Dr. C.V. Raman University
Kargi Road, Kota, BILASPUR, (C. G.),
Ph. : +07753-253801, +07753-253872
E-mail : info@cvru.ac.in | Website : www.cvru.ac.in

Quantitative Method

Institute of **Open** and **Distance Education**

Faculty of Management

Quantitative Method

1MBA2

DR. C.V. RAMAN UNIVERSITY

Chhattisgarh, Bilaspur

A STATUTORY UNIVERSITY UNDER SECTION 2(F) OF THE UGC ACT

1MBA2
Quantitative Method

1MBA2
Quantitative Method

Credit- 2

Subject Expert Team

Dr. Vivek Bajpai, *Dr. C.V. Raman* University, Kota, Bilaspur, Chhattisgarh

Dr. Rajeev H. Peters, *Dr. C.V.* Raman University, Kota, Bilaspur, Chhattisgarh

Dr. Niket Shukla, *Dr. C.V. Raman* University, Kota, Bilaspur, Chhattisgarh

Dr. Satish Sahu, *Dr. C.V.* Raman University, Kota, Bilaspur, Chhattisgarh

Dr. Archana Agrawal, *Dr. C.V. Raman* University, Kota, Bilaspur, Chhattisgarh

Dr. Vikas Kumar Tiwari, *Dr. C.V. Raman* University, Kota, Bilaspur, Chhattisgarh

Course Editor:

- **Dr. Rahul Sharma, Associate Professor**
  School of Management, O.P. Jindal University, Raigarh, Chhattisgarh

Unit Written By:

1 **Dr. Niket Shukla**
Professor, Dr. C. V. Raman University, Bilaspur, Chhattisgarh

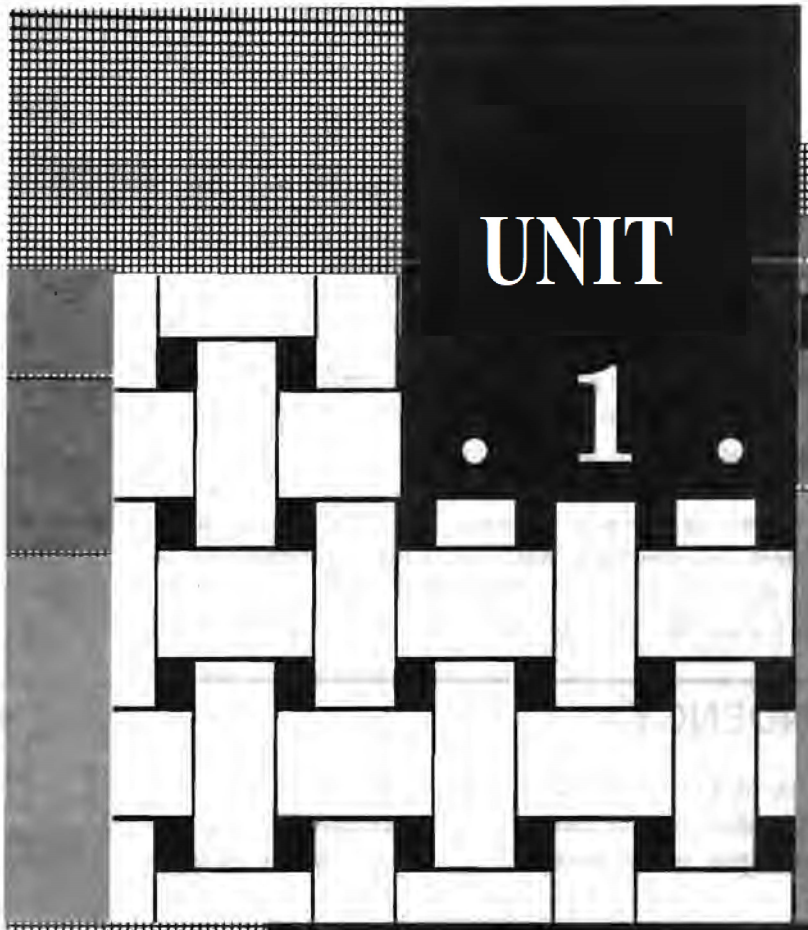
2 **Dr. Priyank Mishra**
Associate Professor, Dr. C. V. Raman University, Bilaspur, Chhattisgarh

3 **Dr. Anshul Shrivastava**
Assistant Professor, Dr. C. V. Raman University, Bilaspur, Chhattisgarh

Published by: Dr. C.V. Raman University Kargi Road, Kota, Bilaspur, (C. G.), Ph. +07753-253801,07753-253872 E-mail: inf o@cvru.ac.in Website: www.cvru.ac.in

CONTENTS

BLOCK – I



UNIT 1

Basic Quantitative Methods

CHAPTER OUTLINE

1.1 INTRODUCTION

Quantitative techniques may be defined as those techniques which provide the decision makes a systematic and powerful means of analysis, based on quantitative data. Quantitative techniques are those statistical and operation research techniques which help in the decision making process especially concerning business and industry. It is a scientific method employed for problem solving and decision making by the management. With the help of quantitative techniques, the decision maker is able to explore policies for attaining the predetermined objectives. In short, quantitative techniques are inevitable in decision-making process. Those techniques which provide the decision maker a systemic means of analysis based on the quantitative data in formulating policies for achieving pre-determined goals.

1.2 MEASURE OF CENTRAL TENDENCY

Measure of central tendency enables us to get an idea of entire data from a single value at which we consider the entire data is concentrated. This single value could be used to represent the entire population. Measure of central tendency also enables us to compare two or more sets of data, for example, average sales figures for two months.

Common Measures of Central Tendency

There are three common measures of central tendency:

1. Mean: The average value
2. Median: The middle value
3. Mode: Most occurring value

Here, we discuss the definitions, concepts and methods of manual calculation. Grouping of discrete data is not necessary for computer calculations. We can directly use the discrete data and get faster as well more accurate results than by grouping of the data. When only grouped data is available, we need to use formulae for grouped data.

1.3 MEAN

There are three types of mean:

1. Arithmetic Mean (AM)
2. Geometric Mean (GM)
3. Harmonic Mean (HM)

Arithmetic Mean

Arithmetic Mean is again of two types, 'Simple Arithmetic Mean' and 'Weighted Arithmetic Mean'.

Simple Arithmetic Mean

Simple Arithmetic Mean for Ungrouped Data (AM)

It is the value obtained by dividing the sum of all the values in data (called data points) by total number of such data points (observations). It is denoted by, $\bar{X}$ (X Bar) or μ depending on the data is a sample or population. Thus,

$$\mu = \frac{x_1 + x_2 + x_3 + ... + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

There is a short cut method for calculations based on a simple concept that, if a constant is subtracted or added to all data points, the Arithmetic Mean (AM) is reduced or increased by that amount. Thus,

$$\mu = A + \frac{\sum_{i=1}^{N} d_i}{N}$$

Where, A = Arbitrarily selected constant value (assumed mean). This value is selected such that it simplifies the values in calculations when deviation of each observation is used instead of the data values. A is selected close to the expected or guess value of mean. Calculations on deviation should be such that we should be able to do it orally.

d_i = Deviation of each observation from the assumed mean.

N = Number of observations.

Note that, when assumed mean 'A' is exactly equal to Arithmetic mean μ or $\bar{X}$, algebraic sum of all deviations is equal to zero. Thus, ***algebraic sum of deviations of all observations about Arithmetic Mean is zero.*** Or,

About Arithmetic Mean, $\sum_{i=1}^{N} d_i = 0$

Now we will solve one example just to demonstrate the method. Calculating arithmetic mean using MS excel is however, very simple.

Example 1

Find the arithmetic mean of 3, 6, 24, and 48.

Solution

Let the assumed mean A = 20

Sl. No.	x_i	Deviation $d_i = (x_i - A)$
1	3	−17
2	6	−14
3	24	4
4	48	28
N = 4	$\Sigma x_i = 81$	$\Sigma d_i = 1$

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{81}{4} = 20.25$$

Or, alternatively by short cut method,

$$\mu = A + \frac{\sum_{i=1}^{N} d_i}{N} = 20 + \frac{1}{4} = 20.25$$ This is same as direct method.

Note: If we take assumed mean as arithmetic mean 20.25,

Sl. No.	x_i	Deviation $d_i = (x_i - A)$
1	3	−17.25
2	6	−14.25
3	24	3.75
4	48	27.75
N = 4	$\Sigma x_i = 81$	$\Sigma d_i = 0$

Example 2

Find the arithmetic mean of 10, 12, 20, 15, 20, 12, 10, 15, 20 and 10

Solution

Arithmetic mean $\mu = \frac{x_1 + x_2 + x_3 + ... + x_N}{N} = \frac{10+12+20+15+20+12+10+15+20+10}{10} = 14.4$

OR

Frequency distribution of the data is,

Sl. No.	x_i	Frequency f_i	$x_i f_i$
1	10	3	30
2	12	2	24
3	15	2	30
4	20	3	60
K = 4		$\Sigma f_i = 10$	$\sum x_i f_i = 144$

Arithmetic Mean $\mu = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} x_i f_i}{N} = \frac{144}{10} = 14.4$

Simple Arithmetic Mean for Grouped Data

In case of grouped data, we consider class mark (Mid point of the class) as a data point (value of observation). In other words, we use mid-value of class for all the observations in that class (since we

don't know exact values of the observations, this is the best we can do keeping grouping errors to the minimum). Multiply the class marks by frequency of that class. Then the weighted average is calculated by dividing sum of these values of class marks with frequency as their weights, by total number of observation (sum of all frequencies). Thus, for grouped data,

$$\mu = \frac{\sum_{i=1}^{k} m_i f_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{k} m_i f_i}{N}$$

Where, m_i = class marks

f_i = Class frequency

$N = \Sigma f_i$ = Total number of observations

k = Number of classes

To make manual calculations easy, we may subtract or add a constant from all class marks (observations). In such case, as discussed earlier, 'Arithmetic Mean' is reduced or increased by that amount. Thus,

$$\mu = A + \frac{\sum_{i=1}^{k} f_i d_i}{\sum_{i=1}^{k} f_i}$$

Where, A = Assumed mean

d_i = $(m_i - A)$ = Deviation of class marks from the assumed mean

f_i = class frequency

N = Σf_i = Total number of observation

m_i = Class marks

k = Number of classes

This method is also called a 'Short Cut Method'. To make manual calculations further easy, we can use the principle, that if all the observations are divided or multiplied by a constant, the 'Arithmetic Mean' is divided or multiplied by that value. We select a convenient number usually the class width or size. Divide all deviations by that number. Then nse following formula to calculate 'Arithmetic Mean'. This method is called as 'Step Division Method'. The formula is:

$$\mu = A + \frac{\sum_{i=1}^{n} f_i d'_i}{\sum_{i=1}^{n} f_i} \times h$$

Where, A = Assumed mean.

$$d'_i = \frac{(m_i - A)}{h} = \frac{d_i}{h}$$

m_i = Class Marks.

h = Step size usually class interval.

N = Σf_i = Total number of observations.

We will now demonstrate the procedure with an example. You are recommended to solve one or two examples by manual calculations (use calculator if necessary). Then onwards you could use MS-Excel, which saves lot of drudgery of calculations and time. If you are using a computer, it is not necessary to group the 'ungrouped data' (discrete data). This will make calculations in MS Excel easier and chances of accurate solution increase. However, if the data is already grouped, we need to write a function in relevant cell and then drag copy.

Example 3

From the following data, compute Arithmetic Mean by direct method, short cut methods and step division method.

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No of students	5	10	25	30	20	10

Solution

Let the Assumed Mean be A = 35 and Step size h = 10

Calculation Table

Marks	Class Mark (m_i)	No. of Students (f_i)	$m_i \cdot f_i$	Deviation $d_i = m_i - A$	$f_i \cdot d_i$	Step Deviation $d_i' = (m_i - A)/h$	$f_i \cdot d_i'$
0-10	5	5	25	-30	-150	-3	-15
10-20	15	10	150	-20	-200	-2	-20
20-30	25	25	625	-10	-250	-1	-25
30-40	35	30	1050	0	0	0	0
40-50	45	20	900	10	200	1	20
50-60	55	10	550	20	200	2	20
Σ		**100**	**3300**		**- 200**		**- 20**

1. **Direct Method**

$$\mu = \frac{\sum_{i=1}^{6} m_i f_i}{\sum_{i=1}^{6} f_i} = \frac{3300}{100} = 33$$

2. **Shortcut Method**

$$\mu = A + \frac{\sum_{i=1}^{6} f_i d_i}{\sum_{i=1}^{6} f_i} = 35 + \frac{-200}{100} = 35 - 2 = 33$$

3. Step Division Method

$$\mu = A + \frac{\sum_{i=1}^{n} f_i d'_i}{\sum_{i=1}^{n} f_i} \times h = 35 + \frac{-20}{100} \times 10 = 33$$

Note: The answer is same irrespective the method used.

Arithmetic Mean of Combined Data

Arithmetic Mean is used very often in business for calculating average sales, average cost, average earnings, etc. If there are two related data groups and their arithmetic means are known, we can calculate arithmetic mean of the combined data without referring to individual data points. If the first group of N_1 items has arithmetic mean of μ_1, the second group of N_2 items has arithmetic mean of μ_2, and so on. We can find the arithmetic mean of combined data as,

$$\mu = \frac{N_1 \times \mu_1 + N_2 \times \mu_2 + \ldots\ldots + N_n \times \mu_n}{N_1 + N_2 + \ldots\ldots + N_n}$$

Example 4

The weekly average salaries paid to all employees in a certain company was Rs. 600. The mean salaries paid to male and female employees were Rs. 620 and Rs. 520 respectively. Obtain the percentage of male and female employees in the company.

Solution

Arithmetic mean of combined data is,

$$\mu = \frac{N_1 \times \mu_1 + N_2 \times \mu_2 + \ldots\ldots + N_n \times \mu_n}{N_1 + N_2 + \ldots\ldots + N_n}$$

In this problem N_1 = number of male employees, N_2 = number of female employees, mean salary of male employees $\mu_1 = 620$, mean salary of female employees $\mu_2 = 520$ and combined mean $\mu = 600$. Therefore,

$$\mu = \frac{N_1 \times \mu_1 + N_2 \times \mu_2}{N_1 + N_2} \Rightarrow 600 = \frac{620 \times N_1 + 520 \times N_2}{N_1 + N_2} \Rightarrow 20 \times N_1 = 80 \times N_2$$

$\therefore \; N_1 : N_2 = 4 : 1$

Thus, percentage of male and female employees in the company is 80% and 20% respectively.

Weighted Arithmetic Mean

There are cases where relative importance of the different items is not the same. In such a case, we need to compute the weighted arithmetic mean. The procedure is similar to the grouped data calculations studied earlier, when we consider frequency as a weight associated with the class-mark. Now suppose

the data values are $x_1, x_2, x_3, \ldots, x_n$ and associated weights are $W_1, W_2, W_3 \ldots W_n$, then the weighted arithmetic mean is:

Direct Method

$$\mu_w = \mu_w = \frac{W_1 \times x_1 + W_2 \times x_2 + \ldots\ldots + W_n \times x_n}{W_1 + W_2 + \ldots\ldots + W_n} = \frac{\sum W_i \times x_i}{\sum W_i}$$

Shortcut Method

$$\mu_w = A_w + \frac{\sum W_i \times d_i}{\sum W_i}$$

Where A_w = Assumed weighted mean.

d_i = $(A_w - x_i)$ Deviation of observations from assumed mean.

Note: For calculations with MS Excel, follow the steps given for grouped data except in place of class-mark, enter the observation values and in place of frequencies use weights.

Utility of Weighted Mean

Some of the common applications where weighted mean is extensively used are:

1. Construction of index numbers, for example, consumer Price Index, BSE sensex, etc., where different weights are associated for different items or shares.
2. Comparison of results of the two companies when their sizes are different.
3. Computation of standardized death and birth rates.

Example 5

The management of hotel has employed 2 managers, 5 cooks and 8 waiters. The monthly salaries of the managers, the cooks and waiters are Rs. 3000, Rs. 1200 and Rs. 1000 respectively. Find the mean salary of the employees. (Note: Although these salaries must be 10 to 15 year old, we will take it only to learn the principle.)

Solution

Here we need to calculate waited average of salary with salaries as weights.

$$\mu_w = \frac{W_1 \times x_1 + W_2 \times x_2 + \ldots\ldots + W_n \times x_n}{W_1 + W_2 + \ldots\ldots + W_n} = \frac{2 \times 3000 + 5 \times 1200 + 8 \times 1000}{2 + 5 + 8}$$

$= 1333.33$ Rs.

Geometric Mean (GM)

It is defined as n^{th} root of the product of 'N' values of data. If $x_1, x_2 \ldots\ldots x_n$ are values of data, then Geometric Mean,

$$GM = \sqrt[n]{x_1 \times x_2 \times \ldots\ldots \times x_n}$$

Now taking log on both sides,

$$\log(GM) = \frac{\sum_{i=1}^{n} \log(x_i)}{n}$$

If different values are not of equal importance and are assigned different weights say $w_1, w_2 \ldots w_n$ then weighted Geometric Mean is given by

$$GM_w = \sqrt[n]{x_1^{w_1} \times x_2^{w_2} \times \ldots\ldots \times x_n^{w_n}}$$

$$\text{Or,} \quad \log(GM_w) = \frac{\sum_{i=1}^{n} w_i \log(x_i)}{\sum_{i=1}^{n} w_i}$$

Geometric Mean is useful to find the average percentage increase in sales, production, population, etc. It is the most representative average in the construction of index numbers.

Geometric mean is antilog of the mean of logarithms of observations. It is useful for graphical representations when the range of the data is very large.

Example 6

A person takes home loan with floating interest, on reducing balance of 10 year term. The interest rates as changed from year to year in percent are 5.5, 6.25, 7.5, 6.75, 8.25, 9.5, 10.5, 9, 8.25 and 7.5. Find the average interest rate? Was it beneficial for him to take fixed interest rate on reducible balance at 7.5% per annum?

Solution

Average interest rate can be found out using G.M. as follows. First, we find the index by dividing percentage rate by 100 and then adding 1. Then we take G.M. of this index as average index. From this, we can find out the average interest rate.

$$\text{Average index (G.M.)} = \sqrt[10]{1.055 \times 1.0625 \times 1.075 \times 1.0675 \times 1.0825 \times 1.095 \times 1.105 \times 1.09 \times 1.0825 \times 1.075}$$

$$= \sqrt[10]{2.137} = 1.0789$$

Thus, Average Interest Rate = 7.89%

Hence it was beneficial for him to take fixed interest rate on reducible balance at 7.5% per annum.

Harmonic Mean (HM)

It is defined as the reciprocal of the arithmetic mean of the reciprocals of the individual observations. Thus, Harmonic Mean is,

$$HM = \frac{n}{\left(\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}\right)} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Example 7

A relay team has four members who have to drive four laps between two fixed points. Average speeds that the members can achieve in Km/hr are 280, 360, 380 and 310. Find average speed of the team to complete the event.

Solution

The average speed can be calculated as Harmonic Mean (HM). Thus, average speed of the team is,

$$HM = \frac{n}{\left(\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}\right)} = \frac{4}{\left(\frac{1}{280} + \frac{1}{360} + \frac{1}{380} + \frac{1}{310}\right)} = 327.69 \text{ Km/hr}$$

Weighted Harmonic Mean

If weight is attached with each observation then the Weighted Harmonic Mean is:

$$HM = \frac{w_1 + w_2 + ... + w_n}{\left(\frac{w_1}{x_1} + \frac{w_2}{x_2} + ... + \frac{w_n}{x_n}\right)} = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{x_i}}$$

Weighted Harmonic Mean is useful in computing the average rate of increase in profits, average speed of journey, average price of articles sold, etc. For example, airplane travels distances w_1, w_2, w_3 $\ldots w_n$ with speeds x_1, x_2, $x_3 \ldots x_n$ km\hr respectively, then the average speed is equal to Weighted Harmonic Mean of speeds, with weights as the distances w_1, w_2, $w_3 \ldots w_n$.

Example 8

An aircraft travels 200 km upto border at speed 700 km/hr (economical), then 250 km upto the target in enemy territory at speed 950 km/hr, then after dropping the bombs travels at runaway speed of 1700 km/hr upto our nearest border at 150 km and then at the speed of 800 km/hr to the base at distance of 300 km. Find the average speed of the sortie. Also find the mission time.

Solution

For the average speed, we need to find the weighted Harmonic Mean. Thus, the average sortie speed is,

$$HM = \frac{w_1 + w_2 + ... + w_n}{\left(\frac{w_1}{x_1} + \frac{w_2}{x_2} + ... + \frac{w_n}{x_n}\right)} = \frac{200 + 250 + 150 + 300}{\left(\frac{200}{700} + \frac{250}{950} + \frac{150}{1700} + \frac{300}{800}\right)} = 889.23 \text{ km/hr}$$

Mission time $1.012 \simeq$ hr approx.

Relationship Among Averages

Arithmetic Mean, Geometric Mean and Harmonic Mean are related through the following relationships.

1. $AM \times HM = (GM)^2$
2. $AM \geq GM \geq HM$

1.4 MEDIAN (M_d)

Median is the value, which divides the distribution of data, arranged in ascending or descending order, into two equal parts. Thus, the 'Median' is a value of the middle observation.

Median for Ungrouped Data

When the series is arranged in order of size or magnitude, and if total number of observations are odd,

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

If the number of observations is even, then the median is the arithmetic mean of two middle observations.

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

Example 9

Students of a class were divided in two groups and undergone tutorial training by different faculty members. There scores in final examination are:

Group A: 80, 70, 50, 20, 30, 90, 10, 40, 60

Group B: 80, 70, 50, 20, 30, 90, 10, 40, 60, 100

Which group showed better performance based on Median?

Solution

First we arrange the scores in ascending order.

Group A: 10, 20, 30, 40, 50, 60, 70, 80, 90

Number of observations is 9 (odd). Therefore,

$$Median\ M_d = \left(\frac{N+1}{2}\right)^{th} = \frac{9+1}{2} = 5^{th} \text{ observation} = 50$$

Group B: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100

Number of observations is 10 (even). Therefore,

$$Median\ M_d = \frac{\left(\frac{N}{2}\right)^{th} observation + \left(\frac{N}{2}+1\right)^{th} observation}{2} = \frac{50+60}{2} = 55$$

Thus, group B has better performance on median.

Median for Grouped Data

In case of grouped data we first find the value $\frac{N}{2}$. Then from the cumulative frequency we find the class in which the $\left(\frac{N}{2}\right)^{th}$ item falls. Such a class is called as Median Class. Then the median is calculated by formula:

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

Where, L = Lower limit of Median class.

N = Total Frequency.

pcf = Preceding cumulative frequency to the median class.

f = Frequency of median class.

h = Class interval of median class.

Let us understand the logic of the formula. Median is value of $\left(\frac{N}{2}\right)^{th}$ observation. But this observation falls in the median class whose lower limit is L. Cumulative frequency of class preceding to the 'median class' is pcf. Thus, the median observation is $\left(\frac{N}{2} - pcf\right)^{th}$ observation in the median class (counted from the lower limit of the median class). Now, if we consider that all f observations in the median class are evenly spaced from lower limit L to upper limit $L+h$, the value of the median can be found out by using ratio proportion.

We will solve one problem on grouped data for demonstrating the procedure. Once you understand the concept you are advised to use MS Excel for finding the median.

Example 10

Calculate the median for the following data.

Age	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60
No. of Workers	14	28	33	30	20	15	13	7

Solution

Age	Frequency f	Cumulative Frequency cf
20-25	14	14
25-30	28	42
30-35	33	75
35-40	30	105
40-45	20	125
45-50	15	140
50-55	13	153
55-60	7	160

Now, $N = 160$

Or, $\frac{N}{2} = 80$

80^{th} item lies in class 35-40.

