent variable under the i^{th} level of factor A and j^{th}level of factor B. Further, let $\mu_{1A}, \mu_{2A}, \ldots, \mu_{rA}$be the means of levels $A_1, A_2, \ldots, A_r$ and $\mu_{1B}, \mu_{2B}, \ldots, \mu_{sB}$ be the means of levels $B_1, B_2, \ldots, B_s$in the population. The observations then can be represented in a table as follows:

TWO-WAY CLASSIFIED DATA

A/B	B_1	B_2	…	B_j	…	Bs	TOTAL	MEAN
A_1	y_{11}	y_{12}	…	y_{1j}	…	y_{1s}	$y_{1.}$	$\square_{1.}$
A_2	y_{21}	y_{22}	…	y_{2j}	…	y_{2s}	$y_{2.}$	$\square_{2.}$
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
A_i	y_{i1}	y_{i2}	…	y_{ij}	…	y_{is}	$y_{i.}$	$\square_{i.}$
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
A_r	Y_{r1}	Y_{r2}	…	Y_{rj}	…	Y_{rs}	$Y_{r.}$	$\square_{r.}$
Total	$y_{.1}$	$y_{.2}$	…	$y_{.j}$	…	$y_{.s}$	$y_{..} = G$	
Mean	$\square_{.1}$	$\square_{.2}$	…	$\square_{.j}$	…	$\square_{.s}$	-	$\square_{..}$

Mathematical Model

Here, the mathematical model may be written as

$y_{ij} = \mu_{ij} + e_{ij}$ wheree_{ij}'s are error terms.

12.5.1 Assumptions of Two-Way ANOVA

For the validity of F-test, the following assumptions should be satisfied:

(i) All the observations y_{ij} are independent;

(ii) Different effects (effects of levels of factor A, effects of levels of factor B and error effect) are additive in 'nature';

(iii) e_{ij}'s are independently and identically distributed normal variables with mean zero and variance σ_e^2 that is, $e_{ij} \approx iid\, N(0, \sigma_e^2)$; and

(iv) There is no interaction between different levels of factor A and B.

12.5.2 Test of significance in Two-way ANOVA

As we have discussed above that if we are interested in studying the simultaneous effect of two independent factors on the dependent variable, we use two-way ANOVA. For example, we may wish to study the simultaneous effects of five varieties of wheat (first criterion) and four different types of fertilizers (second criterion) on the yield (dependent variable) or test the stress level of employees in three different organizations in different regions, and so on. In such situations, we can also apply two separate one-way ANOVA for each treatment/factor. However, it is more advantageous to use two-way ANOVA because the variance can be reduced by introducing the second factor.

In two-way ANOVA, the total variation in the data is divided into three components: variation due to the first criterion (factor), variation due to the second criterion (factor) and variation due to error.

Now let us discuss the testing procedure of two-way ANOVA briefly mentioning the main steps and formulae as follows:

Step 1: We first formulate the null hypothesis (H_0) and alternative hypothesis (H_1). In two-way ANOVA, we can test two hypotheses simultaneously: one for different levels of factor A and the other for different levels of factor B. If factor A has r levels, we can set up the null and alternative hypotheses as follows:

$$H_{0A} : \alpha_1 = \alpha_2 = ... = \alpha_r$$

$$H_{1A} = \text{At least one } \alpha_i \neq \alpha_j (i \neq j = 1,2,..., r)$$

Similarly, if factor B has s levels, we can set up the null and alternative hypotheses as follows:

$$H_{0B} : \beta_1 = \beta_2 = ... = \beta_s$$

$$H_{1B} = \text{At least one } \beta_i \neq \beta_j (i \neq j = 1,2,..., s)$$

Step 2: We calculate the Correction factor (CF) and the raw sum of squares (RSS) using the formulae given below:

$$\text{Correction Factor (CF)} = \frac{G^2}{N} \quad \text{and} \quad \text{Raw Sum of Squares } (RSS) = \sum_{i=1}^{r}\sum_{j=1}^{s} y_{ij}^2$$

where G = sum of all observations, N = the total number of observations, and y_{ij} = observation of the i^{th} level of factor A and the j^{th} level of factor B.

Step 3: We calculate the total sum of squares (TSS), the sum of squares between rows or sum of squares due to factor A (SSA), the sum of squares between columns of sum of squares due to factor B (SSB) and sum of squares due to error (SSE) as follows:

$$\text{Total Sum of Squares(TSS)} = \text{RSS} - \text{CorrectionFactor(CF)}$$

$$\text{Sum of Squares due to Factor A (SSA)} = \frac{1}{q}\sum_{i=1}^{r} y_i^2 - \text{Correction Factor (CF)}$$

$$\text{Sum of Squares due to Factor B (SSB)} = \frac{1}{p}\sum_{i=1}^{s} y_i^2 - \text{Correction Factor (CF)}$$

Sum of Squares due to Error (SSE) = TSS – SSA – SSB

where y_i = the sum of the observations of the i^{th} level of factor A.

y_j = the sum of the observations of the j^{th} level of factor B.