Hence, $pcf = 75$, $f = 30$, $h = 5$ and $L = 35$

Therefore, the Median is,

$$M_d = L + \frac{\frac{N}{2} - pcf}{f} \times h = 35 + \frac{\frac{160}{2} - 75}{30} \times 5$$

$$= 35.83$$

1.5 MODE

The Mode of a data set is the value that occurs most frequently. There are many situations in which arithmetic mean and median fail to reveal the true characteristics of a data (most representative figure), for example, most common size of shoes, most common size of garments etc. In such cases, mode is the best-suited measure of the central tendency. There could be multiple model values, which occur with equal frequency. In some cases, the mode may be absent. For a grouped data, model class is defined as the class with the maximum frequency. Then the mode is calculated as:

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times h$$

Where,

L = Lower limit of modal class.

Δ_1 = Difference between frequency of the modal class and preceding class.

Δ_2 = Difference between frequency of the modal class and succeeding class.

h = Size of the modal class.

Example 11

In a computerized entrance test, 20 candidates appear on a particular day. Their scores are: 9, 6, 12, 10, 13, 15, 16, 14, 14, 16, 17, 16, 24, 21, 22, 18, 19, 18, 20, 17. Find the mode of the data.

Solution

Using manual calculations

Now the value 16 occurs 3 times which is maximum for any observation. Therefore,

Mode = 16

Example 12

In a computerized entrance test, 20 candidates appear on a particular day. Their scores are: 9, 6, 12, 10, 13, 15, 14, 14, 16, 17, 16, 24, 21, 22, 18, 19, 18, 20, 17, 8. Find the mode of the data.

Solution

Using manual calculations

Now the values 14, 16, 17 and 18 occur 2 times which is maximum for any observation. Therefore,

Modes are 14, 16, 17 and 18 (this is a multimodal distribution).

Example 13

In a computerized entrance test 20 candidates appear on a particular day. Their scores are: 9, 6, 12, 10, 13, 15, 14, 16, 24, 21, 22, 19, 18, 20, 17, 8, 11, 26, 2, 5. Find the mode of the data.

Solution

Now there is no value that occurs more than 1 time. Therefore, the data has no Mode.

Relationship among Mean, Median and Mode

A distribution in which the mean, the median, and the mode coincide is known as **symmetrical (bell shaped) distribution.** Normal distribution is one such a symmetric distribution, which is very commonly used.

If the distribution is skewed, the mean, the median and the mode are not equal. In a moderately skewed distribution distance between the mean and the median is approximately one third of the distance between the mean and the mode. This can be expressed as:

Mean – Median = (Mean – Mode) / 3

Mode = 3 * Median – 2 * Mean

Thus, if we know values of two central tendencies, the third value can be approximately determined in any moderately skewed distribution. In any skewed distribution, the median lies between the mean and mode.

In case of right-skewed (positive-skewed) distribution which has a long right tail,

Mode <Median < Mean.

In case of left-skewed (negative-skewed) distribution which has along left tail,

Mean < Median < Mode

1.6 CORRELATION

Correlation is a degree of linear association between two random variables. In these two variables, we do not differentiate them as dependent and independent variables. It may be the case that one is the cause and other is an effect i.e. independent and dependent variables respectively. On the other hand, both may be dependent variables on a third variable. In some cases there may not be any cause-effect relationship at all. Therefore, if we do not consider and study the underlying economic or physical relationship, correlation may sometimes give absurd results. For example, take a case of global average temperature and Indian population. Both are increasing over past 50 years but obviously not related.

Correlation is an analysis of the degree to which two or more variables fluctuate with reference to each other. Correlation is expressed by a coefficient ranging between –1 and +1. Positive (+ve) sign indicates movement of the variables in the same direction. For example, Variation of the fertilizers used

on a farm and yield, observes a positive relationship within technological limits. Whereas negative (–ve) coefficient indicates movement of the variables in the opposite directions, i.e. when one variable decreases, other increases. For example, Variation of price and demand of a commodity have inverse relationship. Absence of correlation is indicated if the coefficient is close to zero. Value of the coefficient close to +1 denotes a very strong linear relationship.

The study of correlation helps managers in following ways:

1. To identify relationship of various factors and decision variables.
2. To estimate value of one variable for a given value of other if both are correlated. For example, estimating sales for a given advertising and promotion expenditure.
3. To understand economic behaviour and market forces.
4. To reduce uncertainty in decision-making to a large extent.

In business, correlation analysis often helps manager to take decisions by estimating the effects of changing the values of the decision variables like promotion, advertising, price, production processes, on the objective parameters like costs, sales, market share, consumer satisfaction, competitive price. The decision becomes more objective by removing subjectivity to certain extent. However, it must be understood that the correlation analysis only tells us about the two or more variables in a data fluctuate together or not. It does not necessarily be due cause and effect relationship. To know if the fluctuations in one of the variables indeed affect other or not, one has to be established with logical understanding of the business environment.

Some of the correlations could be completely nonsense relations like increase in jobs in I.T. and reduction production of wheat over past 3 years in India, or share market bull run of 2004 to 2007 and increase in suicides by farmers in India. There are many reasons to get such spurious correlations. Hence before we use correlation analysis we must check few factors responsible for the apparent relationship. Firstly, the fluctuation may be a chance coincidence. In this case we could look at the data over different periods and also study if one factor affects the other through third factor that we have not considered. Secondly, even when correlation exists the logical analysis may tell us that one variable is independent and other dependent on it. For example, surface temperature of the Pacific Ocean (Al Niño) affects monsoons in India but monsoons do not affect temperatures of the Pacific Ocean. Thirdly, in some cases both variables under study may be fluctuating together due to a variation in the third variables. Thus both variables under correlation analysis may be dependent variables and hence not mutually correlated. In such a case, manager can not vary one of them and expect other variable to vary. For example, correlation in increase in share prices and stronger rupee against dollar may be due to increase in Foreign Direct Investment (FDI). In this case expecting to control falling share prices through selling dollars by the Reserve Bank is incorrect. To control these two variables we need to control FDI. Further, if the falling share prices are due to market sentiments or overheated market, controlling FDI may not help. Thus, the manager needs to analyze the problem in business environment before he/she can apply the correlation analysis in decision-making.

The correlation can be studied as positive and negative, simple and multiple, partial and total, linear and non linear. Further the method to study the correlation is plotting graphs on x-y axis or by algebraic calculation of coefficient of correlation. Graphs are usually scatter diagrams or line diagrams. The correlation coefficients have been defined in different ways, of these Karl Pearson's correlation coefficient; Spearman's Rank correlation coefficient and coefficient of determination are more popular. Drawing scatter diagram was discussed in chapter 3. Here we will discuss coefficients of correlation.

In managerial decision-making, it is a good practice to draw the scatter diagram first, then study the logical relationship to identify the type of correlation and the cause effect relation. Only then manager should calculate the coefficient of correlation for further mathematical analysis. A computer packages, or MS Excel can be used for plotting the scatter diagram as well as finding correlation coefficient. Types of correlation that need to be differentiated before using the correlation coefficient for managerial decision-making are given below.

Positive or Negative Correlation

In positive correlation, both factors increase or decrease together. When we say a perfect correlation, the scatter diagram will show a linear (straight line) plot with all points falling on straight line. If we take appropriate scale, the straight line inclination can be adjusted to 45°, although it is not necessary as long as inclination is not 0° or 90° where there is no correlation at all because value of one variable changes without any change in the value of other variable. In case of negative correlation when one variable increases the other decrease and visa versa. If the scatter diagram shows the points distributed closely around an imaginary line, we say it is high degree of correlation. On the other hand, if we can hardly see any unique imaginary line around which the observations are scattered, we say correlation does not exist. Even in case of imaginary line being parallel to one of the axes we say no correlation exists between the variables. If the imaginary line is a straight line we say the correlation is linear.

Simple or Multiple Correlation

In simple correlation the variation is between only two variables under study and the variation is hardly influenced by any external factor. In other words, if one of the variables remains same, there won't be any change in other variable. For example, variation in sales against price change in case of a price sensitive product under stable market conditions shows a negative correlation. In multiple correlation, more than two variables affect one another. In such a case, we need to study correlation between all the pairs that are affecting each other and study extent to which they have the influence.

Partial or Total Correlation

In case of multiple correlation analysis there are two approaches to study the correlation. In case of partial correlation, we study variation of two variables and excluding the effects of other variables by keeping them under controlled condition. In case of 'total correlation' study we allow all relevant variables to vary with respect to each other and find the combined effect. With few variables, it is feasible to study 'total correlation'. As number of variables increase, it becomes impractical to study the 'total correlation'.

Liner and Nonlinear Correlation

The manager must be careful in analyzing the correlation using coefficients because most of the coefficients are based on assumption of linearity. Hence plotting a scatter diagram is good practice. Scatter diagram not only tell us about linearity or nonlinearity but also whether the data is cyclic. When values of two variables have a constant rate of change it is linear correlation. In such a case, the differential (derivative) of relationship is constant with the graph of the data being a straight line. In case on nonlinear correlation the rate of variation changes as values increase or decrease. The non-linear relationship could be approximated to a polynomial (parabolic, cubic etc.), exponential sinusoidal, etc. In such cases using the correlation coefficients based on linear assumption will be misleading unless used over a very short

data range. Using computers, we could analyze a non-linear correlation to a certain extent, with some simplified assumption.

Practical Application of Correlation

The primary purpose of correlation is to establish an association between any two random variables. The presence of association does not imply causation, but the existence of causation certainly implies association. Statistical evidence can only establish the presence or absence of association between variables. Whether causation exists or not depends merely on reasoning. However, one must be on the guard against spurious or nonsense correlation that may be observed between totally unrelated variables before regression analysis.

Correlation is also used in factor analysis wherein attempts are made to resolve a large set of measured variables in terms of relatively few categories, known as factors. The results could be useful in following three ways:

1. To reveal the underlying or latent factors that determines the relationship between the observed data.
2. To make evident relationship between data that had been obscured before such analysis.
3. To provide a classification scheme when data scored on various rating scales have to be grouped together.

Another major application of correlation is in forecasting with the help of time series models. In past data one has to identify the trend, seasonality and random pattern in the data before an appropriate forecasting model can be built.

1.7 LINEAR SIMPLE CORRELATION

Simple linear correlation is a statistical tool applied in many business situations to find the degree to which two variables vary linearly to one another. Although in many situations even if there are more than two variables involved, two of them may be dominant. In such a case, correlation analysis between these two variables help us to measure the degree of association between these two variables. For example, demand of a particular product depends on number of factors. However, association of demand with price may be dominant. Correlation analysis may also be necessary to eliminate a variable which shows low or hardly any correlation with the variable of our interest. In statistics, there are number of measures to describe degree of association between variables. These are Karl Pearson's Correlation Coefficient, Spearman's rank correlation coefficient, coefficient of determination, Yule's coefficient of association, coefficient of colligation, etc.

The Correlation Coefficient

The correlation coefficient measures the degree of association between two variables X and Y. Karl Pearson's formula for correlation coefficient is given as,

$$r = \frac{\frac{1}{n}\sum (X - \bar{X})(Y - \bar{Y})}{\sigma_X \sigma_Y} \quad \ldots (1)$$

Where r is the 'Correlation Coefficient' or 'Product Moment Correlation Coefficient' between X and Y. σ_X and σ_Y are the standard deviations of X and Y respectively. 'n' is the number of the pairs of variables X and Y in the given data. The expression $\frac{1}{n}\sum(X-\bar{X})(Y-\bar{Y})$ is known as a covariance between the variables X and Y. It is denoted as *Cov(x, y)*. The Correlation Coefficient r is a dimensionless number whose value lies between +1 and –1. Positive values of r indicate positive (or direct) correlation between the two variables X and Y i.e. both X and Y increase or decrease together. Negative values of r indicate negative (or inverse) correlation, thereby meaning that an increase in one variable X or Y results in a decrease in the value of the other variable. A zero correlation means that there is no association between the two variables.

The formula can be modified as,

$$r = \frac{\frac{1}{n}\sum(X-\bar{X})(Y-\bar{Y})}{\sigma_X\sigma_Y} = \frac{\frac{1}{n}\sum(XY-X\bar{Y}-\bar{X}Y+\bar{X}\bar{Y})}{\sigma_X\sigma_Y}$$

$$= \frac{\frac{\sum XY}{n} - \frac{\sum X}{n}\times\frac{\sum Y}{n}}{\sqrt{\frac{\sum X^2}{n}-\left(\frac{\sum X}{n}\right)^2}\sqrt{\frac{\sum Y^2}{n}-\left(\frac{\sum Y}{n}\right)^2}} \quad \ldots (2)$$

$$\frac{E[XY]-E[X]E[Y]}{\sqrt{E[X^2]-\left(E[X]\right)^2}\sqrt{E[Y^2]-\left(E[Y]\right)^2}} \quad \ldots (3)$$

Equations (2) and (3) are alternate forms of equation (1). These have advantage that we don't have to subtract each value from the mean. In any case while using MS Excel, we don't have to use any of these formulae.

Example 14

The data of advertisement expenditure (X) and sales (Y) of a company for past 10 year period is given below. Determine the correlation coefficient between these variables and comment on the correlation.

X	50	50	50	40	30	20	20	15	10	5
Y	700	650	600	500	450	400	300	250	210	200

Solution:

$$\bar{X} = \frac{\sum X}{n} = \frac{290}{10} = 29, \quad \bar{Y} = \frac{\sum Y}{n} = \frac{4260}{10} = 426$$

S. No.	X	Y	$x = (X - \bar{X})$	$y = (Y - \bar{Y})$	x^2	y^2	xy
1	50	700	21	274	441	75076	5754
2	50	650	21	224	441	50176	4704
3	50	600	21	174	441	30276	3654
4	40	500	11	74	121	5476	814
5	30	450	1	24	1	576	24
6	20	400	-9	-26	81	676	234
7	20	300	-9	-126	81	15876	1134
8	15	250	-14	-176	196	30976	2464
9	10	210	-19	-216	361	46656	4104
10	5	200	-24	-226	576	51076	5424
Total $\sum$	290	4260	0	0	2740	306840	28310

Now, $$r = \frac{\frac{1}{n}\sum(X - \bar{X})(Y - \bar{Y})}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\sum xy}{\sqrt{\frac{\sum x^2}{n}}\sqrt{\frac{\sum y^2}{n}}} = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

$$r = \frac{28310}{\sqrt{2740 \times 306840}} = 0.976$$

This value of Karl Pearson's coefficient $r = 0.976$ indicates a high degree of positive association between the variables X and Y.

Effect of shifting origin and change of scale on correlation coefficient

Value of $\bar{X}$ and $\bar{Y}$ may not be integers. In such a case, the calculations become tedious. We can expand the formula as,

$$r = \frac{\frac{1}{n}\sum(X - \bar{X})(Y - \bar{Y})}{\sigma_X \sigma_Y} = \frac{\sum XY - \frac{1}{n}\sum X \sum Y}{\sqrt{\sum X^2 - \frac{1}{n}(\sum X)^2}\sqrt{\sum Y^2 - \frac{1}{n}(\sum Y)^2}}$$

Further simplification in computations can be adopted by calculating the deviation of the observation from an assumed mean rather than the actual mean, and also scaling these deviations conveniently.

Here we use the property that correlation coefficient does not change with shifting of origin i.e. by adding or subtracting any constant from the two variables (X, Y) correlation coefficient remains same. It also remains unchanged if we change the scales by dividing or multiplying the variables by a constant. Let X and Y be the two variable with values $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$. Let us define another two variables obtained by transformation as,

$$U = \frac{X - a}{g} \text{ and } V = \frac{Y - b}{h}$$

Where a, b, g and h are constants.

In this case, we have defined variables U and V through shift of origin from (0, 0) to (a, b) and change the X and Y scale by factors 'g' and 'h' respectively. Thus for every observation pair (x_i, y_i) there is a corresponding pair (u_i, v_i) such that,

$$u_i = \frac{x_i - a}{g} \text{ and } v_i = \frac{y_i - b}{h}$$

$$\text{Now, } \bar{X} = \frac{\Sigma x_i}{n} = \frac{\Sigma(g \times ui + a)}{n} = \frac{g \times \Sigma u_i + n \times a}{n} = g\bar{U} + a$$

Similarly,

$$\bar{Y} = h\bar{V} + b$$

$$\text{Now, } x_i - \bar{X} = (g \times u_i + a) - (g\bar{U} + a) = g(u_i - \bar{U})$$

$$\text{And } y_i - \bar{Y} = h(v_i - \bar{V})$$

$$\text{Hence, } \sigma_X^2 = \frac{\Sigma(x_i - \bar{X})^2}{n} = g^2 \times \frac{\Sigma(u_i - \bar{U})^2}{n} = g^2\sigma_U^2$$

$$\text{And } \sigma_Y^2 = h^2\sigma_V^2$$

$$\text{Now, } r_{XY} = \frac{\frac{1}{n}\Sigma(x_i - \bar{X})(y_i - \bar{Y})}{\sigma_X\sigma_Y} = \frac{\Sigma g \times (u_i - \bar{U}) \times h \times (v_i - \bar{V})}{n \times (g \times \sigma_U)(h \times \sigma_V)}$$

$$= r_{UV}$$

This result is very useful for manual calculations. We can select arbitrary constants a, b, g and h so as to simplify the data and the find r_{UV} which gives the result r_{XY}. Thus, if any constant is added or subtracted to the variables or the variables are multiplied or divided by any constant, the correlation coefficient between these two variables does not change. This leads to a short cut method explained in the following example. This simplification is not necessary for finding correlation coefficient on computer using MS Excel.

Example 15

The data of advertisement expenditure (X) and sales (Y) of a company for past 10 year period is given below. Determine the correlation coefficient between these variables and comment the correlation.

X	50	50	50	40	30	20	20	15	10	5
Y	700	650	600	500	450	400	300	250	210	200

Solution:

Using Manual Calculations

We shall take U to be the deviation of X values from the assumed mean of 30 divided by 5. Similarly, V represents the deviation of Y values from the assumed mean of 400 divided by 10.

Short cut procedure for calculation of correlation coefficient

Sl. No.	$X = x_i$	$Y = y_i$	$U = u_i$	$V = v_i$	$u_i v_i$	u_i^2	v_i^2
1	50	700	4	30	120	16	900
2	50	650	4	25	100	16	625
3	50	600	4	20	80	16	400
4	40	500	2	10	20	4	100
5	30	450	0	5	0	0	25
6	20	400	-2	0	0	4	0
7	20	300	-2	-10	20	4	100
8	15	250	-3	-15	45	9	225
9	10	210	-4	-19	76	16	361
10	5	200	-5	-20	100	25	400
Total			-2	26	561	110	3136

$$r = \frac{\sum_{i=1}^{n} u_i v_i - \frac{1}{n}\sum_{i=1}^{n} u_i \sum_{i=1}^{n} v_i}{\sqrt{\sum_{i=1}^{n} u_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} u_i\right)^2}\sqrt{\sum_{i=1}^{n} v_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} v_i\right)^2}}$$

$$= \frac{561 - \frac{(-2)(26)}{10}}{\sqrt{110 - \frac{4}{10}}\sqrt{3136 - \frac{676}{10}}} = \frac{561 + 5.2}{\sqrt{109.6}\sqrt{3068.4}} = 0.976$$

Check Your Progress 1

Fill in the blanks:

1. ..... is the most representative average in the construction of index numbers.
2. ..... is defined as the reciprocal of the arithmetic mean of the reciprocals of the individual observations.
3. ..... is the value, which divides the distribution of data, arranged in ascending or descending order, into two equal parts.
4. The ..... of a data set is the value that occurs most frequently.
5. ..... is a degree of linear association between two random variables.

1.8 REGRESSION

We need to have statistical model that will extract information from the given data to establish the regression relationship between independent and dependent relationship. The model should capture systematic behaviour of data. The non-systematic behaviour cannot be captured and called as errors. The error is due to random component that cannot be predicted as well as the component not adequately considered in statistical mndel. Good statistical model captures the entire systematic component leaving only random errors.

In any model, we attempt to capture everything which is systematic in data. Random errors cannot be captured in any case. Assuming the random errors are 'Normally distributed' we can specify the confidence level and interval of random errors. Thus, our estimates are more reliable.

If the variables in a bivariate distribution are correlated, the points in scatter diagram approximately cluster around some curve. If the curve is straight line we call it as linear regression. Otherwise, it is curvilinear regression. The equation of the curve which is closest to the observations is called the 'best fit'.

The best fit is calculated as per Legender's principle of least sum squares of deviations of the observed data points from the corresponding values on the 'best fit' curve. This is called s minimum squared error criteria. It may be noted that the deviation (error) can be measured in X direction or Y direction. Accordingly we will get two 'best fit' curves. If we measure deviation in Y direction, i.e. for a given value of data point (x_i, y_i), then we measure corresponding y value on 'beast fit' curve and then take the value of deviation in y, we call it as regression of Y on X. In the other case, if we measure deviations in X direction we call it as regression of X and Y.

Applicability of Regression Analysis

Regression analysis is one of the most popular and commonly used statistical tools in business. With availability of computer packages, it has simplified the use. However, one must be careful before using this tool as it gives only mathematical measure based on available data. It does not check whether the cause effect relationship really exists and if it exists which is dependent and which is dependent variable. Regression analysis helps in the following way:

1. It provides mathematical relationship between two or more variables. This mathematical relationship can then be used for further analysis and treatment of information using more complex techniques.
2. Since most of the business analysis and decisions are based on cause-effect relationships, regression analysis is highly valuable tool to provide mathematical model for this relationship.
3. Most wide use of regression analysis is of course estimation and forecast.
4. Regression analysis is also used in establishing the theories based on relationships of various parameters. Some of the common examples are demand and supply, money supply and expenditure, inflation and interest rates, promotion expenditure and sales, productivity and profitability, health of workers and absenteeism, etc.

Simple Regression

This model is used if we have bivariate distribution i.e., only two variables are considered and the 'best fit' curve is approximated to a straight line. This describes the liner relationship between two variables. Although it appears to be too simplistic, in many bnsiness situations, it is adequate. At least, initial

study can be based on this model for any decision-making situation. Then we could either use other models of some ad hoc methods to cater for the complexity of the business situation. If the system is found to have many nonrandom components we may have to discard this model and use some other model. This model assumes the errors are purely due to randomness and all nonrandom fluctuations are captured by our 'best fit' curve. Thus we can use the regression analysis for prediction of dependent variable for a given value of independent variable or for controlling the independent variable to get the desired results or to explain relationship for reliable predictions.

Simple Linear Regression Model

The linear regression model uses straight line relationship. Equation of a straight line is of the form,

$$\hat{y} = \alpha + \beta x \qquad \ldots (1)$$

Where $\hat{y}$ is the predicted value of Y corresponding to x. α and β are constants. Now if we assume the error (deviation) in Y direction is $\in$, we can write the relationship of X and Y in data points as,

$$y = \alpha + \beta x + \in$$

Error $\in$ is the amonnt by which observation will fall off regression line. Error $\in$ is due to random error 'α' and 'β' are called parameters of the linear regression model whose values are found out from the observed data.

In case of nonlinear equation we use the equation,

$$7y = \alpha + \beta x + \delta x^2 + \ldots + \in$$

The highest power of x is called as order of the model.

Now in model $Y = \alpha + \beta x + \in$ we cannot find $\in$ since it changes from observation to observation. But values of α and β are fixed. However, to know the exact values of α and β we need to know all values of the population which is not the usually feasible. Further, if we know the entire population, regression analysis may not have much utility. Thus, we can only find estimates of α and β from the sample data or past data. We indicate it as 'a' and 'b'.

If we fit a straight line in scattered data points, obviously some of the points would be above the line and some below. The deviation of each point from line is called error. We want the error should be as small as possible. The least square criterion is most commonly used. In this case, we minimize the value of sum of square of the errors. We could also use criteria like sum of minimum absolute deviation. But the least square criterion is superior because,

1. It is simple to interpret.
2. Easy to treat mathematically.
3. Estimate of quality of fit and confidence intervals can be easily stated.