Step 4: We determine the degrees of freedom (df) as

The degrees of freedom (df) for factor A = r – 1

The degrees of freedom (df) for factor B = s – 1

The degrees of freedom (df) for error = (r – 1) (s – 1)

Step 5: We obtain the various mean sums of squares as follows:

$$\text{Mean sum of squares due to factor A (MSSA)} = \frac{\text{SSA}}{r-1}$$

$$\text{Mean sum of squares due to factor B (MSSB)} = \frac{\text{SSB}}{s-1}$$

$$\text{Mean sum of squares due to error (MSSE)} = \frac{\text{SSE}}{(r-1)(s-1)}$$

Step 6: We calculate the value of the test statistics using the formulae given below:

$$F_A = \frac{\text{MSSA}}{\text{MSSE}}$$

$$F_B = \frac{\text{MSSB}}{\text{MSSE}}$$

ANOVA TABLE

Sources of Variation	**DF**	**SS**	**MSS**	**F-Test**
Variation Due to	(r-1)	SSA	$(\text{MSSA}) = \frac{\text{SSA}}{r-1}$	$F_A = \frac{\text{MSSA}}{\text{MSSE}}$

Factor A				
Variation Due to Factor B	(s-1)	SSB	$(MSSB) = \frac{SSB}{s-1}$	$F_B = \frac{MSSB}{MSSE}$
Variation Due to Error	(r-1)(s-1)	SSE	$(MSSE) = \frac{SSE}{(r-1)(s-1)}$	
Total	rs-1	TSS		

Step 7: We take decisions about the null hypotheses for factor A and factor B as explained below:

(i) Compare the calculated value of F_A with tabulated value of F_A at the respective df's. If calculated value is greater than the tabulated value then reject the hypothesis H_{0A}, otherwise it may be accepted.

(ii) Compare the calculated value of F_B with tabulated value of F_B at the respective df's. If calculated value is greater than the tabulated value then reject the hypothesis H_{0B}, otherwise it may be accepted.

Now after discussing the procedure of Two-way ANOVA test let us practice to solve some examples:

Example 3: Future group wishes to enter the frozen shrimp market. They contract a researcher to investigate various methods of groups of shrimp in large tanks. The researcher suspects that temperature and salinity are important factor influencing shrimp yield and conducts a two-way analysis of variance with their levels of temperature and salinity. That is each combination of yield for each (for identical gallon tanks) is measured.The recorded yields are given in the following chart:

Categorical variable Salinity (in pp)

Temperature	**700**	**1400**	**2100**	**Total**	**Mean**
60^0 F	3	5	4	12	4
70^0 F	11	10	12	33	11
80^0 F	16	21	17	54	18
Total	30	36	33	99	11

Compute the ANOVA table for the model.

Solution: Since in each all there is one observation. So we will use the model.

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$$

where, y_{ij} is the yield corresponding to i^{th} temperature and j^{th}salinity, μ is the general mean, α_i is the effect due to i^{th} temperature, β_j is the effect due to j^{th}salinity and $e_{ij} \sim$ i i d N $(0, \sigma^2)$. The hypotheses to be tested are

$H_{0A}: \alpha_1 = \alpha_2 = \alpha_3$against $H_{1A}: \alpha_1 \neq \alpha_2 \neq \alpha_3 \neq 0$

$H_{0B}: \beta_1 = \beta_2 = \beta_3$against $H_{1B}: \beta_1 \neq \beta_2 \neq \beta_3 \neq 0$.

The computations are as follows:

Grand Total (G) = 99

No. of observations(N) = 9

Correction Factor CF = $(99 \times 99) / 9 = 1089$

Raw Sum of Square (RSS) = 1401

Total Sum of Square (TSS) = RSS - CF= 1401-1089 = 312

Sum of Square due to Temperature (SST) = $(12)^2/3+(33)^2/3+(54)^2/3-1089 = 294$

Sum of Square due to Salinity (SSS)= $(30)^2/3+(36)^2/3+(33)^2/3-1089 = 6$

Sum of Square due to Error= TSS – SST – SSS= 312 -294 -6 = 12

ANOVA TABLE

Sources of Variation	**DF**	**SS**	**MSS**	**F-Test**
Due to Temperature	2	294	147	F_T=MSST/MSSE=147/3 =49
Due to Salinity	2	6	3	$F_S = 3/3 = 1$
Due to error	4	12	3	
Total	8	312		

Since tabulated value of $F_{2,4}$ at 5% level of significance is 10.649 which is less than the calculated F_T for testing the significant difference in shrimp yield due to differences in levels of temperature (49). So, H_{0A} isrejected. Hence, there aredifferences in shrimp yield due to temperature at 5% level of significance.