Linear Regression Equation

Suppose the data points are (x_1, y_1) (x_2, y_2) ....(x_n, y_n). Then we can write from regression equation,

$$yi = a + bx_i + \in_i \quad i = 1, 2, \ldots, n. \qquad \ldots (2)$$

$$or, \in i = (y_i - a - bx_i)$$

Thus, sum square of errors is,

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum (y_i - a - bx_i)^2$$

To have minimum sum of squares of errors (SSE) we must have the condition,

$$\frac{\partial S}{\partial a} = \frac{\partial S}{\partial b} = 0$$

Or, $2 \times \sum (y_i - a - bx_i) = 0$

And, $2 \times \sum x_i (y_i - a - bx_i) = 0$

Thus we obtain two linear equations in a and b. These are,

$$a \times n + b \times \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad \ldots (3)$$

$$a \times \sum_{i=1}^{n} x_i + b \sum x_i^2 = \sum_{i=1}^{n} x_i y_i \quad \ldots (4)$$

These two equations are called as **'Normal Equations'**. By solving these equations, we get the values of a and b. Note that these values are estimates of α and β. Alternatively, dividing (3) by n we get,

$$a + \frac{b \times \sum x_i}{n} = \frac{\sum y_i}{n}$$

$$a + b\overline{X} = \overline{Y} \quad \ldots (5)$$

$$a + \overline{Y} - \overline{X} \quad \ldots (6)$$

Substituting (6) in (4) and dividing it by n we get,

$$b \frac{\frac{1}{n} \times \sum x_i y_i - \overline{X} \times \overline{Y}}{\frac{1}{n} \times \sum x_i^2 - \overline{X}^2} \quad \ldots (7)$$

We denote b as b_{yx} only to indicate it is regression of Y on X. b_{yx} is called as Regression Coefficient.

Now equation of regression line is,

$$\hat{y} = a + b_{yx} x$$

Subtracting equation (5) we get

$$(\hat{y} - \overline{Y}) = b_{yx} (x - \overline{X}) \quad \ldots (8)$$

$$\text{And } b_{yx} = \frac{\text{cov}(x,y)}{\sigma_x^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i}{n} \times \frac{\sum_{i=1}^{n} y_i}{n}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{\sum_{i=1}^{n} xi}{n}\right)^2} \quad \ldots (9)$$

For finding regression equation of X on Y we follow similar procedure and get the regression line equation as

$$(\hat{x} - \bar{X}) = b_{xy} (y - \bar{Y}) \quad \ldots (10)$$

$$\text{With } b_{xy} = \frac{\text{cov}(X,Y)}{\sigma_y^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i}{n} \times \frac{\sum_{i=1}^{n} y_i}{n}}{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2} \quad \ldots (11)$$

Further, covariance of (X, Y) is,

$$\text{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - x_i\bar{Y} - \bar{X}y_i - \bar{X}\bar{Y})$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{Y}\frac{\sum_{i=1}^{n} x_i}{n} - \bar{X}\frac{\sum_{i=1}^{n} y_i}{n} + \bar{X}\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{Y}\bar{X} - \bar{X}\bar{Y} + \bar{X}\bar{Y}$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{X}\bar{Y} \quad \ldots (12)$$

Also, variance of X is,

$$\text{var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{X} + \bar{X}^2)$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{X}\frac{\sum_{i=1}^{n} x_i}{n} + \frac{\bar{X}^2}{n}\sum_{i=1}^{n} 1 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{X}^2 + \bar{X}^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{X}^2 \qquad \ldots (13)$$

Substituting (11) & (13) in (7)

$$b_{YX} = \frac{\text{cov}(X,Y)}{\text{var}(X)} = \frac{\text{cov}(X,Y)}{\sigma_X^2} \qquad \ldots (14)$$

Further, we note that

$$\text{cov}(X,Y) = b_{YX}\sigma_X^2 = b_{XY}\sigma_Y^2$$

Also, using correlation coefficient is, $r = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$

Thus we get,

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X} \text{ and } b_{XY} = r\frac{\sigma_X}{\sigma_Y} \qquad \ldots (15)$$

$r^2 = b_{yx} b_{xy}$

Thus, $r = \pm\sqrt{b_{YX} b_{XY}}$ ... (16)

The implications of the above statements are,

1. Slopes of regression lines of Y on X and X on Y viz. b_{yx} and bxy must have same signs (because r^2 cannot be negative).
2. Correlation coefficient is geometric mean of b_{yx} and b_{xy}.
3. If both slopes b_{yx} and b_{xy} are positive correlation coefficient r is positive. If both b_{yx} and b_{xy} are negative the correlation coefficient r is negative.
4. If $b_{YX} = \frac{1}{b_{XY}}, \Rightarrow r = \pm 1$ indicating perfect correlation.
5. Both regression lines intersect at point $(\bar{X}, \bar{Y})$

Example 16

The cost of total output in a factory is linearly related to number of units manufactured. Data collected for 8 months is as follows.

Month	1	2	3	4	5	6	7	8
X('000 Units)	2	3	1	2.5	3.5	4	5	5.5
Y('000 Rs.)	15	16	13	15	17	18	19	20

1. Find best fit linear relationship of cost Y on units X.

2. Compute correlation coefficient and assess whether relation can be deemed as reasonable valid.
3. Estimate the cost for 13,500 units.

Solution

Using Manual Calculations

The calculations are tabulated below.

Month	x_i ('000 Units)	y_i ('000 Rs.)	x_i^2	y_i^2	$x_i y_i$
1	2	15	4	225	30
2	3	16	9	256	48
3	1	13	1	169	13
4	2.5	15	6.25	225	37.5
5	3.5	17	12.25	289	59.5
6	4	18	16	324	72
7	5	19	25	361	95
8	5.5	20	30.25	400	110
Total Σ	26.5	133	103.75	2249	465

1. Now, $$b_{YX} = \frac{\text{cov}(X,Y)}{\sigma_X^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i}{n} \times \frac{\sum_{i=1}^{n} y_i}{n}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{\sum_{i=1}^{n} x_i}{n})^2}$$

$$\text{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i}{n} \times \frac{\sum_{i=1}^{n} y_i}{n} = \frac{465}{8} - \frac{26.5 \times 133}{8 \times 8} = 58.125 - 55.07 = 3.055$$

$$\text{And, } \sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{\sum_{i=1}^{n} x_i}{n})^2 = \frac{103.75}{8} - (3.3125)^2 = 12.96875 - 10.973 = 1.99575$$

Therefore, $\sigma_X = 1.4127$

$$\text{Thus, } b_{YX} = \frac{58.125 - 55.07}{12.96875 - 10.973} = 1.53$$

The regression equation is

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

$$\text{Or, } (\hat{y}-\frac{133}{8})=1.53\times(x-\frac{26.5}{8})\Rightarrow(\hat{y}-16.625)=1.53\times(x-3.3125)$$

$$\text{Or, } \hat{y}=1.53x+11.557 \qquad \textbf{(Ans)}$$

2. Now, $\sigma_y^2=\frac{1}{n}\sum_{i=1}^{n}y_i^2-(\frac{\sum_{i=1}^{n}y_i}{n})^2=\frac{2249}{8}-(\frac{133}{8})^2=281.125\text{-}276.391=4.734$

Therefore, $\sigma_y=2.176$

$$\text{Hence, } r=b_{YX}\times\frac{\sigma_X}{\sigma_Y}=1.53\times\frac{1.4127}{2.176}=0.993 \qquad \textbf{(Ans)}$$

Note: Since correlation coefficient r is close to 1, there is strong association. Hence the relation can be deemed as reasonable valid. We can also find the significance of correlation coefficient with t test as,

$$t=r\times\sqrt{\frac{(n-2)}{(1-r^2)}}=0.993\times20.738=20.593$$

t value as per table of degrees of freedom $df=(n-2)=6$ significance level 5% is, 1.943

Since the calculated value is greater than the value from the table, association is significant. (In the next section we will see how to estimate the goodness of regression line fit).

3. For number of units 13500, $x = 13.5$. The estimated cost of output is,

$$\hat{y}=1.53x+11.557=1.53\times13.5+11.557=32.212 \qquad \textbf{(Ans)}$$

1.9 INDEX NUMBER

Index numbers are statistical measures designed to show changes in a variable or group of related variables with respect to time, geographic location or other characteristics such as income, profession, etc. Index Number is a number that expresses the relative change in price, quantity, or value compared to a base period. A collection of index numbers for different years, locations, etc., is sometimes called an index series. If the index number is used to measure the relative change in just one variable, such as hourly wages in manufacturing, we refer to this as a simple index. It is the ratio of two values of the variable and that ratio converted to a percentage. The following four examples will serve to illustrate the use of index numbers.

Index numbers are meant to study the change in the effects of such factors which cannot be measured directly. According to Bowley, "Index numbers are used to measure the changes in some quantity which we cannot observe directly". For example, changes in business activity in a country are not capable of direct measurement but it is possible to study relative changes in business activity by studying the

variations in the values of some such factors which affect business activity, and which are capable of direct measurement.

Index numbers are commonly used statistical device for measuring the combined fluctuations in a group related variables. If we wish to compare the price level of consumer items today with that prevalent ten years ago, we are not interested in comparing the prices of only one item, but in comparing some sort of average price levels. We may wish to compare the present agricultural production or industrial production with that at the time of independence. Here again, we have to consider all items of production and each item may have undergone a different fractional increase (or even a decrease). How do we obtain a composite measure? This composite measure is provided by index numbers which may be defined as a device for combining the variations that have come in group of related variables over a period of time, with a view to obtain a figure that represents the 'net' result of the change in the constitute variables.

Index numbers may be classified in terms of the variables that they are intended to measure. In business, different groups of variables in the measurement of which index number techniques are commonly used are (i) price, (ii) quantity, (iii) value and (iv) business activity. Thus, we have index of wholesale prices, index of consumer prices, index of industrial output, index of value of exports and index of business activity, etc. Here we shall be mainly interested in index numbers of prices showing changes with respect to time, although methods described can be applied to other cases. In general, the present level of prices is compared with the level of prices in the past. The present period is called the current period and some period in the past is called the base period.

Simple Index Number:

A simple index number is a number that measures a relative change in a single variable with respect to a base.

Composite Index Number:

A composite index number is a number that measures an average relative changes in a group of relative variables with respect to a base.

Types of Index Numbers:

Following types of index numbers are usually used:

Price index Numbers:

Price index numbers measure the relative changes in prices of a commodities between two periods. Prices can be either retail or wholesale.

Quantity Index Numbers:

These index numbers are considered to measure changes in the physical quantity of goods produced, consumed or sold of an item or a group of items.

Constructing an Index

$$\text{Index for any time period n} = \frac{\text{value in period n}}{\text{value in base period}} \times 100$$

Example 17

	Year	Value	Calculation	Index
(Base year is year 1)	1	12380	12380/12380 ×100	100.00
	2	12490	12490/12380 × 100	100.88
	3	12730	12730/12380 × 100	
	4	13145		

Finding values from indexes

If we have a series of index numbers describing index linked values and know the value associated with any one of them, the other values can be found by scaling the known value in the same ratio as the relevant index numbers.

Example 18

Calculate the index number for 2012 taking 2000 as the base for the following data

Commodity	Unit	Prices 2000 (P_0)	Prices 2011 (P_1)
A	Kilogram	2.50	4.00
B	Dozen	5.40	7.20
C	Meter	6.00	7.00
D	Quintal	150.00	200.00
E	Liter	2.50	3.00
Total		166.40	221.20

$$\text{Price index number} = P_{01} = \frac{\sum P_1}{\sum P_0} \times 100 = \frac{221.20}{166.40} \times 100 = 132.93$$

∴ There is a net increase of 32.93% in 2012 as compared to 2000.

Check Your Progress 2

Fill in the blanks:

1. ..... model is used if we have bivariate distribution i.e., only two variables are considered and the 'best fit' curve is approximated to a straight line.
2. ..... is a measure of spread of data about regression line.
3. ..... is a number that expresses the relative change in price, quantity, or value compared to a base period.
4. A ..... is a number that measures a relative change in a single variable with respect to a base.
5. ..... index numbers measure the relative changes in prices of a commodities between two periods.

1.10 SUMMARY

Measures of the central tendency give one of the very important characteristics of the data. According to the situation, one of the various measures of central tendency may be chosen as the most representative. Arithmetic mean is widely used and understood. What characterizes the three measures of centrality, and what are the relative merits of each in the given situation, is the question. Mean summarizes all the information in the data. Mean can be visualized as a single point where all the mass (the weight) of the observations is concentrated. It is like a centre of gravity in physics. Mean also has some desirable mathematical properties that make it useful in the context of statistical inference. Median is the middle value when the data is arranged in order. The median is resistant to the extreme observations. Median is like the geometric centre in physics. In case we want to guard against the influence of a few outlying observations (called outliers), we may use the median. The mode tells our data set's most frequently occurring value.

A scatter plot of the variables may suggest that the two variables are related but the value of the Pearson correlation coefficient r quantifies this association. The correlation coefficient r may assume values between –1 and 1. Regression provides us a measure of the relationship and also facilitates to predict one variable for a value of other variable. Thus, unlike correlation analysis, in regression analysis, one variable is independent and other dependent. Please note that this relationship need not be a cause-effect relationship. A scatter plot helps us in getting rough idea about regression. For regression analysis we need to specify independent and dependent variables clearly. In case of correlation we are only interested in finding whether the relationship exists. Hence the measuring error is only to establish confidence in our analysis. However, in regression our analysis itself is based on the concept of minimizing the errors. Index numbers are statistical measures designed to show changes in a variable or group of related variables with respect to time, geographic location or other characteristics such as income, profession, etc. A collection of index numbers for different years, locations, etc., is sometimes called an index series. If the index number is used to measure the relative change in just one variable, such as hourly wages in manufacturing, we refer to this as a simple index.

1.11 KEYWORDS

- Arithmetic Mean (AM)
- Composite Index Number
- Correlation coefficient
- Harmonic Mean (HM)
- Linear and nonlinear correlation
- Median
- Partial or total correlation
- Practical application of correlation
- Quantity Index Numbers
- Simple Index Number
- Simple or multiple correlations
- Central tendency
- Correlation
- Geometric Mean (GM)
- Index numbers
- Mean
- Mode
- Positive or negative correlation
- Price index Numbers
- Regression
- Simple linear correlation

1.12 REVIEW QUESTIONS

1. What is measure of Central Tendency?
2. What are the common measures of central tendency?
3. Define mean, median and mode.
4. What are the three types of mean?
5. What is the relationship among Mean, Median and Mode?
6. What is correlation?
7. Discuss about positive or negative correlation.
8. What is simple or multiple correlations?
9. Write short note on partial or total correlation, linear and nonlinear correlation.
10. Discuss the practical application of correlation.
11. What is regression?
12. Explain the applicability of regression analysis.
13. What is an index number? What does it measure?
14. Define price index number?
15. Define quantity index number?
16. Compute mean, median, mode quartiles and 90th percentile for data given below:

22	21	37	33	28	42	56	33	32	59
40	47	29	65	45	48	55	43	42	40
37	39	56	54	38	49	60	37	28	27
32	33	47	36	35	42	43	55	53	48
29	30	32	37	43	54	55	47	38	62

17. Compute mean, median, mode, quartiles and 90th percentile for the grouped data of age (years) of employees given below:

Class Interval	20-30	30-40	40-50	50-60	60-70
Frequency	7	16	15	9	3

18. Calculate arithmetic mean and mode from the following:

Monthly salary Rs.	400-600	600-800	800-1000	1000-1200	1200-1400
Number of workers	4	10	12	6	2

19. For the following data find the missing frequency. It is given that mean is 15.38.

Class	9-11	11-13	13-15	15-17	17-19	19-21
Frequency	3	7	12	20	?	5

20. Find Mean and Median for the following data:

41, 42, 43, 42, 43, 42, 44, 44, 40, 45

21. Find the missing frequency, if arithmetic mean for the following data is 28.

Marks	0-10	10-20	20-30	30-40	40-50	50-60
Number of Students	12	18	27	?	17	6

22. Find mode of the following data:

Class	0-20	20-40	40-60	60-80	80-100	100-120
Frequency	12	28	16	5	3	1

23. The mean marks of 100 students were found to be 40. Later on, it was discovered that the score of 53 was misread as 83. Find correct mean corresponding to correct data.

24. The monthly income (in Rs.) of 7 families in a village is as follows:

1200, 1000, 1100, 1250, 950, 1100, 1350

Calculate median and mode.

25. Calculate coefficient of correlation between advertisement cost and sales as per the data given below:

Advertisement cost in '000 Rs.	39	65	62	90	82	75	25	98	36	78
Sales in Lakh Rs.	47	53	58	86	62	68	60	91	51	84

26. Calculate coefficient of correlation between X and Y as per the data given below:

X	14	16	20	22	28	30	34	40	45
Y	97	89	68	65	56	50	37	18	12

27. If $b_{XY} = \frac{3}{2}$ and $b_{YX} = \frac{1}{6}$, find the value of correlation coefficient between X and Y.

28. For the following data

	X	Y
Mean	36	85
Standard Deviation	11	8

The correlation coefficient between X and Y is 0.66. Find regression equation of X on Y, hence estimate the value of X when Y = 80.

29. Construct the price index number for 2012, taking the year 2000 as base year.

Commodity	Price in the year (2000)	Price in the year (2012)
A	60	80
B	50	60
C	70	100
D	120	160
E	100	150

30. Compute the index number for the years 2001, 2002, 2003 and 2004, taking 2000 as base year, from the following data:

Answers to Check Your Progress

Check Your Progress 1

1. Geometric Mean
2. Harmonic Mean
3. Median
4. Mode
5. Correlation

Check Your Progress 2

1. Simple Regression
2. Error variance
3. Index Number
4. Simple index number
5. Price

1.13 REFERENCES AND FURTHER READING

- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2021). Multivariate data analysis (8th ed.). Cengage Learning.
- Pallant, J. (2022). SPSS survival manual: A step by step guide to data analysis using SPSS (7th ed.). Open University Press.
- Boslaugh, S. (2023). Statistics in a nutshell (2nd ed.). O'Reilly Media.
- Cohen, L., Manion, L., & Morrison, K. (2024). Research methods in education (9th ed.). Routledge.

UNIT 2

Probability Distributions

CHAPTER OUTLINE

2.1 INTRODUCTION

Usual manager is forced to make decisions when there is uncertainty as to what will happen after the decisions are made. In this situation the mathematical theory of probability furnishes a tool that can be of great help to the decision maker. A probability function is a rule that assigns probabilities to each element of a set of events that may occur. Probability distribution can either discrete or continuous. A discrete probability distribution is sometimes called a probability mass function and a continuous one is called a probability density function.

2.2 CONCEPT OF PROBABILITY

The concept of probability originated from the analysis of the games of chance in the 17th century. Now the subject has been developed to the extent that it is very difficult to imagine a discipline, be it from social or natural sciences that can do without it. The theory of probability is a study of Statistical or Random Experiments. It is the backbone of Statistical Inference and Decision Theory that are essential tools of the analysis of most of the modern business and economic problems.

Often, in our day-to-day life, we hear sentences like 'it may rain today', 'Mr. X has fifty-fifty chances of passing the examination', 'India may win the forthcoming cricket match against Sri Lanka', 'the chances of making profits by investing in shares of company A are very bright', etc. Each of the above sentences involves an element of uncertainty.

A phenomenon or an experiment which can result into more than one possible outcome is called a random phenomenon or random experiment or statistical experiment. Although, we may be aware of all the possible outcomes of a random experiment, it is not possible to predetermine the outcome associated with a particular experimentation or trial.

Consider, for example, the toss of a coin. The result of a toss can be a head or a tail; therefore, it is a random experiment. Here we know that either a head or a tail would occur as a result of the toss, however, it is not possible to predetermine the outcome. With the use of probability theory, it is possible to assign a quantitative measure, to express the extent of uncertainty, associated with the occurrence of each possible outcome of a random experiment.

Addition Theorem

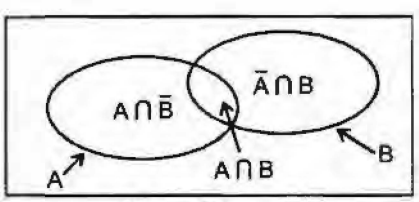
A compound event is any event combining two or more simple events.

The notation for addition rule is: $P(A \text{ or } B)$ = P(event A occurs or event B occurs or they both occur).

When finding the probability that event A occurs or event B occurs, find the total numbers of ways A can occurs and the number of ways B can occurs, but find the total in such a way that no outcome is counted more than once.

General addition rule is:

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



Venn Diagram

Proof: From the Venn diagram, we can write

$$A \cup B = A \cup (\bar{A} \cap B) \text{ or } P(A \cup B) = P[A \cup (\bar{A} \cap B)]$$

Since A and $(\bar{A} \cap B)$ are mutually exclusive, we can write

$$P(A \cup B) = P(A) + P(\bar{A} \cap B)$$

Substituting the value of $P(\bar{A} \cap B)$ from theorem 3, we get

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

Remarks:

1. If A and B are mutually exclusive, i.e., $A \cap B = \phi$, then according to theorem 1, we have $P(A \cap B) = 0$. The addition rule, in this case, becomes $P(A \cup B) = P(A) + P(B)$, which is in conformity with axiom III.
2. The event $A \cup B$ denotes the occurrence of either A or B or both. Alternatively, it implies the occurrence of at least one of the two events.
3. The event $A \cap B$ is a compound event that denotes the simultaneous occurrence of the two events.
4. Alternatively, the event $A \cup B$ is also denoted by $A + B$ and the event $A \cap B$ by AB.

Corollaries:

1. From the Venn diagram, we can write $P(A \cup B) = 1 - P(\bar{A} \cap \bar{B})$, where $P(\bar{A} \cap \bar{B})$ is the probability that none of the events A and B occur simultaneously.
2. $P(\text{exactly one of } A \text{ and } B \text{ occurs}) = P[(A \cap \bar{B}) \cup (\bar{A} \cap B)]$

   $= P(A \cap \bar{B}) + P(\bar{A} \cap B)$ $\quad [\text{Since } (A \cap \bar{B}) \cup (\bar{A} \cap B) = \phi]$

   $= P(A) - P(A \cap B) + P(B) - P(A \cap B)$ $\quad [\text{Since } (\bar{A} \cap B) = P(B) - P(A \cap B)]$

   $= P(A \cup B) - P(A \cap B)$ $\quad [\text{Since } P(A \cup B) = P(A) + P(B) - P(A \cap B)]$
3. The addition theorem can be generalised for more than two events. If A, B and C are three events of a sample space S, then the probability of occurrence of at least one of them is given by

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

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

   Applying theorem 4 on the second and third term, we get

   $$= P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \quad \ldots. (1)$$

   Alternatively, the probability of occurrence of at least one of the three events can also be written as

   $$P(A \cup B \cup C) = 1 - P(\bar{A} \cap \bar{B} \cap \bar{C}) \quad \ldots. (2)$$

If A, B and C are mutually exclusive, then equation (1) can be written as

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) \quad \text{.... (3)}$$

If $A_1, A_2, \ldots\ldots A_n$ are n events of a sample space S, the respective equations (1), (2) and (3) can be modified as

$$P(A_1 \cup A_2 \ldots \cup A_n) = \Sigma P(A_i) - \Sigma\Sigma P(A_i \cap A_j) + \Sigma\Sigma\Sigma P(A_i \cap A_j \cap A_k)$$

$$+(-1)^n P(A_1 \cap A_2 \cap \ldots \cap A_n) \ (i \neq j \neq k, \text{ etc.}) \quad \text{.... (4)}$$

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = 1 - P(\bar{A}_1 \cap \bar{A}_2 \cap \ldots \cap \bar{A}_n) \quad \text{.... (5)}$$

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_{i=1}^{n} P(A_i) \quad \text{.... (6)}$$

(if the events are mutually exclusive)

4. The probability of occurrence of at least two of the three events can be written as:

$$P[(A \cap B) \cup (B \cap C) \cup (A \cap \bar{C})] = P(A \cap B) + P(B \cap C) + P(A \cap \bar{C}) - 3P(A \cap B \cap C) + P(A \cap B \cap C)$$

$$= P(A \cap B) + P(B \cap C) + P(A \cap C) - 2P(A \cap B \cap C)$$

5. The probability of occurrence of exactly two of the three events can be written as:

$$P[(A \cap B \cap \bar{C}) \cup (A \cap \bar{B} \cap C) \cup (\bar{A} \cap B \cap C)] = P[(A \cap B) \cup (B \cap C) \cup (A \cap C)] - P(A \cap B \cap C)$$

(using corollary 2)

$$= P(A \cap B) + P(B \cap C) + P(A \cap C) - 3P(A \cap B \cap C) \text{ (using corollary 4)}$$

6. The probability of occurrence of exactly one of the three events can be written as:

$P[(A \cap \bar{B} \cap \bar{C}) \cup (\bar{A} \cap B \cap \bar{C}) \cup (\bar{A} \cap \bar{B} \cap C)]$ = P(at least one of the three events occur) – P(at least two of the three events occur).

$$= P(A) + P(B) + P(C) - 2P(A \cap B) - 3P(B \cap C) - 2P(A \cap C) + 3P(A \cap B \cap C).$$

Example 2.1: In a group of 1,000 persons, there are 650 who can speak Hindi, 400 can speak English and 150 can speak both Hindi and English. If a person is selected at random, what is the probability that he speaks (i) Hindi only, (ii) English only, (iii) only one of the two languages, (iv) at least one of the two languages?