Since tabulated value of $F_{2,4}$ at 5% level of significance is 10.649 which is greater than the calculated F_S for testing the significant difference in shrimp yield due to difference in the level of salinity (1). So, H_{0B} is not rejected. Hence there is no any difference in shrimp yield due the salinity level of 5% level of significance.

CHECK YOUR PROGRESS 3

Note: i) Check your answers with those given at the end of the unit.

5) An experiment was conducted to determine the effect of different dates of planting and different methods of planting on the field of

sugar-cane. The data below show the fields of sugar cane for four different data and the methods of planting:

Dates of Planting

Method of Planting	October	November	February	March
I	7.10	3.69	4.70	1.90
II	10.29	4.79	4.50	2.64
III	8.30	3.58	4.90	1.80

Carry out an analysis of the above data.

6) A researcher wants to test four diets A, B, C, D on growth rate in mice. These animals are divided into 3 groups according to their weights. Heaviest 4, next 4 and lightest 4 are put in Block I, Block II and Block III respectively. Within each block, one of the diets is given at random to the animals. After 15 days increase in weight is noted and given in the following table:

Blocks	Treatments/Diets			
	A	B	C	D
I	12	8	6	5
II	15	12	9	6
III	14	10	8	5

Perform a two-way ANOVA to test whether the data indicate any significant difference between the four diets due to different blocks.

12.6 LET US SUM UP

In this unit, we have discussed:

1. Basic ideas and concepts related to ~~and~~the technique of Analysis of Variance as applied to test the equality of several population means simultaneously;
2. Basic underlying assumptions of ANOVA, terminologies and notations which are frequently used in Analysis of Variance;
3. Different applications of ANOVA;
4. Types of data which can be analyzed through ANOVA technique and the meaning of one -way and two-way classified data as well as the concept of one-way and two-way ANOVA;
5. Formation of null and alternative hypotheses to be tested and different steps of One-way ANOVAfor testing the hypothesis under consideration;
6. Null and alternative hypotheses to be tested and different steps of two-way ANOVA for testing the hypothesis under consideration;

12.7 KEY WORDS

Analysis of variance (ANOVA): An overall test of the null hypothesis for more than two population means.

Two-factor ANOVA: A more complex type of analysis of variance that tests whether differences exist among population means categorized by two factors or independent variables.

12.8 SUGGESTED FURTHER READING/ REFERENCES

Witte, R., & Witte, J. (2017). *Statistics*. Hoboken, NJ: John Wiley & Sons.

12.9 ANSWERS TO CHECK YOUR PROGRESS

1) Please refer to section 12.6excluding sub-sections 12.3.1 and 12.3.2.

2) Please refer to section 12.3.3.

3) The null hypothesis states that there are no differences among the mean scores of the three sections(denoted by μ_i, i = 1, 2, 3), so that we have:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_1 : \text{At least two means differ.}$$

a) As we know, the sum squares between the samples is given by:

$$SSB = \sum_{i=1}^{k} n_i (\overline{X}_i - \overline{\overline{X}})^2$$

In our case, we have 3 samples, therefore,

$$\overline{X}_1 = \frac{119}{7} = 17,\ \overline{X}_2 = \frac{84}{7} = 12,\ \overline{X}_3 = \frac{112}{7} = 16$$

And hence, $\overline{\overline{X}} = \frac{45}{3} = 15$

Then, $SSB = 7(17-15)^2 + 7(12-15)^2 + 7(16-15)^2 = 28 + 63 + 7 = 98.$

b) Sum of squares within (SSW) samples is calculated as follows:

$$\begin{aligned} SSW &= \sum_{I=1}^{n_1} (X_{i1} - \overline{X}_1)^2 + \sum_{i=1}^{n_2} (X_{i2} - \overline{X}_2)^2 + \sum_{i=1}^{n_k} (X_{ik} - \overline{X}_k)^2 \\ &= (20-17)^2 + (18-17)^2 + (18-17)^2 + (16-17)^2 \\ &+ (14-17)^2 + (18-17)^2 + (15-17)^2 \\ &+ (12-12)^2 + (11-12)^2 + (10-12)^2 + (14-12)^2 \\ &+ (15-12)^2 + (12-12)^2 + (10-12)^2 \\ &+ (16-16)^2 + (15-16)^2 + (18-16)^2 + (16-16)^2 \\ &+ (16-16)^2 + (17-16)^2 + (14-16)^2 \\ &= 58 \end{aligned}$$