Solution: Let A denote the event that a person selected at random speaks Hindi and B denotes the event that he speaks English.

Thus, we have $n(A) = 650$, $n(B) = 400$, $n(A \cap B) = 150$ and $n(S) = 1000$,

where $n(A)$, $n(B)$, etc. denote the number of persons belonging to the respective event.

(i) The probability that a person selected at random speaks Hindi only, is given by

$$P(A\cap\bar{B})=\frac{n(A)}{n(S)}-\frac{n(A\cap B)}{n(S)}=\frac{650}{1000}-\frac{150}{1000}=\frac{1}{2}$$

(ii) The probability that a person selected at random speaks English only, is given by

$$P(\bar{A}\cap B)=\frac{n(B)}{n(S)}-\frac{n(A\cap B)}{n(S)}=\frac{400}{1000}-\frac{150}{1000}=\frac{1}{4}$$

(iii) The probability that a person selected at random speaks only one of the languages, is given by

$$P\left[(A\cap\bar{B})\cup(\bar{A}\cap B)\right]=P(A)+P(B)-2P(A\cap B) \quad \text{(see corollary 2)}$$

$$=\frac{n(A)+n(B)-2n(A\cap B)}{n(S)}=\frac{650+400-300}{1000}=\frac{3}{4}$$

(iv) The probability that a person selected at random speaks at least one of the languages, is given by

$$P(A\cup B)=\frac{650+400-150}{1000}=\frac{9}{10}$$

Alternative Method: The above probabilities can easily be computed by the following nine-square table:

	B	$\bar{B}$	*Total*
A	150	500	650
$\bar{A}$	250	100	350
Total	400	600	1000

From the above table, we can write

(i) $P(A\cap\bar{B})=\frac{500}{1000}=\frac{1}{2}$

(ii) $P(\bar{A}\cap B)=\frac{250}{1000}=\frac{1}{4}$

(iii) $P\left[(A\cap\bar{B})\cup(\bar{A}\cap B)\right]=\frac{500+250}{1000}=\frac{3}{4}$

(iv) $P(A\cup B)=\frac{150+500+250}{1000}=\frac{9}{10}$

This can, alternatively, be written as $P(A\cup B)=1-P(\bar{A}\cap\bar{B})=1-\frac{100}{1000}=\frac{9}{10}$.

Multiplication or Compound Probability Theorem

A compound event is the result of the simultaneous occurrence of two or more events. For convenience, we assume that there are two events; however, the results can be easily generalized. The probability of the compound event would depend upon whether the events are independent or not. Thus, we shall discuss two theorems; (a) *Conditional Probability Theorem*, and (b) *Multiplicative Theorem for Independent Events*.

(a) ***Conditional Probability Theorem:*** For any two events A and B in a sample space S, the probability of their simultaneous occurrence, is given by

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

or equivalently $= P(B)P(A/B)$

Here, $P(B/A)$ is the conditional probability of B given that A has already occurred. Similar interpretation can be given to the term $P(A/B)$.

Proof: Let all the outcomes of the random experiment be equally likely. Therefore,

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{\text{no. of elements in } (A \cap B)}{\text{no. of elements in sample space}}$$

For the event B/A, the sample space is the set of elements in A and out of these the number of cases favourable to B is given by $n(A \cap B)$.

$$\therefore \quad P(B/A) = \frac{n(A \cap B)}{n(A)}.$$

If we multiply the numerator and denominator of the above expression by $n(S)$, we get

$$P(B/A) = \frac{n(A \cap B)}{n(A)} \times \frac{n(S)}{n(S)} = \frac{P(A \cap B)}{P(A)}$$

or $P(A \cap B) = P(A).P(B/A)$.

The other result can also be shown in a similar way.

Note: To avoid mathematical complications, we have assumed that the elementary events are equally likely. However, the above results will hold true even for the cases where the elementary events are not equally likely.

(b) ***Multiplicative Theorem for Independent Events:*** If A and B are independent, the probability of their simultaneous occurrence is given by $P(A \cap B) = P(A).P(B)$.

Proof: We can write $A = (A \cap B) \cup (A \cap \overline{B})$.

Since $(A \cap B)$ and $(A \cap \overline{B})$ are mutually exclusive, we have

$P(A) = P(A \cap B) + P(A \cap \overline{B})$ (by axiom III)

$= P(B).P(A/B) + P(\overline{B}).P(A/\overline{B})$

If A and B are independent, then proportion of A's in B is equal to proportion of A's in $\overline{B}$'s, i.e., $P(A/B) = P(A/\overline{B})$.

Thus, the above equation can be written as

$$n(B)=\frac{600\times 30}{100}+\frac{400\times 5}{100}=200$$

Substituting this value in the formula of conditional probability theorem, we get

$$P(A\cap B)=P(A).P(B).$$

Corollaries:

1. (i) If A and B are mutually exclusive and $P(A).P(B) > 0$, then they cannot be independent since $P(A\cap B)=0$.

   (ii) If A and B are independent and $P(A).P(B) > 0$, then they cannot be mutually exclusive since $P(A\cap B)>0$.

2. Generalisation of Multiplicative Theorem :

   If A, B and C are three events, then

   $$P(A\cap B\cap C)=P(A).P(B/A).P\left[C/(A\cap B)\right]$$

   Similarly, for n events A_1, A_2, ... A_n, we can write

   $$P(A_1\cap A_2\cap \ldots \cap A_n)=P(A_1).P(A_2/A_1).P\left[A_3/(A_1\cap A_2)\right]$$
   $$\ldots P\left[A_n/(A_1\cap A_2\cap \ldots \cap A_{n-1})\right]$$

   Further, if A_1, A_2, ... A_n are independent, we have

   $$P(A_1\cap A_2\cap \ldots \cap A_n)=P(A_1).P(A_2)\ldots P(A_n).$$

3. If A and B are independent, then A and $\overline{B}$, $\overline{A}$ and B, $\overline{A}$ and $\overline{B}$ are also independent.

   We can write $P(A\cap \overline{B})=P(A)-P(A\cap B)$ (by theorem 3)

   $=P(A)-P(A).P(B)=P(A)[1-P(B)]=P(A).P(\overline{B})$, which shows that A and $\overline{B}$ are independent. The other results can also be shown in a similar way.

4. The probability of occurrence of at least one of the events A_1, A_2, ... A_n, is given by

   $$P(A_1\cup A_2\cup \ldots \cup A_n)=1-P(\overline{A}_1\cap \overline{A}_2\cap \ldots \cap \overline{A}_n).$$

   If A_1, A_2, ... A_n are independent then their compliments will also be independent, therefore, the above result can be modified as

   $$P(A_1\cup A_2\cup \ldots \cup A_n)=1-P(\overline{A}_1).P(\overline{A}_2)\ldots P(\overline{A}_n).$$

Pair-wise and Mutual Independence

Three events A, B and C are said to be mutually independent if the following conditions are simultaneously satisfied :

$$P(A\cap B)=P(A).P(B),\ P(B\cap C)=P(B).P(C),\ P(A\cap C)=P(A).P(C)$$

and $P(A \cap B \cap C) = P(A).P(B).P(C)$.

If the last condition is not satisfied, the events are said to be pair-wise independent.

From the above we note that mutually independent events will always be pair-wise independent but not vice-versa.

Example 2.2: Among 1,000 applicants for admission to M.A. economics course in a University, 600 were economics graduates and 400 were non-economics graduates; 30% of economics graduate applicants and 5% of non-economics graduate applicants obtained admission. If an applicant selected at random is found to have been given admission, what is the probability that he/she is an economics graduate?

Solution: Let A be the event that the applicant selected at random is an economics graduate and B be the event that he/she is given admission.

We are given $n(S) = 1000$, $n(A) = 600$, $n(\overline{A}) = 400$

Also, $n(B) = \frac{600 \times 30}{100} + \frac{400 \times 5}{100} = 200$ and $n(A \cap B) = \frac{600 \times 30}{100} = 180$

Thus, the required probability is given by $P(A/B) = \frac{n(A \cap B)}{n(B)} = \frac{180}{200} = \frac{9}{10}$

Alternative Method: Writing the given information in a nine-square table, we have :

	B	$\overline{B}$	*Total*
A	180	420	600
$\overline{A}$	20	380	400
Total	200	800	1000

From the above table we can write $P(A/B) = \frac{180}{200} = \frac{9}{10}$

Example 2.3: A bag contains 2 black and 3 white balls. Two balls are drawn at random one after the other without replacement. Obtain the probability that (a) Second ball is black given that the first is white, (b) First ball is white given that the second is black.

Solution: First ball can be drawn in any one of the 5 ways and then a second ball can be drawn in any one of the 4 ways. Therefore, two balls can be drawn in 5 × 4 = 20 ways. Thus, $n(S) = 20$.

(a) Let A_1 be the event that first ball is white and A_2 be the event that second is black. We want to find $P(A_2 / A_1)$.

First white ball can be drawn in any of the 3 ways and then a second ball can be drawn in any of the 4 ways, $\therefore$ $n(A_1) = 3 \times 4 = 12$.

Further, first white ball can be drawn in any of the 3 ways and then a black ball can be drawn in any of the 2 ways, $\therefore$ $n(A_1 \cap A_2) = 3 \times 2 = 6$.

Thus, $P(A_2/A_1)=\frac{n(A_1\cap A_2)}{n(A_1)}=\frac{6}{12}=\frac{1}{2}$.

(b) Here we have to find $P(A_1/A_2)$.

The second black ball can be drawn in the following two mutually exclusive ways:

(i) First ball is white and second is black or

(ii) both the balls are black.

Thus, $n(A_2) = 3\times 2 + 2\times 1 = 8$, $\therefore P(A_1/A_2)=\frac{n(A_1\cap A_2)}{n(A_2)}=\frac{6}{8}=\frac{3}{4}$.

Alternative Method: The given problem can be summarised into the following nine-square table:

	B	$\bar{B}$	*Total*
A	6	6	12
$\bar{A}$	2	6	8
Total	8	12	20

The required probabilities can be directly written from the above table.

2.3 BAYES THEOREM OR INVERSE PROBABILITY RULE

The probabilities assigned to various events on the basis of the conditions of the experiment or by actual experimentation or past experience or on the basis of personal judgement are called prior probabilities. One may like to revise these probabilities in the light of certain additional or new information. This can be done with the help of Bayes Theorem, which is based on the concept of conditional probability. The revised probabilities, thus obtained, are known as *posterior* or inverse *probabilities*. Using this theorem it is possible to revise various business decisions in the light of additional information.

Bayes' Theorem

If an event D can occur only in combination with any of the n mutually exclusive and exhaustive events $A_1, A_2, \ldots\ldots A_n$ and if, in an actual observation, D is found to have occurred, then the probability that it was preceded by a particular event A_k is given by

$$P(A_k/D)=\frac{P(A_k).P(D/A_k)}{\sum_{i=1}^{n}P(A_i).P(D/A_i)}$$

Proof: Since $A_1, A_2, \ldots\ldots A_n$ are n exhaustive events, therefore,

$$S = A_1 \cup A_2 \ldots\ldots \cup A_n.$$

Since D is another event that can occur in combination with any of the mutually exclusive and exhaustive events $A_1, A_2, \ldots\ldots A_n$, we can write

$$D = (A_1 \cap D) \cup (A_2 \cap D) \cup \ldots\ldots \cup (A_n \cap D)$$

Taking probability of both sides, we get

$$P(D) = P(A_1 \cap D) + P(A_2 \cap D) + \ldots\ldots + P(A_n \cap D)$$

We note that the events $(A_1 \cap D), (A_2 \cap D)$, etc. are mutually exclusive.

$$P(D) = \sum_{i=1}^{n} P(A_i \cap D) = \sum_{i=1}^{n} P(A_i).P(D/A_i) \quad \ldots. (1)$$

The conditional probability of an event A_k given that D has already occurred, is given by

$$P(A_k/D) = \frac{P(A_k \cap D)}{P(D)} = \frac{P(A_k).P(D/A_k)}{P(D)} \quad \ldots. (2)$$

Substituting the value of $P(D)$ from (1), we get

$$P(A_k/D) = \frac{P(A_k).P(D/A_k)}{\sum_{i=1}^{n} P(A_i).P(D/A_i)} \quad \ldots. (3)$$

Example 2.4: A manufacturing firm purchases a certain component, for its manufacturing process, from three sub-contractors *A*, *B* and *C*. These supply 60%, 30% and 10% of the firm's requirements, respectively. It is known that 2%, 5% and 8% of the items supplied by the respective suppliers are defective. On a particular day, a normal shipment arrives from each of the three suppliers and the contents get mixed. A component is chosen at random from the day's shipment :

(a) What is the probability that it is defective?

(b) If this component is found to be defective, what is the probability that it was supplied by (i) *A*, (ii) *B*, (iii) *C* ?

Solution: Let *A* be the event that the item is supplied by *A*. Similarly, *B* and *C* denote the events that the item is supplied by *B* and *C* respectively. Further, let *D* be the event that the item is defective. It is given that :

$$P(A) = 0.6,\ P(B) = 0.3,\ P(C) = 0.1,\ P(D/A) = 0.02$$

$$P(D/B) = 0.05,\ P(D/C) = 0.08.$$

(a) We have to find $P(D)$

From equation (1), we can write

$$P(D) = P(A \cap D) + P(B \cap D) + P(C \cap D)$$

$$= P(A)P(D/A) + P(B)P(D/B) + P(C)P(D/C)$$

$$= 0.6 \times 0.02 + 0.3 \times 0.05 + 0.1 \times 0.08 = 0.035$$

(b) (i) We have to find $P(A/D)$

$$P(A/D)=\frac{P(A)P(D/A)}{P(D)}=\frac{0.6\times 0.02}{0.035}=0.343$$

Similarly, (ii) $P(B/D)=\frac{P(B)P(D/B)}{P(D)}=\frac{0.3\times 0.05}{0.035}=0.429$

and (iii) $P(C/D)=\frac{P(C)P(D/C)}{P(D)}=\frac{0.1\times 0.08}{0.035}=0.228$

Alternative Method: The above problem can also be attempted by writing various probabilities in the form of following table :

	A	B	C	*Total*
D	$P(A\cap D)$ $=0.012$	$P(B\cap D)$ $=0.015$	$P(C\cap D)$ $=0.008$	0.035
$\bar{D}$	$P(A\cap \bar{D})$ $=0.588$	$P(B\cap \bar{D})$ $=0.285$	$P(C\cap \bar{D})$ $=0.092$	0.965
Total	0.600	0.300	0.100	1.000

Thus $P(A/D)=\frac{0.012}{0.035}$ etc.

Example 2.5: A box contains 4 identical dice out of which three are fair and the fourth is loaded in such a way that the face marked as 5 appears in 60% of the tosses. A die is selected at random from the box and tossed. If it shows 5, what is the probability that it was a loaded die?

Solution: Let A be the event that a fair die is selected and B be the event that the loaded die is selected from the box.

Then, we have $P(A)=\frac{3}{4}$ and $P(B)=\frac{1}{4}$.

Further, let D be the event that 5 is obtained on the die, then

$$P(D/A)=\frac{1}{6} \text{ and } P(D/B)=\frac{6}{10}$$

Thus, $P(D) = P(A).P(D/A) + P(B).P(D/B) = \frac{3}{4}\times\frac{1}{6}+\frac{1}{4}\times\frac{6}{10}=\frac{11}{40}$

We want to find $P(B/D)$, which is given by

$$P(B/D)=\frac{P(B\cap D)}{P(D)}=\frac{1}{4}\times\frac{6}{10}\times\frac{40}{11}=\frac{6}{11}$$

2.4 RANDOM VARIABLES

Given any random variable, corresponding to a sample space, it is possible to associate probabilities to each of its possible values. For example, in the toss of 3 coins, assuming that they are unbiased, the probabilities of various values of the random variable X, can be written as:

$$P(X=0)=\frac{1}{8},\ P(X=1)=\frac{3}{8},\ P(X=2)=\frac{3}{8} \text{ and } P(X=3)=\frac{1}{8}.$$

The set of all possible values of the random variable X alongwith their respective probabilities is termed as Probability Distribution of X. The probability distribution of X, defined in example 1 above, can be written in a tabular form as given below:

X	:	0	1	2	3	*Total*
$p(X)$	:	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

Note that the total probability is equal to unity.

In general, the set of n possible values of a random variable X, i.e., $\{X_1, X_2, \ldots\ldots X_n\}$ along with their respective probabilities $p(X_1)$, $p(X_2)$, ... $p(X_n)$, where $\sum_{i=1}^{n} p(X_i)=1$, is called a probability distribution of X. The expression $p(X)$ is called the probability function of X.

Discrete and Continuous Probability Distributions

Like any other variable, a random variable X can be discrete or continuous. If X can take only finite or countably infinite set of values, it is termed as a discrete random variable. On the other hand, if X can take an uncountable set of infinite values, it is called a continuous random variable.

The random variable defined in example 1 is a discrete random variable. However, if X denotes the measurement of heights of persons or the time interval of arrival of a specified number of calls at a telephone desk, etc., it would be termed as a continuous random variable.

The distribution of a discrete random variable is called the Discrete Probability Distribution and the corresponding probability function $p(X)$ is called a Probability Mass Function. In order that any discrete function $p(X)$ may serve as probability function of a discrete random variable X, the following conditions must be satisfied :

(i) $p(X) \geq 0\ \forall\ i = 1, 2, \ldots\ldots n$ and

(ii) $\sum_{i=1}^{n} p(X_i)=1$

In a similar way, the distribution of a continuous random variable is called a *Continuous Probability Distribution* and the corresponding probability function $p(X)$ is termed as the *Probability Density Function.* The conditions for any function of a continuous variable to serve as a probability density function are :

(i) $p(X) \geq 0\ \forall$ real values of X, and

(ii) $\int_{-\infty}^{\infty} p(X)dX=1$

Remarks:

1. When X is a continuous random variable, there are an infinite number of points in the sample space and thus, the probability that X takes a particular value is always defined to be zero even though the event is not regarded as impossible. Hence, we always talk of the probability of a continuous random variable lying in an interval.
2. The concept of a probability distribution is not new. In fact it is another way of representing a frequency distribution. Using statistical definition, we can treat the relative frequencies of various values of the random variable as the probabilities.

Example 2.6: Two unbiased die are thrown. Let the random variable X denote the sum of points obtained. Construct the probability distribution of X.

Solution: The possible values of the random variable are :

2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

The probabilities of various values of X are shown in the following table :

Probability Distribution of X

X	2	3	4	5	6	7	8	9	10	11	12	*Total*
$p(X)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	1

Example 2.7: Three marbles are drawn at random from a bag containing 4 red and 2 white marbles. If the random variable X denotes the number of red marbles drawn, construct the probability distribution of X.

Solution: The given random variable can take 3 possible values, i.e., 1, 2 and 3. Thus, we can compute the probabilities of various values of the random variable as given below :

$$P(X = 1, \text{ i.e., } 1R \text{ and } 2\ W \text{ marbles are drawn}) = \frac{{}^4C_1 \times {}^2C_2}{{}^6C_3} = \frac{4}{20}$$

$$P(X = 2, \text{ i.e., } 2R \text{ and } 1W \text{ marbles are drawn}) = \frac{{}^4C_2 \times {}^2C_1}{{}^6C_3} = \frac{12}{20}$$

$$P(X = 3, \text{ i.e., } 3R \text{ marbles are drawn}) = \frac{{}^4C_3}{{}^6C_3} = \frac{4}{20}$$

Note: In the event of white balls being greater than 2, the possible values of the random variable would have been 0, 1, 2 and 3.

Cumulative Probability Function or Distribution Function

This concept is similar to the concept of cumulative frequency. The distribution function is denoted by F(x).

For a discrete random variable X, the distribution function or the cumulative probability function is given by $F(x) = P(X \leq x)$.

If X is a random variable that can take values, say 0, 1, 2, ..., then

F(1) = P(X = 0) + P(X =1), F(2) = P(X = 0) + P(X =1) +P(X = 2), etc.

Similarly, if X is a continuous random variable, the distribution function or cumulative probability density function is given by

$$F(x) = P(X \le x) = \int_{-\infty}^{x} p(X)dX$$

2.5 MEAN AND VARIANCE OF A RANDOM VARIABLE

The mean and variance of a random variable can be computed in a manner similar to the computation of mean and variance of the variable of a frequency distribution.

Mean

If X is a discrete random variable which can take values X_1, X_2, ... X_n, with respective probabilities as $p(X_1)$, $p(X_2)$, ... $p(X_n)$, then its mean, also known as the Mathematical Expectation or Expected Value of X, is given by:

$$E(X) = X_1p(X_1) + X_2p(X_2) + ... + X_np(X_n) = \sum_{i=1}^{n} X_i p(X_i)$$

The mean of a random variable or its probability distribution is often denoted by μ, i.e., $E(X) = \mu$

Remarks: The mean of a frequency distribution can be written as $X_1.\frac{f_1}{N} + X_2.\frac{f_2}{N} + ... + X_n.\frac{f_n}{N}$, which is identical to the expression for expected value.

Variance

The concept of variance of a random variable or its probability distribution is also similar to the concept of the variance of a frequency distribution.

The variance of a frequency distribution is given by

$$\sigma^2 = \frac{1}{N}\sum f_i\left(X_i - \bar{X}\right)^2 = \sum\left(X_i - \bar{X}\right)^2.\frac{f_i}{N} = \text{Mean of } \left(X_i - \bar{X}\right)^2 \text{ values.}$$

The expression for variance of a probability distribution with mean μ can be written in a similar way, as given below :

$$\sigma^2 = E(X-\mu)^2 = \sum_{i=1}^{n}\left(X_i-\mu\right)^2 p\left(X_i\right), \text{ where X is a discrete random variable.}$$

Remarks: If X is a continuous random variable with probability density function p(X), then

$$E(X) = \int_{-\infty}^{\infty} X.p(X)dX$$

$$\sigma^2 = E(X-\mu)^2 = \int_{-\infty}^{\infty}(X-\mu)^2.p(X)dX$$

Moments

The rth moment of a discrete random variable about its mean is defined as:

$$\mu_r = E(X-\mu)^r = \sum_{i=1}^{n}(X_i-\mu)^r\, p(X_i)$$

Similarly, the rth moment about any arbitrary value A, can be written as

$$\mu_r' = E(X-A)^r = \sum_{i=1}^{n}(X_i-A)^r\, p(X_i)$$

The expressions for the central and the raw moments, when X is a continuous random variable, can be written as

$$\mu_r = E(X-\mu)^r = \int_{-\infty}^{\infty}(X-\mu)^r . p(X)dX$$

and $\mu_r' = E(X-A)^r = \int_{-\infty}^{\infty}(X-A)^r . p(X)dX$ respectively.

2.6 EXPECTED VALUE

Theorem 1

Expected value of a constant is the constant itself, i.e., E(b) = b, where b is a constant.

Proof

The given situation can be regarded as a probability distribution in which the random variable takes a value b with probability 1 and takes some other real value, say a, with probability 0.

Thus, we can write E(b) = b × 1 + a × 0 = b

Theorem 2

E(aX) = aE(X), where X is a random variable and a is constant.

Proof

For a discrete random variable X with probability function p(X), we have :

$E(aX) = aX_1.p(X_1) + aX_2.p(X_2) + \ldots\ldots + aX_n.p(X_n)$

$$= a\sum_{i=1}^{n} X_i.p(X_i) = aE(X)$$

Combining the results of theorems 1 and 2, we can write

E(aX + b) = aE(X) + b

Remarks: Using the above result, we can write an alternative expression for the variance of X, as given below :

$$\sigma^2 = E(X - \mu)^2 = E(X^2 - 2\mu X + \mu^2)$$

$$= E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - 2\mu^2 + \mu^2$$

$$= E(X^2) - \mu^2 = E(X^2) - [E(X)]^2$$

= Mean of Squares - Square of the Mean

We note that the above expression is identical to the expression for the variance of a frequency distribution.

Theorems on Variance

Theorem 1

The variance of a constant is zero.

Proof

Let b be the given constant. We can write the expression for the variance of b as:

$Var(b) = E[b - E(b)]^2 = E[b - b]| = 0.$

Theorem 2

$Var(X + b) = Var(X).$

Proof

We can write $Var(X + b) = E[X + b - E(X + b)]^2 = E[X + b - E(X) - b]^2$

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

Similarly, it can be shown that $Var(X - b) = Var(X)$

Remarks: The above theorem shows that variance is independent of change of origin.

Theorem 3

$Var(aX) = a^2Var(X)$

Proof

We can write $Var(aX) = E[aX - E(aX)]^2 = E[aX - aE(X)]^2$

$$= a^2E[X - E(X)]^2 = a^2Var(X).$$

Combining the results of theorems 2 and 3, we can write

$Var(aX + b) = a^2Var(X).$

This result shows that the variance is independent of change origin but not of change of scale.

Remarks:

1. On the basis of the theorems on expectation and variance, we can say that if X is a random variable, then its linear combination, aX + b, is also a random variable with mean aE(X) + b and Variance equal to $a^2Var(X)$.
2. The above theorems can also be proved for a continuous random variable.

Example 2.8

Compute mean and variance of the probability distributions of following conditions.

(i)

X	0	1	2	3
$p(X)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Solution

From the above distribution, we can write

$$E(X)=0\times\frac{1}{8}+1\times\frac{3}{8}+2\times\frac{3}{8}+3\times\frac{1}{8}=1.5$$

To find variance of X, we write

Var(X) = $E(X^2)$ - $[E(X)]^2$, where $E\left(X^2\right)=\sum X^2 p(X)$

Now, $E\left(X^2\right)=0\times\frac{1}{8}+1\times\frac{3}{8}+4\times\frac{3}{8}+9\times\frac{1}{8}=3$

Thus, Var(X) = 3 - $(1.5)^2$ = 0.75

(ii)

X	2	3	4	5	6	7	8	9	10	11	12	*Total*
$p(X)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	1

Solution

$$\therefore E(X)=2\times\frac{1}{36}+3\times\frac{2}{36}+4\times\frac{3}{36}+5\times\frac{4}{36}+6\times\frac{5}{36}+7\times\frac{6}{36}$$

$$+8\times\frac{5}{36}+9\times\frac{4}{36}+10\times\frac{3}{36}+11\times\frac{2}{36}+12\times\frac{1}{36}=\frac{252}{36}=7$$

Further, $E(X^2)=4\times\frac{1}{36}+9\times\frac{2}{36}+16\times\frac{3}{36}+25\times\frac{4}{36}+36\times\frac{5}{36}+49\times\frac{6}{36}$

$$+64\times\frac{5}{36}+81\times\frac{4}{36}+100\times\frac{3}{36}+121\times\frac{2}{36}+144\times\frac{1}{36}=\frac{1974}{36}=54.8$$

Thus, Var(X) = 54.8 - 49 = 5.8

(iii)

X	1	2	3
$p(X)$	$\frac{4}{20}$	$\frac{12}{20}$	$\frac{4}{20}$

Solution

From the above, we can write

$$E(X)=1\times\frac{4}{20}+2\times\frac{12}{20}+3\times\frac{4}{20}=2$$

and $$E(X^2)=1\times\frac{4}{20}+4\times\frac{12}{20}+9\times\frac{4}{20}=4.4$$

$\therefore$ Var(X) = 4.4 - 4 = 0.4

1. Expected Monetary Value (EMV)

When a random variable is expressed in monetary units, its expected value is often termed as expected monetary value and symbolized by EMV.

Example 2.9: If it rains, an umbrella salesman earns Rs 100 per day. If it is fair, he loses Rs 15 per day. What is his expectation if the probability of rain is 0.3?

Solution: Here the random variable X takes only two values, X_1 = 100 with probability 0.3 and X_2 = – 15 with probability 0.7.

Thus, the expectation of the umbrella salesman

= 100 × 0.3 – 15 × 0.7 = 19.5

The above result implies that his average earning in the long run would be Rs 19.5 per day.

Example 2.10: A person plays a game of throwing an unbiased die under the condition that he could get as many rupees as the number of points obtained on the die. Find the expectation and variance of his winning. How much should he pay to play in order that it is a fair game?

Solution: The probability distribution of the number of rupees won by the person is given below :

X(*Rs*)	1	2	3	4	5	6
$p(X)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Thus, $$E(X)=1\times\frac{1}{6}+2\times\frac{1}{6}+3\times\frac{1}{6}+4\times\frac{1}{6}+5\times\frac{1}{6}+6\times\frac{1}{6}=Rs\,\frac{7}{2}$$

and $$E(X^2)=1\times\frac{1}{6}+4\times\frac{1}{6}+9\times\frac{1}{6}+16\times\frac{1}{6}+25\times\frac{1}{6}+36\times\frac{1}{6}=\frac{91}{6}$$

$\therefore$ $$\therefore\quad \sigma^2=\frac{91}{6}-\left(\frac{7}{2}\right)^2=\frac{35}{12}=2.82\text{ . Note that the unit of }s^2\text{ will be (Rs)}^2.$$

Since $E(X)$ is positive, the player would win Rs 3.5 per game in the long run. Such a game is said to be favourable to the player. In order that the game is fair, the expectation of the player should be zero. Thus, he should pay Rs 3.5 before the start of the game so that the possible values of the random variable become 1 – 3.5 = – 2.5, 2 – 3.5 = – 1.5, 3 – 3.5 = – 0.5, 4 – 3.5 = 0.5, etc. and their expected value is zero.

2. Expected Value with Perfect Information (EVPI)

Example 2.11: The payoffs (in Rs) of three Acts A_1, A_2 and A_3 and the possible states of nature S_1, S_2 and S_3 are given below :

Acts → / *States of Nature* ↓	A_1	A_2	A_3
S_1	−20	−50	200
S_2	200	−100	−50
S_3	400	600	300

The probabilities of the states of nature are 0.3, 0.4 and 0.3 respectively. Determine the optimal act using the Bayesian Criterion.

Solution:

Computation of Expected Monetary Value

	S_1	S_2	S_3	
$P(S)$	0.3	0.4	0.3	*EMV*
A_1	−20	200	400	$-20 \times 0.3 + 200 \times 0.4 + 400 \times 0.3 = 194$
A_2	−50	−100	600	$-50 \times 0.3 - 100 \times 0.4 + 600 \times 0.3 = 125$
A_3	200	−50	300	$200 \times 0.3 - 50 \times 0.4 + 300 \times 0.3 = 130$

From the above table, we find that the act A_1 is optimal.

The problem can alternatively be attempted by finding minimum EOL, as shown below:

Computation of Expected Opportunity Loss

	S_1	S_2	S_3	
$P(S)$	0.3	0.4	0.3	*EOL*
A_1	220	0	200	$220 \times 0.3 + 0 \times 0.4 + 200 \times 0.3 = 126$
A_2	250	300	0	$250 \times 0.3 + 300 \times 0.4 + 0 \times 0.3 = 195$
A_3	0	250	300	$0 \times 0.3 + 250 \times 0.4 + 300 \times 0.3 = 190$

This indicates that the optimal act is again A_1.

2.7 EXPECTED VALUE WITH PERFECT INFORMATION (EVPI)

The expected value with perfect information is the amount of profit foregone due to uncertain conditions affecting the selection of a course of action.

Given the perfect information, a decision maker is supposed to know which particular state of nature will be in effect. Thus, the procedure for the selection of an optimal course of action, for the decision problem given in example 2.11, will be as follows:

If the decision maker is certain that the state of nature S_1 will be in effect, he would select the course of action A_3, having maximum payoff equal to Rs 200.

Similarly, if the decision maker is certain that the state of nature S_2 will be in effect, his course of action would be A_1 and if he is certain that the state of nature S_3 will be in effect, his course of action would be A_2. The maximum payoffs associated with the actions are Rs 200 and Rs 600 respectively.

The weighted average of these payoffs with weights equal to the probabilities of respective states of nature is termed as Expected Payoff under Certainty (EPC).

Thus, $EPC = 200 \times 0.3 + 200 \times 0.4 + 600 \times 0.3 = 320$

The difference between EPC and EMV of optimal action is the amount of profit foregone due to uncertainty and is equal to EVPI.

Thus, EVPI = EPC – EMV of optimal action = 320 – 194 = 126

It is interesting to note that EVPI is also equal to EOL of the optimal action.

Cost of Uncertainty

This concept is similar to the concept of EVPI. Cost of uncertainty is the difference between the EOL of optimal action and the EOL under perfect information.

Given the perfect information, the decision maker would select action with minimum opportunity loss under each state of nature. Since minimum opportunity loss under each state of nature is zero, therefore,

EOL under certainty $= 0 \times 0.3 + 0 \times 0.4 + 0 \times 0.3 = 0$.

Thus, the cost of uncertainty = EOL of optimal action = EVPI

Example 2.12: A group of students raise money each year by selling souvenirs outside the stadium of a cricket match between teams *A* and *B*. They can buy any of three different types of souvenirs from a supplier. Their sales are mostly dependent on which team wins the match. A conditional payoff (in Rs.) table is as under :

Type of Souvenir →	*I*	*II*	*III*
Team A wins	1200	800	300
Team B wins	250	700	1100

(i) Construct the opportunity loss table.

(ii) Which type of souvenir should the students buy if the probability of team A's winning is 0.6?

(iii) Compute the cost of uncertainty.

Solution:

(i) The Opportunity Loss Table

Actions → / *Events* ↓	*Type of Souvenir bought*		
	I	*II*	*III*
Team A wins	0	400	900
Team B wins	850	400	0

(ii) EOL of buying type I Souvenir $= 0 \times 0.6 + 850 \times 0.4 = 340$

EOL of buying type II Souvenir $= 400 \times 0.6 + 400 \times 0.4 = 400$.

EOL of buying type III Souvenir $= 900 \times 0.6 + 0 \times 0.4 = 540$.

Since the EOL of buying Type I Souvenir is minimum, the optimal decision is to buy Type I Souvenir.

(iii) Cost of uncertainty = EOL of optimal action = Rs. 340

Binomial Distribution

Binomial distribution is a theoretical probability distribution which was given by James Bernoulli. This distribution is applicable to situations with the following characteristics:

1. An experiment consists of a finite number of repeated trials.
2. Each trial has only two possible, mutually exclusive, outcomes which are termed as a 'success' or a 'failure'.
3. The probability of a success, denoted by p, is known and remains constant from trial to trial. The probability of a failure, denoted by q, is equal to $1 - p$.
4. Different trials are independent, i.e., outcome of any trial or sequence of trials has no effect on the outcome of the subsequent trials.

The sequence of trials under the above assumptions is also termed as Bernoulli Trials.

Probability Function or Probability Mass Function

Let n be the total number of repeated trials, p be the probability of a success in a trial and q be the probability of its failure so that $q = 1 - p$.

Let r be a random variable which denotes the number of successes in n trials. The possible values of r are 0, 1, 2, ... n. We are interested in finding the probability of r successes out of n trials, i.e., $P(r)$.

To find this probability, we assume that the first r trials are successes and remaining $n - r$ trials are failures. Since different trials are assumed to be independent, the probability of this sequence is

$$\underbrace{p.p. \ldots . p}_{r \text{ times}} \; \underbrace{q.q. \ldots . q}_{(n-r) \text{ times}} \text{ i.e. } p^r q^{n-r}.$$

Since out of n trials any r trials can be success, the number of sequences showing any r trials as success and remaining $(n - r)$ trials as failure is ${}^nC_r$, where the probability of r successes in each trial is $p^r q^{n-r}$. Hence, the required probability is $P(r) = {}^nC_r p^r q^{n-r}$, where $r = 0, 1, 2, \ldots . n$.

Writing this distribution in a tabular form, we have

r	0	1	2	.....	n	*Total*
$P(r)$	${}^nC_0 p^0 q^n$	${}^nC_1 p\, q^{n-1}$	${}^nC_2 p^2 q^{n-2}$	.....	${}^nC_n p^n q^0$	1

It should be noted here that the probabilities obtained for various values of r are the terms in the binomial expansion of $(q + p)^n$ and thus, the distribution is termed as Binomial Distribution.

$P(r) = {}^nC_r p^r q^{n-r}$ is termed as the probability function or probability mass function (p.m.f.) of the distribution.

Summary Measures of Binomial Distribution

(a) *Mean:* The mean of a binomial variate r, denoted by μ, is equal to $E(r)$, i.e.,

$$\mu = E(r) = \sum_{r=0}^{n} rP(r) = \sum_{r=1}^{n} r.{}^nC_r p^r q^{n-r} \text{ (note that the term for } r = 0 \text{ is 0)}$$

$$= \sum_{r=1}^{n} \frac{r.n!}{r!(n-r)!} . p^r q^{n-r} = \sum_{r=1}^{n} \frac{n.(n-1)!}{(r-1)!(n-r)!} . p^r q^{n-r}$$

$$= np \sum_{r=1}^{n} \frac{(n-1)!}{(r-1)!(n-r)!} . p^{r-1} q^{n-r} = np(q+p)^{n-1} = np \qquad [\because q+p=1]$$

(b) *Variance:* The variance of r, denoted by s^2, is given by

$$\sigma^2 = E\left[r - E(r)\right]^2 = E\left[r - np\right]^2 = E\left[r^2 - 2npr + n^2p^2\right]$$

$$= E(r^2) - 2npE(r) + n^2p^2 = E(r^2) - 2n^2p^2 + n^2p^2$$

$$= E(r^2) - n^2p^2 \qquad \text{.... (1)}$$

Thus, to find σ^2, we first determine $E(r^2)$.

Now, $E(r^2) = \sum_{r=1}^{n} r^2 .{}^nC_r p^r q^{n-r} = \left[r(r-1) + r\right] {}^nC_r p^r q^{n-r}$

$$= \sum_{r=2}^{n} r(r-1) {}^nC_r p^r q^{n-r} + \sum_{r=1}^{n} r.{}^nC_r p^r q^{n-r} = \sum_{r=2}^{n} \frac{r(r-1)n!}{r!(n-r)!} . p^r q^{n-r} + np$$

$$= \sum_{r=2}^{n} \frac{n!}{(r-2)!(n-r)!} . p^r q^{n-r} + np = \sum_{r=2}^{n} \frac{n(n-1).(n-2)!}{(r-2)!(n-r)!} . p^r q^{n-r} + np$$

$$= n(n-1)p^2 \sum_{r=2}^{n} \frac{(n-2)!}{(r-2)!(n-r)!} . p^{r-2} q^{n-r} + np$$

$$= n(n-1)p^2 (q+p)^{n-2} + np = n(n-1)p^2 + np$$

Substituting this value in equation (1), we get

$$\sigma^2 = n(n-1)p^2 + np - n^2p^2 = np(1-p) = npq$$

Or the standard deviation $= \sqrt{npq}$

Remarks: $\sigma^2 = npq = mean \times q$, which shows that $\sigma^2 < mean$, since $0 < q < 1$.

(c) The values of μ_3, μ_4, β_1 and β_2

Proceeding as above, we can obtain

$$\mu_3 = E(r-np)^3 = npq(q-p)$$

$$\mu_4 = E(r-np)^4 = 3n^2p^2q^2 + npq(1-6pq)$$

Also $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3} = \dfrac{n^2p^2q^2(q-p)^2}{n^3p^3q^3} = \dfrac{(q-p)^2}{npq}$

The above result shows that the distribution is symmetrical when

$p = q = \dfrac{1}{2}$, negatively skewed if $q < p$, and positively skewed if $q > p$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3n^2p^2q^2 + npq(1-6pq)}{n^2p^2q^2} = 3 + \frac{(1-6pq)}{npq}$$

The above result shows that the distribution is leptokurtic if $6pq < 1$, platykurtic if $6pq > 1$ and mesokurtic if $6pq = 1$.

(d) ***Mode:*** Mode is that value of the random variable for which probability is maximum.

If r is mode of a binomial distribution, then we have

$P(r-1) \le P(r) \ge P(r+1)$

Consider the inequality $P(r) \ge P(r+1)$

or ${}^nC_r p^r q^{n-r} \ge {}^nC_{r+1} p^{r+1} q^{n-r-1}$

or $\dfrac{n!}{r!(n-r)!} p^r q^{n-r} \ge \dfrac{n!}{(r+1)!(n-r-1)!} p^{r+1} q^{n-r-1}$

or $\dfrac{1}{(n-r)}.q \ge \dfrac{1}{(r+1)}.p$ or $qr + q \ge np - pr$

Solving the above inequality for r, we get

$$r \ge (n+1)p - 1 \quad \text{.... (1)}$$

Similarly, on solving the inequality $P(r-1) \le P(r)$ for r, we can get

$$r \le (n+1)p \quad \text{.... (2)}$$

Combining inequalities (1) and (2), we get

$$(n+1)p-1 \le r \le (n+1)p$$

Case I: When $(n + 1)p$ is not an integer

When $(n + 1)p$ is not an integer, then $(n + 1)\ p - 1$ is also not an integer. Therefore, mode will be an integer between $(n + 1)p - 1$ and $(n + 1)p$ or mode will be an integral part of $(n + 1)p$.

Case II: When $(n + 1)p$ is an integer

When $(n + 1)p$ is an integer, the distribution will be bimodal and the two modal values would be $(n + 1)\ p - 1$ and $(n + 1)p$.

Example 2.13: An unbiased die is tossed three times. Find the probability of obtaining (a) no six, (b) one six, (c) at least one six, (d) two sixes and (e) three sixes.

Solution: The three tosses of a die can be taken as three repeated trials which are independent. Let the occurrence of six be termed as a success. Therefore, r will denote the number of six obtained. Further, $n = 3$ and $p = \frac{1}{6}$.

(a) Probability of obtaining no six, i.e.,

$$P(r=0) = {}^3C_0 p^0 q^3 = 1.\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^3 = \frac{125}{216}$$

(b) $$P(r=1) = {}^3C_1 p^1 q^2 = 3.\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^2 = \frac{25}{72}$$

(c) Probability of getting at least one six = $1 - P(r = 0)$ $$= 1 - \frac{125}{216} = \frac{91}{216}$$

(d) $$P(r=2) = {}^3C_2 p^2 q^1 = 3.\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right) = \frac{5}{72}$$

(e) $$P(r=3) = {}^3C_3 p^3 q^0 = 3.\left(\frac{1}{6}\right)^3 = \frac{1}{216}$$

Example 2.14: Assuming that it is true that 2 in 10 industrial accidents are due to fatigue, find the probability that:

(a) Exactly 2 of 8 industrial accidents will be due to fatigue.

(b) At least 2 of the 8 industrial accidents will be due to fatigue.

Solution: Eight industrial accidents can be regarded as Bernoulli trials each with probability of success $p = \frac{2}{10} = \frac{1}{5}$. The random variable r denotes the number of accidents due to fatigue.

(a) $$P(r=2) = {}^8C_2\left(\frac{1}{5}\right)^2\left(\frac{4}{5}\right)^6 = 0.294$$

(b) We have to find $P(r \ge 2)$. We can write

$P(r \ge 2) = 1 - P(0) - P(1)$, thus, we first find $P(0)$ and $P(1)$.

We have $P(0) = {}^8C_0\left(\frac{1}{5}\right)^0\left(\frac{4}{5}\right)^8 = 0.168$

and $P(1) = {}^8C_1\left(\frac{1}{5}\right)^1\left(\frac{4}{5}\right)^7 = 0.336$

$\therefore \quad P(r \geq 2) = 1 - 0.168 - 0.336 = 0.496$

Check Your Progress 1

Fill in the blanks:

1. A phenomenon or an experiment which can result into more than one possible outcome is called .....
2. The occurrence or non-occurrence of a phenomenon is called an .....
3. When a random variable is expressed in monetary units, its expected value is often termed as .....

2.8 POISSON

This distribution was derived by a noted mathematician, Simon D. Poisson, in 1837. He derived this distribution as a limiting case of binomial distribution, when the number of trials n tends to become very large and the probability of success in a trial p tends to become very small such that their product np remains a constant. This distribution is used as a model to describe the probability distribution of a random variable defined over a unit of time, length or space. For example, the number of telephone calls received per hour at a telephone exchange, the number of accidents in a city per week, the number of defects per meter of cloth, the number of insurance claims per year, the number breakdowns of machines at a factory per day, the number of arrivals of customers at a shop per hour, the number of typing errors per page etc.

Poisson Process

Let us assume that on an average 3 telephone calls are received per 10 minutes at a telephone exchange desk and we want to find the probability of receiving a telephone call in the next 10 minutes. In an effort to apply binomial distribution, we can divide the interval of 10 minutes into 10 intervals of 1 minute each so that the probability of receiving a telephone call (i.e., a success) in each minute (i.e., trial) becomes 3/10 (note that p = m/n, where m denotes mean). Thus, there are 10 trials which are independent, each with probability of success = 3/10. However, the main difficulty with this formulation is that, strictly speaking, these trials are not Bernoulli trials. One essential requirement of such trials, that each trial must result into one of the two possible outcomes, is violated here. In the above example, a trial, i.e. an interval of one minute, may result into 0, 1, 2, ... successes depending upon whether the exchange desk receives none, one, two, ... telephone calls respectively.

One possible way out is to divide the time interval of 10 minutes into a large number of small intervals so that the probability of receiving two or more telephone calls in an interval becomes almost zero. This is illustrated by the following table which shows that the probabilities of receiving two calls decreases sharply as the number of intervals are increased, keeping the average number of calls, 3 calls in 10 minutes in our example, as constant.

n	$P(\text{one call is received})$	$P(\text{two calls are received})$
10	0.3	0.09
100	0.03	0.0009
1,000	0.003	0.000009
10,000	0.0003	0.00000009

Using symbols, we may note that as n increases then p automatically declines in such a way that the mean m (= np) is always equal to a constant. Such a process is termed as a Poisson Process. The chief characteristics of Poisson process can be summarised as given below:

1. The number of occurrences in an interval is independent of the number of occurrences in another interval.
2. The expected number of occurrences in an interval is constant.
3. It is possible to identify a small interval so that the occurrence of more than one event, in any interval of this size, becomes extremely unlikely.