Further, we have N = 21, k = 3 so that k-1= 2 and N-k= 18. Therefore, calculated

$$F = \frac{SSB/df}{SSW/df} = \frac{98/2}{58/18} = 15.21$$

The tabulated $F_{2,18}$ at 5% level of significance is 3.55. Hence, since $F_{tab}<F_{cal}$; H_0 is rejected implying that mean scores of three sections are not same.

4) In the usual notations,our null hypothesis is

H_0: $\mu_1 = \mu_2 = \mu_3$, and H_1 at least two means are unequal.

Here $\overline{X}_1 = \frac{40}{4} = 10 \qquad \overline{X}_2 = \frac{28}{4} = 7 \quad \overline{X}_3 = \frac{52}{4} = 13$

so that, $\overline{\overline{X}} = \frac{10+7+13}{3} = \frac{30}{3} = 10.$

Then, $SSB = 4(10-10)^2 + 4(7-10)^2 + 4(13-10)^2 = 0 + 36 + 36 = 72$

Degrees of freedom = df = (k – 1) = (3 – 1) = 2.

$$SSW = \sum_{i=1}^{4}(X_{i1} - \overline{X}_1)^2 + \sum_{i=1}^{4}(X_{i2} - \overline{X}_2)^2 + \sum_{i=1}^{4}(X_{i3} - \overline{X}_3)^2$$

$$= (8-10)^2 + (9-10)^2 + (10-10)^2 + (12-10)^2$$

$$+ (6-7)^2 + (8-7)^2 + (10-7)^2 + (4-7)^2$$

$$+ (14-13)^2 + (12-13)^2 + (18-13)^2 + (8-13)^2 = 82$$

Degrees of freedom = df = (N – k) = (12 – 3) = 9.

Then, the F-ratio is given as:

$$F = \frac{SSB/df}{SSW/df} = \frac{72/2}{82/9} = \frac{36}{9.1} = 3.95$$

The F-ratio from the table at 95% confidence level and degrees of freedom 2 and 9 respectively is given as 4.26.

Since our calculated value of F is less than the tabulated value of F, we cannot reject the null hypothesis.

5) Here we have H_{0A} : There is no differencein the average yield due to different methods of planting, that is,

H_{0A} : $\alpha_1 = \alpha_2 = \alpha_3$ against, H_{1A}: $\alpha_1 \neq \alpha_2 \neq \alpha_3$

H_{0B} : There is no difference in the average yield due to different dates of planting, that is,

H_{0B}: $\beta_1 = \beta_2 = \beta_3 = \beta_4$

H_{1B}: $\beta_1 \neq \beta_2 \neq \beta_3 \neq \beta_4$

G= $\sum\sum y_{ij}$ = Total of all observations

=7.10+3.69+4.70+1.90+10.29+4.79+4.58+2.64+8.30+ 3.58+4.90+1.80

=58.28

N = No. of observations = 12

Correction Factor = CF = $\frac{G^2}{N} = \frac{58.28 \times 58.28}{12} = 283.0465$

Raw Sum of Square (RSS) =

$$(7.10)^2 + (3.69)^2 + (4.70)^2 + (1.90)^2 + (10.29)^2 + (4.79)^2 + (4.58)^2 + (2.64)^2 + (8.30)^2 + (3.58)^2 + (4.90)^2 + (1.80)^2 = 355.5096$$

Total Sum of Square (TSS) = RSS – CF = 355.5096-283.0465= 72.4631

Sum of Square due to Dates of Planting (SSD)

$$= \frac{D_1^2}{3} + \frac{D_2^2}{3} + \frac{D_3^2}{3} + \frac{D_4^2}{3} - CF$$

$$= \frac{(25.69)^2}{3} + \frac{(12.06)^2}{3} + \frac{(14.18)^2}{3} + \frac{(6.35)^2}{3} - 283.0465$$