Probability Mass Function of Poisson Distribution

The probability mass function (p.m.f.) of Poisson distribution can be derived as a limit of p.m.f. of binomial distribution when $n \to \infty$ such that m (= np) remains constant. Thus, we can write

$$P(r) = \lim_{n\to\infty} {}^{n}C_{r}\left(\frac{m}{n}\right)^{r}\left(1-\frac{m}{n}\right)^{n-r} = \lim_{n\to\infty} \frac{n!}{r!(n-r)!}\left(\frac{m}{n}\right)^{r}\left(1-\frac{m}{n}\right)^{n-r}$$

$$= \frac{m^{r}}{r!}.\lim_{n\to\infty}\left[n(n-1)(n-2)\ \ldots.(n-r+1)\frac{1}{n^{r}}\left(1-\frac{m}{n}\right)^{n-r}\right.$$

$$= \frac{m^{r}}{r!}.\lim_{n\to\infty}\left[\frac{\frac{n}{n}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots.\left(1-\frac{(r-1)}{n}\right)\left(1-\frac{m}{n}\right)^{n}}{\left(1-\frac{m}{n}\right)^{r}}\right]$$

$$= \frac{m^{r}}{r!}\lim_{n\to\infty}\left(1-\frac{m}{n}\right)^{n}, \text{ since each of the remaining terms will tend to unity as } n\to\infty$$

$$= \frac{m^{r}.e^{-m}}{r!}, \text{ since } \lim_{n\to\infty}\left(1-\frac{m}{n}\right)^{n} = \lim_{n\to\infty}\left\{\left(1-\frac{m}{n}\right)^{\frac{n}{m}}\right\}^{m} = e^{-m}.$$

Thus, the probability mass function of Poisson distribution is

$$P(r) = \frac{e^{-m}.m^{r}}{r!}, \text{ where } r = 0,1,2, \ldots\ldots\infty.$$

Here e is a constant with value = 2.71828... . Note that Poisson distribution is a discrete probability distribution with single parameter m.

$$\text{Total probability } = \sum_{r=0}^{\infty} \frac{e^{-m}.m^r}{r!} = e^{-m}\left(1+\frac{m}{1!}+\frac{m^2}{2!}+\frac{m^3}{3!}+\ ....\right)$$

$$= e^{-m}.e^{m} = 1.$$

Summary Measures of Poisson Distribution

(a) ***Mean:*** The mean of a Poisson variate r is defined as

$$E(r) = \sum_{r=0}^{\infty} r.\frac{e^{-m}.m^r}{r!} = e^{-m}\sum_{r=1}^{\infty}\frac{m^r}{(r-1)!} = e^{-m}\left[m+m^2+\frac{m^3}{2!}+\frac{m^4}{3!}+\ ....\right]$$

$$= me^{-m}\left[1+m\ +\frac{m^2}{2!}+\frac{m^3}{3!}+\ ....\right] = me^{-m}e^{m} = m$$

(b) ***Variance:*** The variance of a Poisson variate is defined as

$\text{Var}(r) = E(r-m)^2 = E(r^2) - m^2$

$$\text{Now } E\left(r^2\right) = \sum_{r=0}^{\infty} r^2 P(r) = \sum_{r=0}^{\infty}\left[r(r-1)+r\right]P(r) = \sum_{r=0}^{\infty}\left[r(r-1)\right]P(r)+\sum_{r=0}^{\infty} r\,P(r)$$

$$= \sum_{r=2}^{\infty}\left[r(r-1)\right]\frac{e^{-m}.m^r}{r!}+m = e^{-m}\sum_{r=2}^{\infty}\frac{m^r}{(r-2)!}+m$$

$$= m+e^{-m}\left(m^2+m^3+\frac{m^4}{2!}+\frac{m^5}{3!}+\ ....\right)$$

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

Thus, $\text{Var}(r) = m + m^2 - m^2 = m$.

Also standard deviation $\sigma = \sqrt{m}$.

(c) The values of β_3, β_4, β_1 and β_2

It can be shown that $\mu_3 = m$ and $\mu_4 = m + 3m^2$.

$$\therefore\ \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{m^2}{m^3} = \frac{1}{m}$$

Since m is a positive quantity, therefore, β_1 is always positive and hence the Poisson distribution is always positively skewed. We note that $\beta_1 \to 0$ as $m \to \infty$, therefore the distribution tends to become more and more symmetrical for large values of m.

Further, $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{m+3m^2}{m^2} = 3 + \frac{1}{m} \to 3$ as $m \to \infty$. This result shows that the distribution becomes normal for large values of m.

(d) ***Mode:*** As in binomial distribution, a Poisson variate r will be mode if

$$P(r-1) \le P(r) \ge P(r+1)$$

The inequality $P(r-1) \le P(r)$ can be written as

$$\frac{e^{-m}.m^{r-1}}{(r-1)!} \le \frac{e^{-m}.m^{r}}{r!} \Rightarrow 1 \le \frac{m}{r} \Rightarrow r \le m \quad \ldots (1)$$

Similarly, the inequality $P(r) \ge P(r+1)$ can be shown to imply that

$$r \ge m - 1 \quad \ldots (2)$$

Combining (1) and (2), we can write $m - 1 \le r \le m$.

Case I: When m is not an integer

The integral part of m will be mode.

Case II: When m is an integer

The distribution is bimodal with values m and $m - 1$.

Example 2.15: The average number of customer arrivals per minute at a super bazaar is 2. Find the probability that during one particular minute (i) exactly 3 customers will arrive, (ii) at the most two customers will arrive, (iii) at least one customer will arrive.

Solution: It is given that $m = 2$. Let the number of arrivals per minute be denoted by the random variable r. The required probability is given by

(i) $$P(r=3) = \frac{e^{-2}.2^3}{3!} = \frac{0.13534 \times 8}{6} = 0.18045$$

(ii) $$P(r \le 2) = \sum_{r=0}^{2} \frac{e^{-2}.2^r}{r!} = e^{-2}\left[1 + 2 + \frac{4}{2}\right] = 0.13534 \times 5 = 0.6767.$$

(iii) $$P(r \ge 1) = 1 - P(r=0) = 1 - \frac{e^{-2}.2^0}{0!} = 1 - 0.13534 = 0.86464.$$

Example 2.16: An executive makes, on an average, 5 telephone calls per hour at a cost which may be taken as Rs 2 per call. Determine the probability that in any hour the telephone calls' cost (i) exceeds Rs 6, (ii) remains less than Rs 10.

Solution: The number of telephone calls per hour is a random variable with mean = 5. The required probability is given by

(i) $$P(r>3)=1-P(r\le 3)=1-\sum_{r=0}^{3}\frac{e^{-5}.5^r}{r!}$$

$$=1-e^{-5}\left[1+5+\frac{25}{2}+\frac{125}{6}\right]=1-0.00678\times\frac{236}{6}=0.7349.$$

(ii) $$P(r\le 4)=\sum_{r=0}^{4}\frac{e^{-5}.5^r}{r!}=e^{-5}\left[1+5+\frac{25}{2}+\frac{125}{6}+\frac{625}{24}\right]=0.00678\times\frac{1569}{24}=0.44324.$$

2.9 HYPERGEOMETRIC DISTRIBUTION

The binomial distribution is not applicable when the probability of a success p does not remain constant from trial to trial. In such a situation the probabilities of the various values of r are obtained by the use of Hypergeometric distribution.

Let there be a finite population of size N, where each item can be classified as either a success or a failure. Let there be k successes in the population. If a random sample of size n is taken from this population, then the probability of r successes is given by $P(r)=\frac{\left({}^{k}C_{r}\right)\left({}^{N-k}C_{n-r}\right)}{{}^{N}C_{n}}$. Here r is a discrete random variable which can take values 0, 1, 2, ... n. Also $n \le k$.

It can be shown that the mean of r is np and its variance is

$$\left(\frac{N-n}{N-1}\right)npq, \text{ where } p=\frac{k}{N} \text{ and } q = 1 - p.$$

Example 2.17: A retailer has 10 identical television sets of a company out which 4 are defective. If 3 televisions are selected at random, construct the probability distribution of the number of defective television sets.

Solution: Let the random variable r denote the number of defective televisions. In terms of notations, we can write $N = 10$, $k = 4$ and $n = 3$.

Thus, we can write $$P(r)=\frac{{}^{4}C_{r}\times{}^{6}C_{3-r}}{{}^{10}C_{3}}, \quad r=0,1,2,3$$

The distribution of r is hypergeometric. This distribution can also be written in a tabular form as given below :

r	0	1	2	3	*Total*
$P(r)$	$\frac{5}{30}$	$\frac{15}{30}$	$\frac{9}{30}$	$\frac{1}{30}$	1

Binomial Approximation to Hypergeometric Distribution

In sampling problems, where sample size n (total number of trials) is less than 5% of population size N, i.e., $n < 0.05N$, the use of binomial distribution will also give satisfactory results. The reason for this is that the smaller the sample size relative to population size, the greater will be the validity of the requirements of independent trials and the constancy of p.

2.10 NORMAL DISTRIBUTION

The normal probability distribution occupies a place of central importance in *Modern Statistical Theory*. This distribution was first observed as the *normal law of errors* by the statisticians of the eighteenth century. They found that each observation X involves an error term which is affected by a large number of small but independent chance factors. This implies that an observed value of X is the sum of its true value and the net effect of a large number of independent errors which may be positive or negative each with equal probability. The observed distribution of such a random variable was found to be in close conformity with a continuous curve, which was termed as the *normal curve of errors* or simply the *normal curve*.

Since Gauss used this curve to describe the theory of accidental errors of measurements involved in the calculation of orbits of heavenly bodies, it is also called as Gaussian curve.

Conditions of Normality

In order that the distribution of a random variable X is normal, the factors affecting its observations must satisfy the following conditions:

(i) *A large number of chance factors:* The factors, affecting the observations of a random variable, should be numerous and equally probable so that the occurrence or non-occurrence of any one of them is not predictable.

(ii) *Condition of homogeneity:* The factors must be similar over the relevant population although, their incidence may vary from observation to observation.

(iii) *Condition of independence:* The factors, affecting observations, must act independently of each other.

(iv) *Condition of symmetry:* Various factors operate in such a way that the deviations of observations above and below mean are balanced with regard to their magnitude as well as their number.

Random variables observed in many phenomena related to economics, business and other social as well as physical sciences are often found to be distributed normally. For example, observations relating to the life of an electrical component, weight of packages, height of persons, income of the inhabitants of certain area, diameter of wire, etc., are affected by a large number of factors and hence, tend to follow a pattern that is very similar to the normal curve. In addition to this, when the number of observations become large, a number of probability distributions like Binomial, Poisson, etc., can also be approximated by this distribution.

Probability Density Function

If X is a continuous random variable, distributed normally with mean m and standard deviation σ, then its p.d.f. is given by

$$p(X) = \frac{1}{\sigma\sqrt{2\pi}} . e^{\frac{-1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} \quad \text{where} \quad -\infty < X < \infty.$$

Here π and σ are absolute constants with values 3.14159.... and 2.71828.... respectively.

It may be noted here that this distribution is completely known if the values of mean m and standard deviation σ are known. Thus, the distribution has two parameters, viz. mean and standard deviation.

Shape of Normal Probability Curve

For given values of the parameters, m and s, the shape of the curve corresponding to normal probability density function $p(X)$ is as shown in Figure. 2.1

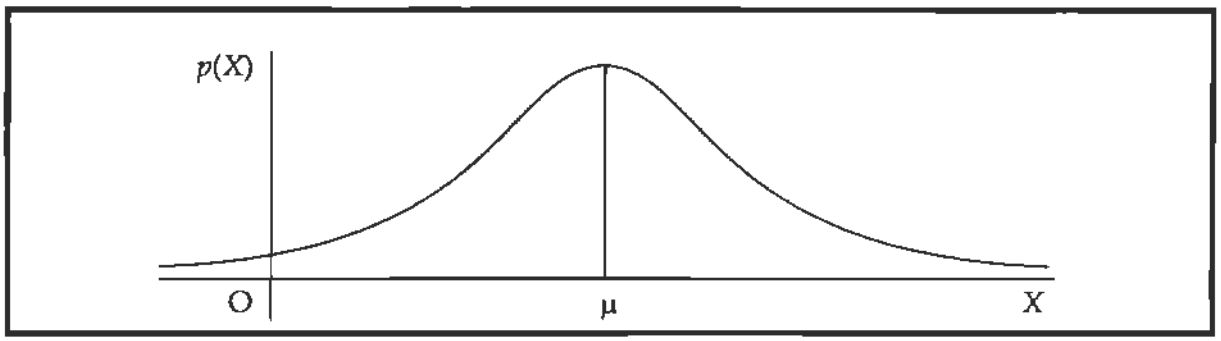


Figure 2.1

Normal Probability Curve

It should be noted here that although we seldom encounter variables that have a range from - ∞ to ∞, as shown by the normal curve, nevertheless the curves generated by the relative frequency histograms of various variables closely resembles the shape of normal curve.

Properties of Normal Probability Curve

A normal probability curve or normal curve has the following properties :

1. It is a bell shaped symmetrical curve about the ordinate at $X = \mu$. The ordinate is maximum at $X = \mu$.
2. It is unimodal curve and its tails extend infinitely in both directions, i.e., the curve is asymptotic to X axis in both directions.
3. All the three measures of central tendency coincide, i.e.,

   mean = median = mode
4. The total area under the curve gives the total probability of the random variable taking values between $-\infty$ to ∞. Mathematically, it can be shown that

$$P(-\infty < X < \infty) = \int_{-\infty}^{\infty} p(X)\, dX = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} dX = 1.$$

5. Since median = m, the ordinate at $X = \mu$ divides the area under the normal curve into two equal parts, i.e.,

$$\int_{-\infty}^{\mu} p(X)dX = \int_{\mu}^{\infty} p(X)dX = 0.5$$

6. The value of $p(X)$ is always non-negative for all values of X, i.e., the whole curve lies above X axis.

7. The points of inflexion (the point at which curvature changes) of the curve are at $X = \mu \pm \sigma$.

8. The quartiles are equidistant from median, i.e., $M_d - Q_1 = Q_3 - M_d$, by virtue of symmetry. Also $Q_1 = \mu - 0.6745\sigma$, $Q_3 = \mu + 0.6745\sigma$, quartile deviation = $\sigma 0.6745$ and mean deviation = 0.8σ, approximately.

9. Since the distribution is symmetrical, all odd ordered central moments are zero.

10. The successive even ordered central moments are related according to the following recurrence formula

$\mu_{2n} = (2n - 1)\sigma^2\mu_{2n-2}$ for = 1, 2, 3, ...

11. The value of moment coefficient of skewness β_1 is zero.

12. The coefficient of kurtosis $\beta_2 = \dfrac{\mu_4}{\mu_2^2} = \dfrac{3\sigma^4}{\sigma^4} = 3.$

Note that the above expression makes use of property 10.

13. Additive or reproductive property

If $X_1, X_2, \ldots\ldots X_n$ are n independent normal variates with means $\mu_1, \mu_2, \ldots\ldots \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \ldots\ldots \sigma_n^2$, respectively, then their linear combination $a_1X_1 + a_2X_2 + \ldots\ldots + a_nX_n$ is also a normal variate with mean $\sum_{i=1}^{n} a_i\mu_i$ and variance $\sum_{i=1}^{n} a_i^2\sigma_i^2$.

In particular, if $a_1 = a_2 = \ldots\ldots = a_n = 1$, we have $\sum X_i$ is a normal variate with mean $\sum \mu_i$ and variance $\sum \sigma_i^2$. Thus the sum of independent normal variates is also a normal variate.

14. *Area property:* The area under the normal curve is distributed by its standard deviation in the following manner :

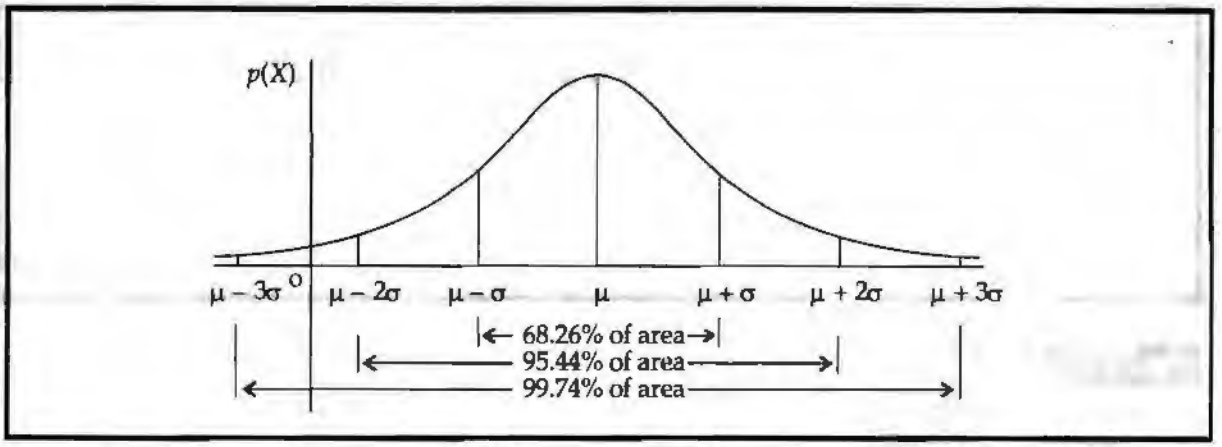


Figure 2.2

Area under the normal curve

(i) The area between the ordinates at $\mu - \sigma$ and $\mu + \sigma$ is 0.6826. This implies that for a normal distribution about 68% of the observations will lie between $\mu - \sigma$ and $\mu + \sigma$.

(ii) The area between the ordinates at $\mu - 2\sigma$ and $\mu + 2\sigma$ is 0.9544. This implies that for a normal distribution about 95% of the observations will lie between $\mu - 2\sigma$ and $\mu + 2\sigma$.

(iii) The area between the ordinates at $\mu - 3\sigma$ and $\mu + 3\sigma$ is 0.9974. This implies that for a normal distribution about 99% of the observations will lie between $\mu - 3\sigma$ and $\mu + 3\sigma$. This result shows that, practically, the range of the distribution is 6σ although, theoretically, the range is from $-\infty$ to ∞.

Probability of Normal Variate in an Interval

Let X be a normal variate distributed with mean μ and standard deviation σ, also written in abbreviated form as $X \sim N(\mu, \sigma)$ The probability of X lying in the interval (X_1, X_2) is given by

$$P(X_1 \le X \le X_2) = \int_{X_1}^{X_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} dX$$

In terms of figure, this probability is equal to the area under the normal curve between the ordinates at $X = X_1$ and $X = X_2$ respectively.

Note: It may be recalled that the probability that a continuous random variable takes a particular value is defined to be zero even though the event is not impossible.

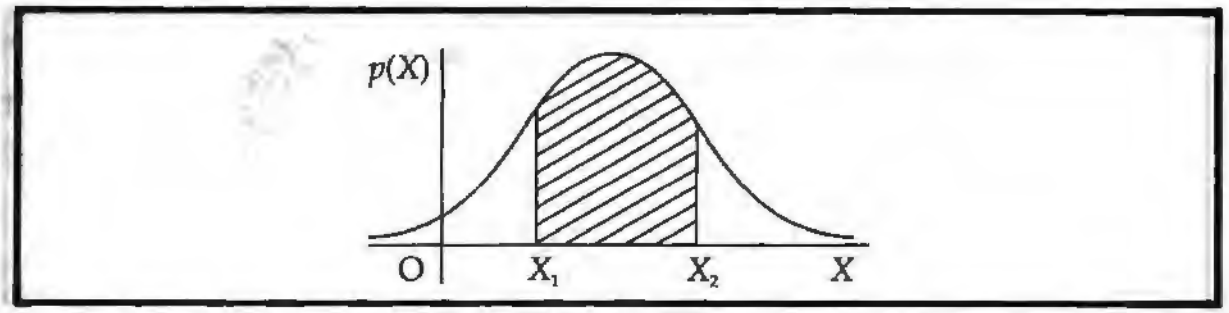


Figure 2.3

Area under the normal curve between the ordinates at $X = X_1$ and $X = X_2$

It is obvious from the above that, to find $P(X_1 \le X \le X_2)$, we have to evaluate an integral which might be cumbersome and time consuming task. Fortunately, an alternative procedure is available for performing this task. To devise this procedure, we define a new variable $z = \frac{X-\mu}{\sigma}$.

We note that $E(z) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}\left[E(X) - \mu\right] = 0$

and $Var(z) = Var\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2} Var(X-\mu) = \frac{1}{\sigma^2} Var(X) = 1.$

Further, from the reproductive property, it follows that the distribution of z is also normal.

Thus, we conclude that if X is a normal variate with mean m and standard deviation σ, then $z = \frac{X-\mu}{\sigma}$ is a normal variate with mean zero and standard deviation unity. Since the parameters of the distribution of z are fixed, it is a known distribution and is termed as *standard normal distribution* (s.n.d.). Further, z is termed as a *standard normal variate* (s.n.v.).

It is obvious from the above that the distribution of any normal variate X can always be transformed into the distribution of standard normal variate z. This fact can be utilised to evaluate the integral given above.

We can write $P(X_1 \le X \le X_2) = P\left[\left(\frac{X_1-\mu}{\sigma}\right) \le \left(\frac{X-\mu}{\sigma}\right) \le \left(\frac{X_2-\mu}{\sigma}\right)\right]$

$= P(z_1 \le z \le z_2)$, where $z_1 = \frac{X_1-\mu}{\sigma}$ and $z_2 = \frac{X_2-\mu}{\sigma}$

In terms of figure, this probability is equal to the area under the standard normal curve between the ordinates at $z = z_1$ and $z = z_2$. Since the distribution of z is fixed, the probabilities of z lying in various intervals are tabulated. These tables can be used to write down the desired probability.

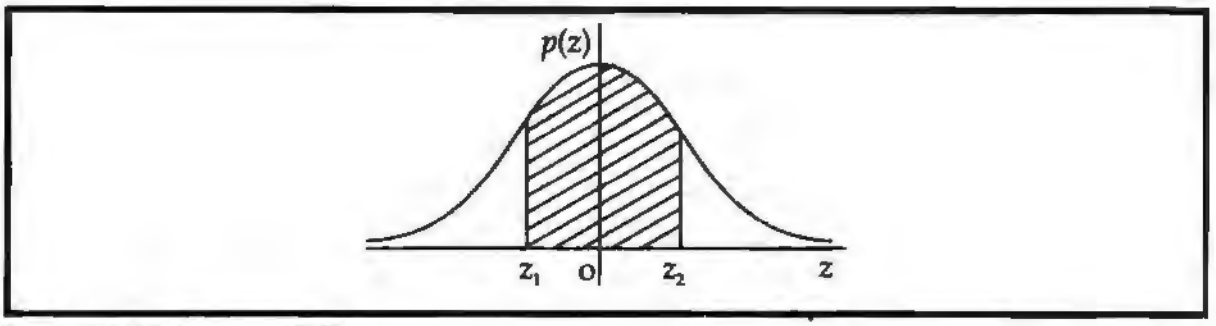


Figure 2.4

Area under the standard normal curve between the ordinates at $z = z_1$ and $z = z_2$

Example 2.18: Using the table of areas under the standard normal curve, find the following probabilities:

(i) $P(0 \le z \le 1.3)$ (ii) $P(-1 \le z \le 0)$ (iii) $P(-1 \le z \le 12)$

(iv) $P(z \ge 1.54)$ (v) $P(|z| > 2)$ (vi) $P(|z| < 2)$

Solution: The required probability, in each question, is indicated by the shaded are of the corresponding figure.

(i) From the table, we can write $P(0 \le z \le 1.3) = 0.4032$.

(ii) We can write $P(-1 \le z \le 0) = P(0 \le z \le 1)$, because the distribution is symmetrical.

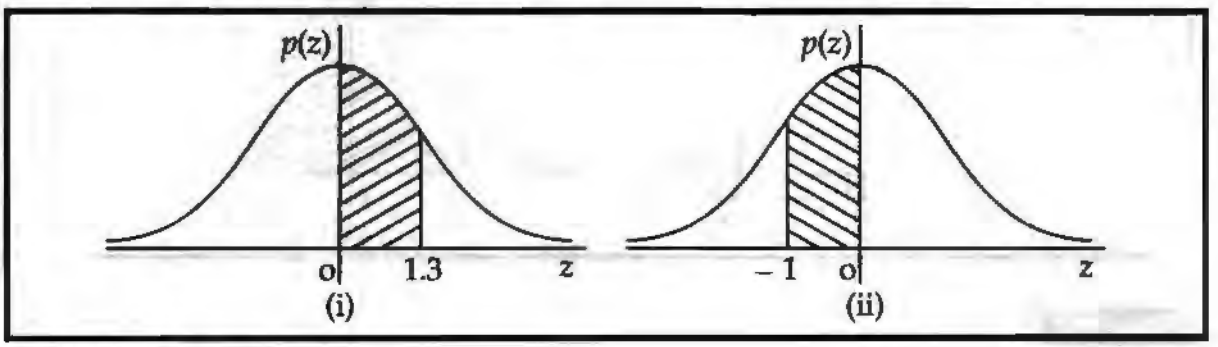


Figure 2.5

Areas under the standard normal curve

From the table, we can write $P(-1 \le z \le 0) = P(0 \le z \le 1) = 0.3413$.

(iii) We can write $P(-1 \le z \le 2) = P(-1 \le z \le 0) + P(0 \le z \le 2)$

$= P(0 \le z \le 1) + P(0 \le z \le 2) = 0.3413 + 0.4772$

$= 0.8185$.

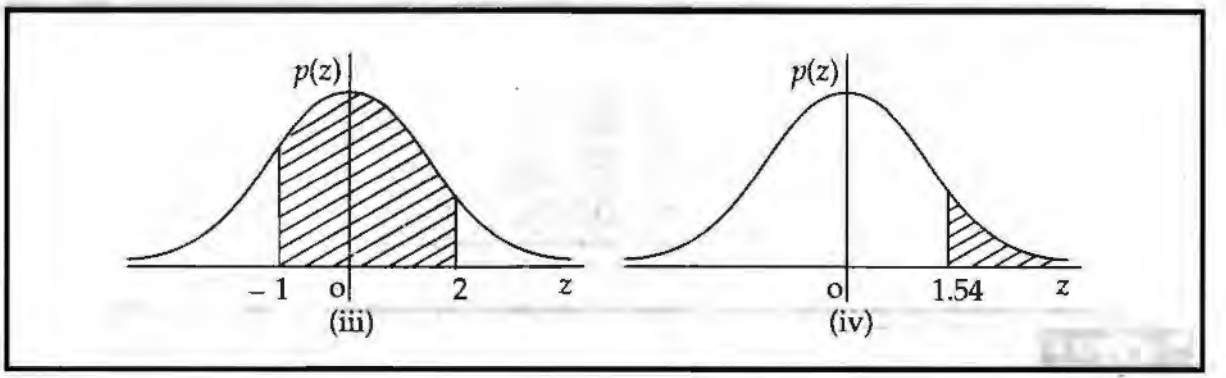


Figure 2.6

Areas under the standard normal curve

(iv) We can write

$P(z \geq 1.54) = 0.5000 - P(0 \leq z \leq 1.54) = 0.5000 - 0.4382 = 0.0618.$

(v) $P(|z| > 2) = P(z > 2) + P(z < -2) = 2P(z > 2) = 2[0.5000 - P(0 \leq z \leq 2)]$

$= 1 - 2P(0 \leq z \leq 2) = 1 - 2 \times 0.4772 = 0.0456.$

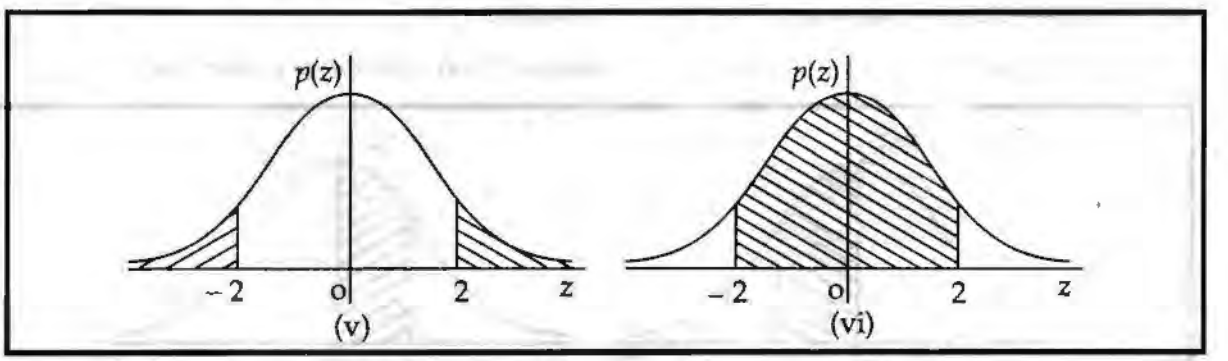


Figure 2.7

Areas under the standard normal curve

(vi) $P(|z| < 2) = P(-2 \leq z \leq 0) + P(0 \leq z \leq 2) = 2P(0 \leq z \leq 2) = 2 \times 0.4772 = 0.9544.$

Example 2.19: Determine the value or values of z in each of the following situations:

(a) Area between 0 and z is 0.4495.

(b) Area between $-\infty$ to z is 0.1401.

(c) Area between $-\infty$ to z is 0.6103.

(d) Area between -1.65 and z is 0.0173.

(e) Area between -0.5 and z is 0.5376.

Solution:

(a) On locating the value of z corresponding to an entry of area 0.4495 in the table of areas under the normal curve, we have $z = 1.64$. We note that the same situation may correspond to a negative value of z. Thus, z can be 1.64 or – 1.64.

(b) Since the area between $-\infty$ to $z < 0.5$, z will be negative. Further, the area between z and 0 = 0.5000 – 0.1401 = 0.3599. On locating the value of z corresponding to this entry in the table, we get $z = -1.08$.

(c) Since the area between $-\infty$ to $z > 0.5000$, z will be positive. Further, the area between 0 to z = 0.6103 – 0.5000 = 0.1103. On locating the value of z corresponding to this entry in the table, we get $z = 0.28$.

(d) Since the area between –1.65 and $z <$ the area between –1.65 and 0 (which, from table, is 0.4505), z is negative. Further z can be to the right or to the left of the value –1.65. Thus, when z lies to the right of –1.65, its value, corresponds to an area (0.4505 – 0.0173) = 0.4332, is given by $z = -1.5$ (from table). Further, when z lies to the left of – 1.65, its value, corresponds to an area (0.4505 + 0.0173) = 0.4678, is given by $z = -1.85$ (from table).

(e) Since the area between –0.5 to $z >$ area between –0.5 to 0 (which, from table, is 0.1915), z is positive. The value of z, located corresponding to an area (0.5376 – 0.1915) = 0.3461, is given by 1.02.

2.11 JOINT PROBABILITY DISTRIBUTION

When two or more random variables X and Y are studied simultaneously on a sample space, we get a joint probability distribution. Consider the experiment of throwing two unbiased dice. If X denotes the number on the first and Y denotes the number on the second die, then X and Y are random variables having a joint probability distribution. When the number of random variables is two, it is called a bi-variate probability distribution and if the number of random variables become more than two, the distribution is termed as a multivariate probability distribution.

Let the random variable X take values $X_1, X_2, \ldots\ldots X_m$ and Y take values $Y_1, Y_2, \ldots\ldots Y_n$. Further, let p_{ij} be the joint probability that X takes the value X_i and Y takes the value Y_j, i.e., $P[X = X_i \text{ and } Y = Y_j] = p_{ij}$ (i = 1 to m and j = 1 to n). This hi-variate probability distribution can be written in a tabular form as follows:

	Y_1	Y_2	...	...	Y_n	*Marginal Probabilities of X*
X_1	p_{11}	p_{12}	...	...	p_{1n}	P_1
X_2	p_{21}	p_{22}	...	...	p_{2n}	P_2
.	.	.	...	...	.	.
.	.	.	...	...	.	.
X_m	p_{m1}	p_{m2}	...	...	p_{mn}	P_m
Marginal Probabilities of Y	P_1'	P_2'	...	...	P_n'	1

Marginal Probability Distribution

In the above table, the probabilities given in each row are added and shown in the last column. Similarly, the sum of probabilities of each column are shown in the last row of the table. These probabilities are termed as marginal probabilities. The last column of the table gives the marginal probabilities for various values of random variable X. The set of all possible values of the random variable X along with their respective marginal probabilities is termed as the marginal probability distribution of X. Similarly, the marginal probabilities of the random variable Y are given in the last row of the above table.

Remarks: If X and Y are independent random variables, by multiplication theorem of probability we have

$P(X = X_i \text{ and } Y = Y_j) = P(X = X_i).P(Y = Y_j) \quad \forall$ i and j

Using notations, we can write $p_{ij} = P_i.P'_j$

The above relation is similar to the relation between the relative frequencies of independent attributes.

Conditional Probability Distribution

Each column of the above table gives the probabilities for various values of the random variable X for a given value of Y, represented by it. For example, column 1 of the table represents that $P(X_1,Y_1) = p_{11}$, $P(X_2,Y_1) = p_{21}$, ... $P(X_m,Y_1) = p_{m1}$, where $P(X_i,Y_1) = p_{i1}$ denote the probability of the event that $X = X_i$ (i = 1 to m) and $Y = Y_1$. From the conditional probability theorem, we can write

$$P(X = X_i / Y = Y_1) = \frac{\textit{Joint probability of } X_i \textit{ and } Y_1}{\textit{Marginal probability of } Y_1} = \frac{p_{ij}}{P'_j} \text{ (for i = 1, 2, ... m).}$$

This gives us a conditional probability distribution of X given that $Y = Y_1$. This distribution can be written in a tabular form as shown below :

X	X_1	X_2	... ...	X_m	*Total Probability*
Probability	$\frac{p_{11}}{P'_1}$	$\frac{p_{21}}{P'_1}$	... ...	$\frac{p_{m1}}{P'_1}$	1

The conditional distribution of X given some other value of Y can be constructed in a similar way. Further, we can construct the conditional distributions of Y for various given values of X.

Remarks:

It can be shown that if the conditional distribution of a random variable is same as its marginal distribution, the two random variables are independent. Thus, if for the conditional distribution of X given Y_1 we have $\frac{p_i}{P'_1} = P_i$ for $\forall$ i, then X and Y are independent. It should be noted here that if one conditional distribution satisfies the condition of independence of the random variables, then all the conditional distributions would also satisfy this condition.

Example 2.20

Let two unbiased dice be tossed. Let a random variable X take the value 1 if first die shows 1 or 2, value 2 if first die shows 3 or 4 and value 3 if first die shows 5 or 6. Further, Let Y be a random variable which denotes the number obtained on the second die. Construct a joint probability distribution of X and Y. Also determine their marginal probability distributions and find E(X) and E(Y) respectively. Determine the conditional distribution of X given Y = 5 and of Y given X = 2. Find the expected values of these conditional distributions. Determine whether X and Y are independent?

Solution

For the given random experiment, the random variable X takes values 1, 2 and 3 and the random variable Y takes values 1, 2, 3, 4, 5 and 6. Their joint probability distribution is shown in the following table:

$X \downarrow \backslash Y \rightarrow$	1	2	3	4	5	6	*Marginal Dist. of X*
1	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{3}$
2	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{3}$
3	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{3}$
Marginal Dist. of Y	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

From the above table, we can write the marginal distribution of X as given below :

X	1	2	3	*Total*
P_i	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1

Thus, the expected value of X is $E(X) = 1.\frac{1}{3} + 2.\frac{1}{3} + 3.\frac{1}{3} = 2$

Similarly, the probability distribution of Y is

Y	1	2	3	4	5	6	*Total*
P_j'	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

and $E(Y) = 1.\frac{1}{6} + 2.\frac{1}{6} + 3.\frac{1}{6} + 4.\frac{1}{6} + 5.\frac{1}{6} + 6.\frac{1}{6} = \frac{21}{6} = 3.5$

The conditional distribution of X when Y = 5 is

X	1	2	3	*Total*
$P_i / Y = 5$	$\frac{1}{18} \times \frac{6}{1} = \frac{1}{3}$	$\frac{1}{18} \times \frac{6}{1} = \frac{1}{3}$	$\frac{1}{18} \times \frac{6}{1} = \frac{1}{3}$	1

$$\therefore \quad E(X/Y=5)=\frac{1}{3}(1+2+3)=2$$

The conditional distribution of Y when X = 2 is

Y	1	2	3	4	5	6	*Total*
$P_j'/X=2$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

$$\therefore \quad E(Y/X=2)=\frac{1}{6}(1+2+3+4+5+6)=3.5$$

Since the conditional distribution of X is same as its marginal distribution (or equivalently the conditional distribution of Y is same as its marginal distribution), X and Y are independent random variables.

Example 2.21

Two unbiased coins are tossed. Let X be a random variable which denotes the total number of heads obtained on a toss and Y be a random variable which takes a value 1 if head occurs on first coin and takes a value 0 if tail occurs on it. Construct the joint probability distribution of X and Y. Find the conditional distribution of X when Y = 0. Are X and Y independent random variables?

Solution

There are 4 elements in the sample space of the random experiment. The possible values that X can take are 0, 1 and 2 and the possible values of Y are 0 and 1. The joint probability distribution of X and Y can be written in a tabular form as follows:

$X\downarrow \backslash Y\rightarrow$	0	1	*Total*
0	$\frac{1}{4}$	0	$\frac{1}{4}$
1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{2}{4}$
2	0	$\frac{1}{4}$	$\frac{1}{4}$
Total	$\frac{2}{4}$	$\frac{2}{4}$	1

The conditional distribution of X when Y = 0, is given by

X	0	1	2	*Total*
$P(X/Y=0)$	$\frac{1}{2}$	$\frac{1}{2}$	0	1

Also, the marginal distribution of X, is given by

X	0	1	2	*Total*
P_i	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

Since the conditional and the marginal distributions are different, X and Y are not independent random variables.

Expectation of the Sum or Product of two Random Variables

Theorem 1

If X and Y are two random variables, then E(X + Y) = E(X) + E(Y).

Proof

Let the random variable X takes values $X_1, X_2, \ldots\ldots X_m$ and the random variable Y takes values $Y_1, Y_2, \ldots\ldots Y_n$ such that $P(X = X_i \text{ and } Y = Y_j) = p_{ij}$ (i = 1 to m, j = 1 to n).

By definition of expectation, we can write

$$E(X+Y)=\sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_i+Y_j\right)p_{ij}=\sum_{i=1}^{m}\sum_{j=1}^{n}X_i p_{ij}+\sum_{i=1}^{m}\sum_{j=1}^{n}Y_j p_{ij}=\sum_{i=1}^{m}X_i\sum_{j=1}^{n}p_{ij}+\sum_{i=1}^{n}Y_j\sum_{j=1}^{m}p_{ij}$$

$$=\sum_{i=1}^{m}X_iP_i+\sum_{j=1}^{n}Y_jP_j' \qquad \left(\text{Here } \sum_{j=1}^{n}p_{ij}=P_i \text{ and } \sum_{i=1}^{m}p_{ij}=P_j'\right)$$

$$=E(X)+E(Y)$$

The above result can be generalised. If there are k random variables $X_1, X_2, \ldots\ldots X_k$, then $E(X_1 + X_2 + \ldots\ldots + X_k) = E(X_1) + E(X_2) + \ldots\ldots E(X_k)$.

Remarks: The above result holds irrespective of whether $X_1, X_2, \ldots\ldots X_k$ are independent or not.

Theorem 2

If X and Y are two independent random variables, then

E(X.Y) = E(X).E(Y)

Proof

Let the random variable X takes values $X_1, X_2, \ldots\ldots X_m$ and the random variable Y takes values $Y_1, Y_2, \ldots\ldots Y_n$ such that $P(X = X_i \text{ and } Y = Y_j) = p_{ij}$ (i = 1 to m, j = 1 to n).

By definition $E(XY)=\sum_{i=1}^{m}\sum_{j=1}^{n}X_iY_jp_{ij}$

Since X and Y are independent, we have $p_{ij} = P_i.P_j'$

$$\therefore \quad E(XY)=\sum_{i=1}^{m}\sum_{j=1}^{n}X_iY_jP_i.P_j'=\sum_{i=1}^{m}X_iP_i\times\sum_{j=1}^{n}Y_jP_j'$$

$$= E(X).E(Y).$$

The above result can be generalised. If there are k independent random variables $X_1, X_2, \ldots\ldots X_k$, then

$E(X_1. X_2. \ldots\ldots X_k) = E(X_1).E(X_2). \ldots\ldots E(X_k)$

Expectation of a Function of Random Variables

Let $f(X,Y)$ be a function of two random variables X and Y. Then we can write

$$E[\phi(X,Y)] = \sum_{i=1}^{m}\sum_{j=1}^{n}\phi(X_i,Y_j)p_{ij}$$

Expression for Covariance

As a particular case, assume that $\phi(X_i,Y_j) = (X_i - \mu_X)(Y_j - \mu_Y)$, where $E(X) = \mu_X$ and $E(Y) = \mu_Y$

Thus, $E[(X - \mu_X)(Y - \mu_Y)] = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_i - \mu_X)(Y_j - \mu_Y)p_{ij}$

The above expression, which is the mean of the product of deviations of values from their respective means, is known as the Covariance of X and Y denoted as Cov(X, Y) or σ_{XY}. Thus, we can write

$$Cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

An alternative expression of Cov(X, Y)

$$Cov(X,Y) = E[\{X - E(X)\}\{Y - E(Y)\}]$$

$$= E[X.\{Y - E(Y)\} - E(X).\{Y - E(Y)\}]$$

$$= E[X.Y - X.E(Y)] = E(X.Y) - E(X).E(Y)$$

Note that E[{Y – E(Y)}] = 0, the sum of deviations of values from their arithmetic mean.

Remarks:

1. If X and Y are independent random variables, the right hand side of the above equation will be zero. Thus, covariance between independent variables is always equal to zero.
2. COV(a + bX, c + dY) = bd COV(X, Y)
3. COV(X, X) = VAR(X)

Mean and Variance of a Linear Combination

Let $Z = \phi(X,Y) = aX + bY$ be a linear combination of the two random variables X and Y, then using the theorem of addition of expectation, we can write

$$\mu_Z = E(Z) = E(aX + bY) = aE(X) + bE(Y) = a\mu_X + b\mu_Y$$

Further, the variance of Z is given by

$$\sigma_Z^2 = E[Z - E(Z)]^2 = E[aX + bY - a\mu_X - b\mu_Y]^2 = E[a(X - \mu_X) + b(Y - \mu_Y)]^2$$

$$= a^2E(X-\mu_X)^2 + b^2E(Y-\mu_Y)^2 + 2abE(X-\mu_X)(Y-\mu_Y)$$

$$= a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}$$

Remarks:

1. The above results indicate that any function of random variables is also a random variable.
2. If X and Y are independent, then $\sigma_{XY} = 0$, $\therefore \quad \sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2$
3. If Z = aX - bY, then we can write $\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 - 2ab\sigma_{XY}$. However, $\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2$, if X and Y are independent.
4. The above results can be generalised. If $X_1, X_2, \ldots X_k$ are k independent random variables with means $\mu_1, \mu_2, \ldots \mu_k$ and variances $\sigma_1^2, \sigma_2^2, \ldots \sigma_k^2$ respectively, then

$$E(X_1 \pm X_2 \pm \ldots \pm X_k) = \mu_1 \pm \mu_2 \pm \ldots \pm \mu_k$$

and $$Var(X_1 \pm X_2 \pm \ldots \pm X_k) = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_k^2$$

Notes:

1. The general result on expectation of the sum or difference will hold even if the random variables are not independent.
2. The above result can also be proved for continuous random variables.

Example 2.22

A random variable X has the following probability distribution :

X	:	-2	-1	0	1	2
Probability	:	$\frac{1}{6}$	p	$\frac{1}{4}$	p	$\frac{1}{6}$

(i) Find the value of p.

(ii) Calculate E(X + 2) and $E(2X^2 + 3X + 5)$.

Solution

Since the total probability under a probability distribution is equal to unity, the value of p should be such that $\frac{1}{6} + p + \frac{1}{4} + p + \frac{1}{6} = 1$.

This condition gives $p = \frac{5}{24}$

Further, $$E(X) = -2.\frac{1}{6} - 1.\frac{5}{24} + 0.\frac{1}{4} + 1.\frac{5}{24} + 2.\frac{1}{6} = 0,$$

$$E(X^2) = 4.\frac{1}{6} + 1.\frac{5}{24} + 0.\frac{1}{4} + 1.\frac{5}{24} + 4.\frac{1}{6} = \frac{7}{4},$$

$$E(X+2) = E(X) + 2 = 0 + 2 = 2$$

and $$E(2X^2 + 3X + 5) = 2E(X^2) + 3E(X) + 5 = 2.\frac{7}{4} + 0 + 5 = 8.5$$

Example 2.23

A dealer of ceiling fans has estimated the following probability distribution of the price of a ceiling fan in the next summer season:

Price (P)	:	800	825	850	875	900
Probability (p)	:	0.15	0.25	0.30	0.20	0.10

If the demand (x) of his ceiling fans follows a linear relation x = 6000 - 4P, find expected demand of fans and expected total revenue of the dealer.

Solution

Since P is a random variable, therefore, x = 6000 - 4P, is also a random variable. Further, Total Revenue TR = P.x = 6000P - $4P^2$ is also a random variable.

From the given probability distribution, we have

$E(P)$ = 800 × 0.15 + 825 × 0.25 + 850 × 0.30 + 875 × 0.20 + 900 × 0.10

= Rs 846.25 and

$E(P^2)$ = $(800)^2$ × 0.15 + $(825)^2$ × 0.25 + $(850)^2$ × 0.30 + $(875)^2$ × 0.20 + $(900)^2$ × 0.10 = 717031.25

Thus, $E(X)$ = 6000 - 4E(P) = 6000 - 4 × 846.25 = 2615 fans.

And $E(TR)$ = 6000E(P) - $4E(P^2)$

= 6000 × 846.25 - 4 × 717031.25 = Rs 22,09,375.00

Example 2.24

A person applies for equity shares of Rs 10 each to be issued at a premium of Rs 6 per share; Rs 8 per share being payable along with the application and the balance at the time of allotment. The issuing company may issue 50 or 100 shares to those who apply for 200 shares, the probability of issuing 50 shares being 0.4 and that of issuing 100 shares is 0.6. In either case, the probability of an application being selected for allotment of any shares is 0.2 The allotment usually takes three months and the market price per share is expected to be Rs 25 at the time of allotment. Find the expected rate of return of the person per month.

Solution

Let A be the event that the application of the person is considered for allotment, B_1 be the event that he is allotted 50 shares and B_2 be the event that he is allotted 100 shares. Further, let R_1 denote the rate of return (per month) when 50 shares are allotted, R_2 be the rate of return when 100 shares are allotted and $R = R_1 + R_2$ be the combined rate of return.

We are given that P(A) = 0.2, $P(B_1/A)$ = 0.4 and $P(B_2/A)$ = 0.6.

(a) When 50 shares are allotted

The return on investment in 3 months = (25 - 16)50 = 450

$\therefore$ Monthly rate of return $= \frac{450}{3} = 150$

The probability that he is allotted 50 shares $= P(A \cap B_1) = P(A).P(B_1/A) = 0.2 \times 0.4 = 0.08$

Thus, the random variable R_1 takes a value 150 with probability 0.08 and it takes a value 0 with probability 1 - 0.08 = 0.92

$\therefore$ $E(R_1) = 150 \times 0.08 + 0 = 12.00$

(b) When 100 shares are allotted

The return on investment in 3 months = (25 - 16).100 = 900

$\therefore$ Monthly rate of return $= \frac{900}{3} = 300$

The probability that he is allotted 100 shares $= P(A \cap B_2) = P(A).P(B_2/A) = 0.2 \times 0.6 = 0.12$

Thus, the random variable R_2 takes a value 300 with probability 0.12 and it takes a value 0 with probability 1 - 0.12 = 0.88

$\therefore$ $E(R_2) = 300 \times 0.12 + 0 = 36$

Hence, $E(R) = E(R_1 + R_2) = E(R_1) + E(R2) = 12 + 36 = 48$

Example 2.25

What is the mathematical expectation of the sum of points on n unbiased dice?

Solution

Let X_i denote the number obtained on the i th die. Therefore, the sum of points on n dice is $S = X_1 + X_2 + \ldots\ldots + X_n$ and

$E(S) = E(X_1) + E(X_2) + \ldots\ldots + E(X_n)$.

Further, the number on the i th die, i.e., X_i follows the following distribution :

X_i	:	1	2	3	4	5	6
$p(X_i)$	:	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$\therefore$ $E(X_i) = \frac{1}{6}(1+2+3+4+5+6) = \frac{7}{2}$ (i = 1, 2, ... n)

Thus, $E(S) = \frac{7}{2} + \frac{7}{2} + \ldots. + \frac{7}{2}$ (n times) $= \frac{7n}{2}$

Example 2.26

If X and Y are two independent random variables with means 50 and 120 and variances 10 and 12 respectively, find the mean and variance of Z = 4X + 3Y.

Solution

$E(Z) = E(4X + 3Y) = 4E(X) + 3E(Y) = 4 \times 50 + 3 \times 120 = 560$

Since X and Y are independent, we can write

$Var(Z) = Var(4X + 3Y) = 16Var(X) + 9Var(Y) = 16 \times 10 + 9 \times 12 = 268$

Example 2.27

It costs Rs 600 to test a machine. If a defective machine is installed, it costs Rs 12,000 to repair the damage resulting to the machine. Is it more profitable to install the machine without testing if it is known that 3% of all the machines produced are defective? Show by calculations.