Sum of Square to Method of Planting (SSM) = $\frac{M_1^2}{4} + \frac{M_2^2}{4} + \frac{M_3^2}{4} - CF$

$$= \frac{(17.39)^2}{4} + \frac{(22.31)^2}{4} + \frac{(15.58)^2}{4} - 283.0465$$

= 286.3412-283.0465 = 3.2947

Sum of Square due to Error (SSE)= TSS-SSD-SSM

= 72.4631-3.2947-65.8917 = 3.2767

MSSD = $\frac{65.8917}{3} = 21.9639$, MSSM = $\frac{3.2947}{2} = 1.6473$,

MSSE = $\frac{3.2767}{6} = 0.5461$

For testing $H_{0A} : F_M$ is $\frac{MSSM}{MSSE} = \frac{1.6473}{0.5461} = 3.02$

For testing $H_{0B} : F_D$ is $\frac{MSSD}{MSSE} = \frac{21.9639}{0.5461} = 40.22$

The tabulated value of $F_{2,6}$ at 5% level of significance is 5.14 which is greater than the calculated value of F_M (3.02) so H_{0A} is accepted. So, we conclude that there is no significant difference among the different methods of planting.

The tabulated value of $F_{3,6}$ at 5% level of significance is 4.76 which is less than calculated value of F_D (40.22). So, we reject the null hypothesis H_{0B} . Hence there is a significant difference among the dates of planting.

6) Null Hypotheses are:

H_{01}: There is no significant difference between mean effects of diets.

H_{02}: There is no significant difference between mean effects of different blocks.

Against the alternative hypothesis

H_{11}: There is significant difference between mean effects of diets

H_{12}: There is significant difference between mean effects of different blocks.

Blocks	Treatments/Diets				
	A	B	C	D	Totals
I	12	8	6	5	31
II	15	12	9	6	42
III	14	10	8	5	37
Totals	$T_{1.}$=41	$T_{2.}$=30	$T_{3.}$=23	$T_{4.}$=16	110

Squares of observations

Blocks	**Treatments/Diets**				**Totals**
	A	B	C	D	
I	144	64	36	25	269
II	225	144	81	36	486
III	196	100	64	25	385
Totals	565	308	181	86	1140

Grand Total = G = $\sum\sum y_{ij}$ = 110

Correction Factor (CF) = $\frac{G^2}{N} = \frac{(110)^2}{12} = 1008.3333$

Raw Sum of Squares (RSS) = $\sum\sum y_{ij}^{\ 2} = 1140$

Total Sum Squares (TSS) = RSS - CF = 1140 - 1008.3333 = 131.6667

Sum of Squares due to Treatments/ Diets (SST)

$$= \frac{T_{1.}^2}{3} + \frac{T_{2.}^2}{3} + \frac{T_{3.}^2}{3} + \frac{T_{4.}^2}{3} - CF$$

$$= \frac{1}{3}[(41)^2 + (30)^2 + (23)^2 + (16)^2] - 1008.3333$$

$$= \frac{1}{3}[1681 + 900 + 529 + 256] - 1008.3333$$

$$= 1122 - 1008.3333 = 113.6667$$

Sum of squares due to block (SSB) = $\frac{1}{4}\left(T_{.1}^2 + T_{.2}^2 + T_{.3}^2\right) - CF$

$$= \frac{1}{4}[(31)^2 + (42)^2 + (37)^2] - 1008.3333$$

$$= 1023.5 - 1008.3333 = 15.1667$$

Sum of Squares due to Errors (SSE) = TSS – SST − SSB

$$= 131.6667 - 113.6667 - 15.1667 = 2.8333$$

Mean Sum of Squares due to Treatments (MSST)

$$= \frac{SST}{df} = \frac{113.6667}{3} = 37.8889$$

Mean Sum of Squares due to Blocks (MSSB)

$$= \frac{SSB}{df} = \frac{15.1667}{2} = 7.58335$$

Mean Sum of Squares due to Errors (MSSE)

$$= \frac{SSE}{df} = \frac{2.8333}{6} = 0.4722$$

ANOVA TABLE

Source of Variation	**SS**	**df**	**MSS**	**F**
Between Treatments/ Diets	113.6667	3	$\frac{113.6667}{3} = 37.8889$	$F_1 = \frac{37.8889}{0.4722} = 80.2391$
Between Blocks	15.1667	2	$\frac{15.1667}{2} = 7.58335$	$F_2 = \frac{7.58335}{0.4722} = 160.596$
Due to Errors	2.8333	6	$\frac{2.8333}{6} = 0.4722$	
Total	131.6667	11		

Tabulated F at 5% level of significance with (3, 6) degree of freedom is 4.76 & tabulated F at 5% level of significance with (2, 6) degree of freedom is 5.14

Conclusion: Since calculated F_1 > Tabulated F_1, so we reject H_{01} and conclude that there is significant difference between mean effect of diets.

Also calculated F_2 is greater than tabulated F_2, so we reject H_{02} and conclude that there is significant difference between mean effect of different blocks.

ignou
THE PEOPLE'S
UNIVERSITY