Solution

Here X is a random variable which takes a value 12,000 with probability 0.03 and a value 0 with probability 0.97.

$\therefore$ $E(X) = 12000 \times 0.03 + 0 \times 0.97 = Rs\ 360.$

Since E(X) is less than Rs 600, the cost of testing the machine, hence, it is more profitable to install the machine without testing.

Check Your Progress 2

Fill in the blanks:

1. ..... is used as a model to describe the probability distribution of a random variable defined over a unit of time, length or space.
2. The number of occurrences in an interval is .....of the number of occurrences in another interval .
3. Normal distribution was first observed as the .....by the statisticians of the eighteenth century.

2.12 SUMMARY

The concept of probability originated from the analysis of the games of chance in the 17th century. It is the backbone of Statistical Inference and Decision Theory that are essential tools of the analysis of most of the modern business and economic problems.

A phenomenon or an experiment which can result into more than one possible outcome is called a random phenomenon or random experiment or statistical experiment. If n is the number of equally likely, mutually exclusive and exhaustive outcomes of a random experiment out of which m outcomes are favorable to the occurrence of an event A, then the probability that A occurs, denoted by $P(A)$.

Distribution	*p.m.f. / p.d.f.*	*Range of R.V.*	*Parameters*
(i) Binomial	${}^{n}C_r p^r q^{n-r}$	$0, 1, 2, \ldots . n$	n and p
(ii) Hyper - geometric	$\frac{\left({}^{k}C_r\right)\left({}^{N-k}C_{n-r}\right)}{{}^{N}C_r}$	$0, 1, 2, \ldots . n$	n,N and k
(iii) Poisson	$\frac{e^{-m}.m^r}{r!}$	$0, 1, 2, \ldots . \infty$	m
(iv) Exponential	$m.e^{-mt}$	$0 < t < \infty$	m
(v) Normal	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$	$-\infty < X < \infty$	μ and σ
(vi) S.Normal	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$	$-\infty < z < \infty$	0 and 1

2.13 KEYWORDS

- Binomial Distribution
- Poisson distribution
- Expected value
- Functions
- Random Variable
- Normal Distribution
- Hypergeometric distribution

2.14 REVIEW QUESTIONS

Section A (2 or 3 Marks Questions)

1. What is probability?
2. Write down the concept of addition theorem.
3. What are the parameters of a binomial distribution?
4. State the conditions under which binomial probability model is appropriate.
5. What is a 'Poisson Process'?
6. Write some business and economic situations where Poisson probability model is appropriate.
7. An unbiased coin is tossed 5 times. Find the probability of getting (i) two heads, (ii) at least two heads.
8. Assume that the probability that a bomb dropped from an aeroplane will strike a target is 1/5. If six bombs are dropped, find the probability that (i) exactly two will strike the target, (ii) at least two will strike the target.
9. Write short notes on:
   (a) Fitting of Binomial Distribution
   (b) Poisson Distribution
   (c) Normal Distribution

10. In an army battalion 60% of the soldiers are known to be married and remaining unmarried. If $p(r)$ denotes the probability of getting r married soldiers from 5 soldiers, calculate $p(0)$, $p(1)$, $p(2)$, $p(3)$, $p(4)$ and $p(5)$. If there are 500 rows each consisting of 5 soldiers, approximately how many rows are expected to contain

    (i) All married soldiers, (ii) all unmarried soldiers?

11. A local politician claims that the assessed value of houses, for house tax purposes by the Municipal Corporation of Delhi, is not correct in 90% of the cases. Assuming this claim to be true, what is the probability that out of a sample of 4 houses selected at random (i) at least one will be found to be correctly assessed (ii) at least one will be found to be wrongly assessed?

12. The administrator of a large airport is interested in the number of aircraft departure delays that are attributable to inadequate control facilities. A random sample of 10 aircraft take off is to be thoroughly investigated. If the true proportion of such delays in all departures is 0.40, what is the probability that 4 of the sample departures are delayed because of control inadequacies? Also find mean, variance and mode of the random variable.

13. Fit a binomial distribution to the following data :

x:	0	1	2	3	4
f:	28	62	46	10	4

14. Five coins were tossed 100 times and the outcomes are recorded as given below.

    Compute the expected frequencies.

No. of heads :	0	1	2	3	4	5
Observed frequency :	2	10	24	38	18	8

15. A company manufactures batteries and guarantees them for a life of 24 months.

    (i) If the average life has been found in tests to be 33 months with a standard deviation of 4 months, how many batteries will have to be replaced under guarantee if the life of the batteries follows a normal distribution?

    (ii) If annual sales are 10,000 batteries at a profit of Rs 50 each and each replacement costs the company Rs 100, find the net profit.

    (iii) Would it be worth its while to extend the guarantee to 27 months if the sales were to be increased by this extra offer to 12,000 batteries?

16. (a) From the past experience, a committee for admission to certain course consisting of 200 seats, has estimated that 5% of those granted admission do not turn up. If 208 letters of intimation of admission are issued, what is the probability that seat is available for all those who turn up? Use normal approximation to the binomial.

    (b) The number of customer arrivals at a bank is a Poisson process with average of 6 customers per 10 minutes. (a) What is the probability that the next customer will arrive within 3 minutes? (b) What is the probability that the time until the next customer arrives will be from 2 to 3 minutes? (c) What is the probability that the next customer will arrive after more than 4 minutes?

17. The marks obtained in a certain examination follow normal distribution with mean 45 and standard deviation 10. If 1,000 students appeared at the examination, calculate the number of students scoring (i) less than 40 marks, (ii) more than 60 marks and (iii) between 40 and 50 marks.

Answers to Check Your Progress

Check Your Progress 1

1. A random phenomenon or random experiment or statistical experiment.
2. Event
3. Expected monetary value

Check Your Progress 2

1. Poisson distribution
2. Independent
3. Normal law of errors

2.15 REFERENCES AND FURTHER READING

- Oakshott, L. (2021). Essential quantitative methods: For business, management, and finance (6th ed.). Macmillan. ISBN: 9781137610890.
- Bell, M. L. (2020). Research methods and quantitative techniques (2nd ed.). Routledge. ISBN: 9781138473876.

UNIT 3

Sampling and Sampling Distributions

CHAPTER OUTLINE

3.1 INTRODUCTION

The most important task in carrying out a survey is to select the sample. Sample selection is undertaken for practical impossibility to survey the population. By applying rationality in selection of samples, we generalize the findings of our research. There are different types of sampling, which you would study in this lesson.

A theoretical probability distribution is constructed on the basis of the specification of the conditions of a random experiment. In contrast to this, if the construction of the probability distribution is based upon the random experiment of obtaining a sample from a population, the resulting distribution is termed as a sampling distribution.

As we know that the main aim of obtaining a sample from a population is to draw certain conclusions about it. The process of drawing such conclusions, known as 'Statistical Inference', is based upon the rules or the framework provided by various sampling distributions.

Meaning of Sampling Distribution

A sampling distribution is the probability distribution, under repeated sampling of the population, of a given statistic (numerical quantity calculated from the data values in a sample). Involves items selected at random from a population and used to test hypotheses about the population sampling is an important tool for determining the characteristics of a population. We usually don't know the population's parameters (mean, standard deviation, etc.), but often want reliable estimates of them. Ensuring random (representative) sampling free of bias and sampling errors is important. An important rule to remember is: No randomization, no generalization.

Population and Samples

A population is any entire collection of people, animals, plants or things from which we may collect data. It is the entire group we are interested in, which we wish to describe or draw conclusions about. A large population may be impractical and costly to study; collecting data from every member of the population is not possible. A sample is more manageable and easier to study. A sample is a *part* of the population of interest, a sub-collection selected from a population. In order to make any generalizations about a population, a sample, that is meant to be representative of the population, is often studied. For each population there are many possible samples. A sample statistic gives information about a corresponding population parameter. For example, the sample mean for a set of data would give information about the overall population mean. It is important that the population is carefully and completely defined before collecting the sample.

Example 3.1: The population for a study of infant health might be all children born in the India in the 1980's. The sample might be all babies born on 10 June in any of the years.

After collecting and organizing the data, a summary is made such as average values. Hopefully valid conclusions can be made on the whole population based on the sample data. Therefore it is important that the sample data collected be representative of the population. Otherwise conclusions may be invalid. Conclusions are only as reliable as the sampling process, and information can change from sample to sample.

Distinction between population and sample

A population or a universe is the totality of the units under the field of investigation. These units are also called items or objects or individuals or sampling units, which may be animates or inanimate. According to Simpson and Kafka, "A universe or a population may be defined as an aggregate of items possessing a common trait or traits." The term 'population', in contrast to its common meaning, has a wider meaning in statistics. According to G. Kalton, "In statistical usage, the term population does not necessarily refer to people but is a technical term used to describe the complete group of persons or objects for which the results are to apply." For example, if we want to study the marks obtained by students of B. Com. of a university, the population will be all the B. Com. Students of that university. Further, if we wish to determine average yield of wheat per acre in a particular year, the population will be all those acres of land which were under wheat crop in that year. In the words of Norma Gilbert, "A population consists of all the individuals or objects in a well defined group about which information is needed to answer a question."

Parameters and Statistics

Parameters

Parameter is a value, usually unknown (and which therefore has to be estimated), used to represent a certain population characteristic. For example, the population mean is a parameter that is often used to indicate the average value of a quantity. Within a population, a parameter is a fixed value which does not vary. Each sample drawn from the population has its own value of any statistic that is used to estimate this parameter. For example, the mean of the data in a sample is used to give information about the overall mean in the population from which that sample was drawn. Parameters are often assigned Greek letters (e.g. σ), whereas statistics are assigned Roman letters (e.g. σ).

Statistic

A statistic is a quantity that is calculated from a sample of data. It is used to give information about unknown values in the corresponding population. For example, the average of the data in a sample is used to give information about the overall average in the population from which that sample was drawn.

It is possible to draw more than one sample from the same population and the value of a statistic will in general vary from sample to sample. For example, the average value in a sample is a statistic. The average values in more than one sample, drawn from the same population, will not necessarily be equal. Statistics are often assigned Roman letters (e.g. m and s), whereas the equivalent unknown values in the population (parameters) are assigned Greek letters (e.g. μ and σ).

3.2 TYPES OF SAMPLING

Sampling is divided into two types:

- ***Probability sampling:*** In probability sample, every unit in the population has equal chances for being selected as a sample unit.
- ***Non-probability sampling:*** In non-probability sampling, units in the population have unequal or zero chances for being selected as a sample unit.

Probability Sampling Techniques

1. Random sampling
2. Systematic sampling
3. Stratified random sampling
4. Cluster sampling

Random Sampling

Simple random sample is a process in which every item of the population has equal probability of being chosen.

There are two methods used in random sampling –

1. Lottery method
2. Using random number table.

1. ***Lottery method:*** Take a population containing 4 departmental stores: *A, B, C* & *D*. Suppose we need to pick a sample of two stores from the population using simple random procedure. We write down all possible sample of two. Six different combinations each contain two stores from the population. Combination is *AB, AD, AC, BC BD, CD*. We can now write down 6 sample combination on six identical pieces of paper, fold the piece of paper so that they cannot be distinguished. Put them in a box. Mix it and pull one at random. This procedure is the lottery method of making random selection.
2. ***Using random number table:*** A Random number table consists of a group of digits that are arranged in random order, i.e. any row, column, or diagonal in such a table contains digits that are not in any systematic order. There are three tables for random numbers (a) Tippet's table (b) Fisher and Yate's table (c) Kendall and Raington table.

Table for random number is as follows:

Table 3.1: Random Number

40743	39672
80833	18496
10743	39431
88103	23016
53946	43761
31230	41212
24323	18054

Example 3.2: Taking the earlier example of stores we first number the stores.

1 *A* 2 *B* 3 *C* 4 *D*

The stores *A, B, C, D* has been numbered as 1,2,3,4.

In order to select 2 shops out of 4 randomly, we proceed as follows:

Suppose we start with second row in the first column of the table and decide to read diagonally. The starting digit is 8. There is no departmental stores with number 8 in the population. There are only 4

stores. Move to the next digit on the diagonal, which is 0. Ignore it since it does not correspond to any stores in the population. The next digit on the diagonal is 1 which corresponds to store *A*. Pick *A* and proceed until we get 2 samples. In this case the 2 departmental stores are 1 and 4. Sample derived from this consists of departmental stores *A* and *D*.

In random sampling there are two possibilities (1) Equal probability (2) Varying probability.

(a) ***Equal Probability:*** This is also called as random sampling with replacement.

Example 3.3: Put 100 chits in a box numbered 1 to 100. Pick one No. at random. Now the population has 99 chits. Now, when a Second number is picked, there are 99 chits. In order to provide equal probability, the sample selected is replaced in the population.

(b) ***Varying Probability:*** This is also called random sampling without replacement. Once a number is picked, it is not included again. Therefore the probability of selecting a unit varies from the other. In our example it is 1/100, 1/99, 1/98, 1/97 if we select 4 samples out of 100.

Systematic Random Sampling

There are three steps:

1. Sampling interval K is determined

$$K = \frac{\text{No. of units in the population}}{\text{No. of units desired in the sample}}$$

2. One unit between the first and Kth unit in the population list is randomly chosen.
3. Add Kth unit to the randomly chosen number.

Example 3.4: Consider 1000 households, from which we want to select 50 units. Calculate $k = 1000/50 = 20$

To select the first unit, we randomly pick one number between 1 to 20 say 17. So our sample is starting with 17, 37, 57..... Please note that only first item was randomly selected. The rest are systematically selected. This is a very popular method because; we need only one random number.

Stratified Random Sampling

A probability sampling procedure in which simple random sub-samples are drawn from within different strata that are more or less equal on some characteristics. Stratified sampling is of two types:

1. ***Proportionate stratified sampling:*** The number of sampling units drawn from each stratum is in proportion to the population size of that stratum.
2. ***Disproportionate stratified sampling:*** The number of sampling units drawn from each stratum is based on the analytical consideration, hut not in proportion to the population size of that stratum.

Sampling process is as follows:

1. The population to be sampled is divided into groups (stratified)
2. A simple random sample is chosen

Reason for stratified sampling: Sometimes marketing professionals want information about the component part of the population. Assume there are 3 stores. Each store forms a strata and sampling from within

each strata is selected. The result might be used to plan different promotional activities for each store strata.

Suppose a researcher wishes to study the retail sale of product such as tea in a universe of 1000 grocery stores (Kirana shops included). The researcher will first divide this universe into say 3 strata based on store size. This bench mark for size could be only one of the following (a) Floor space (b) Sales volumes (c) Variety displayed etc.

Table 3.2: Strata based on store size

Stores size	No. of stores	Percentage of stores
Large stores	2000	20
Medium stores	3000	30
Small stores	5000	50
Total	**10,000**	**100**

Suppose we need 12 stores, and then choose 4 from each strata. Choose 4 stores at random. If there was no stratification, simple random sampling from the population would be expected to choose 2 large stores (20 percent of 12) about 4 medium stores (30 percent of 12) and about 6 small stores (50 percent of 12).

As can be seen, each store can be studied separately using stratified sample.

Stratified sampling can be carried out with

1. Same proportion across strata called proportionate stratified sample
2. Varying proportion across strata called disproportionate stratified sample.

Example 3.5:

Table 3.3: Strata based on stores size

Stores size	No. of stores (Population)	Sample Proportionate	Sample Disproportionate
Large	2000	20	25
Medium	3000	30	35
Small	5000	50	40
Total	**10,000**	**100**	**100**

Estimation of universe mean with a stratified sample

Example 3.6:

Table 3.4: Strata based on stores size

Stores size	Sample Mean Sales per store	No. of stores	Percent of stores
Large	200	2000	20
Medium	80	3000	30
Small	40	5000	50
Total		**10,000**	**100**

The population mean of monthly sales is calculated by multiplying the sample mean by its relative weight.

$$200 \times 0.2 + 80 \times 0.3 + 40 \times 0.5 = 84$$

Sample Proportionate: If N is the size of the population.

n is the size of the strain.

i represents 1, 2, 3, ... k [number of strata in the population]

Proportionate sampling

$$P = \frac{n_1}{N_1} = \frac{n_2}{N_2} = \ldots\ldots\ldots\ldots = \frac{n_k}{N_k} = \frac{n}{N}$$

$$\frac{n_1}{N_1} = \frac{n}{N}$$

$$n_1 = \frac{n}{N} \times n_1$$

n_1 is the sample size to be drawn from stratum 1

$n_1 + n_2 + \ldots\ldots\ldots\ldots n_k = n$ [Total sample size of the all strata]

Illustration 3.1: A survey is planned to analyse the perception of people towards their own religious practices. Population consists of various religious, viz, Hindu, Muslim, Christian, Sikh, Jain assume total population is 10000. Hindu, Muslim, Christian, Sikh and Jains consists of 6000, 2000, 1000, 500 and 500 respectively. Determine the sample size of each stratum by applying proportionate stratified sampling. If the sample size required is 200.

Solution: Total population, $N = 10000$

Population in the strata of Hindus $N_1 = 6000$

Population in the strata of Muslims $N_2 = 2000$

Population in the strata of Christians $N_3 = 1000$

Population in the strata of Sikhs $N_4 = 500$

Population in the strata of Jains $N_5 = 500$

Proportionate stratified sampling

$$P = \frac{n_1}{N_1} = \frac{n_2}{N_2} = \frac{n_3}{N_3} = \frac{n_4}{N_4} = \frac{n_5}{N_5} = \frac{n}{N}$$

∴ Let us determine the sample size of strata N_1

$$\frac{n_1}{N_1} = \frac{n}{N} \times N_1 = \frac{200}{10000} \times 6000$$

$$= 20 \times 6$$

$$= 120.$$

$$n_2 = \frac{n}{N} \times N_2 = \frac{200}{10000} \times 2000$$
$$= 40.$$

$$n_3 = \frac{n}{N} \times N_3 = \frac{200}{10000} \times 1000$$
$$= 20$$

$$n_4 = \frac{n}{N} \times N_4 = \frac{200}{10000} \times 500$$
$$= 10$$

$$n_5 = \frac{n}{N} \times N_5 = \frac{200}{10000} \times 500 = 10$$

$$n = n_1 + n_2 + n_3 + n_4 + n_5$$
$$= 120 + 40 + 20 + 10 + 10$$
$$= 200.$$

Sample Disproportion: Let σi is the variance of the stratum i,

where i = 1, 2, 3 ..... k.

Formula to compute the sample size of the stratum i is.

is the variance of the stratum i,

where size of stratum i

r_i = Sample size of stratum i

r_i = Ratio of the size of he stratum I with that of the population.

N_i = Population of stratum i

N = Total population.

Illustration 3.2: Govt. of India wants to study the performance of women self help groups (WSHG) in three region viz. North, South and west. Total WSHG's are 1500. Number of groups in North, South and West are 600, 500 and 400 respectively. Govt. found more variation between WSHG's in North, South and West regions. The variance of performance of WSHG's in there regions are 64, 25 and 16 respectively. If the disappropriate stratified sampling is to be sued with the sample size of 100, determine the number of sampling units for each regions.

Solution: Toral Population N = 1500

Size of the stratum 1, N_1 = 600

Size of the stratum 2, N_2 = 500

Size of the stratum 3, N_3 = 400

Variance of stratum 1, $\sigma 1^2$ = 64

Variance of stratum 2, $\sigma_2^2 = 25$

Variance of stratum 3, $\sigma_3^2 = 16$

Sample size $n = 100$

Table 3.5: Cluster Sampling

Stratum Number	Size of the stratum N_i	$r_i = \frac{N_i}{N}$	σ_i	$r_i\sigma_i n$	$r_i\sigma_i n = \frac{r_i\sigma_i n}{\sum_1^3 r_i\sigma_i}$
1	600	0.4	8	3.2	54
2	500	0.33	5	1.65	28
3	400	0.26	4	1.04	18
Total					100

Cluster Sampling

Following steps are followed.

1. Population is divided into clusters
2. A simple random sample of few clusters selected
3. All the units in the selected cluster is studied.

Step 1: Mentioned above of cluster sampling is similar to the first step of stratified random sampling. But the 2 sampling methods are different. The key to cluster sampling is decided by how homogeneous or heterogeneous the clusters are.

Major advantage of simple cluster sampling is the case of sample selection. Suppose we have a population of 20,000 units from which we want to select 500 units. Choosing a sample of that size is a very time consuming process, if we use Random Numbers table. Suppose the entire population is divided into 80 clusters of 250 units, we can choose two sample clusters (2 × 250=500) easily by using cluster sampling. The most difficult job is to form clusters. In marketing the researcher forms clusters so that he can deal each cluster differently.

Example 3.7: Assume there are 20 household in a locality

Cross	Houses			
1	X_1	X_2	X_3	X_4
2	X_5	X_6	X_7	X_8
3	X_9	X_{10}	X_{11}	X_{12}
4	X_{13}	X_{14}	X_{15}	X_{16}

We need to select 8 houses. We can choose 8 houses at random. Alternatively, 2 clusters each containing 4 houses can be chosen. In this method, every possible sample of eight houses would have a known probability of being chosen - i.e. chance of one in two. We must remember that in the cluster each house has the same characteristics. With cluster sampling, it is impossible for certain random sample to be selected. For example, in the cluster sampling process described above, the following combination of houses could not occur: $X_1\ X_2\ X_5\ X_6\ X_9\ X_{10}\ X_{13}\ X_{14}$. This is because the original universe of 16 houses have been redefined as a universe of 4 clusters. So only clusters can be chosen as sample.

Check Your Progress 1

Fill in the blanks:

1. .....is also called as random sampling with replacement.
2. Stratified sampling is of two types, i.e. *Proportionate stratified sampling and* .....

3.3 SAMPLING DISTRIBUTION

A sampling distribution is the probability distribution, under repeated sampling of the population, of a given statistic (numerical quantity calculated from the data values in a sample). Involves items selected at random from a population and used to test hypotheses about the population sampling is an important tool for determining the characteristics of a population. We usually don't know the population's parameters (mean, standard deviation, etc.), but often want reliable estimates of them. Ensuring random (representative) sampling free of bias and sampling errors is important. An important rule to remember is: No randomization, no generalization.

Sampling Distribution of Sample Mean

We know that $\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$. In the previous section we have shown that if the sample is random, then each of the X_i's are random variable with mean μ and variance σ^2. Since $\overline{X}$ is a linear combination of these random variables, therefore, it is also a random variable with mean equal to $E(\overline{X}) = \frac{1}{n}\left[E(X_1) + E(X_2) + \cdots + E(X_n)\right] = \frac{1}{n} \cdot n\mu = \mu$ and variance equal to

$$Var(\overline{X}) = E(\overline{X} - \mu)^2 = E\left[\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right]^2$$

$$= E\left[\frac{(X_1 + X_2 + \cdots + X_n) - n\mu}{n}\right]^2 = \frac{1}{n^2} E\left[\sum (X_i - \mu)\right]^2$$

$$= \frac{1}{n^2} E\left[\sum (X_i - \mu)^2 + \sum\sum_{i \neq j} (X_i - \mu)(X_j - \mu)\right]$$

$$= \frac{1}{n^2}\left[\sum E(X_i - \mu)^2 + \sum\sum_{i \neq j} E(X_i - \mu)(X_j - \mu)\right]$$

$$= \frac{1}{n^2}\left[n\sigma^2 + \sum\sum_{i \neq j} Cov(X_i, X_j)\right]$$

Case I. If the sample is drawn with replacement, then X_1, X_2 ..... X_n are independent random variates and hence, $Cov(X_i, X_j) = 0$. Thus, we have

$$Var(\bar{X}) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Case II. If the sample is drawn without replacement, then

$$Cov(X_i, X_j) = -\frac{\sigma^2}{N-1}, \text{ therefore,}$$

$$Var(\bar{X}) = \frac{1}{n^2}\left[n\sigma^2 - n(n-1)\frac{\sigma^2}{N-1}\right] = \frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}$$

We note that if $N \to \infty$ (i.e., population becomes large), $\frac{N-n}{N-1} \to 1$ and therefore, in this case also, $Var(\bar{X}) = \frac{\sigma^2}{n}$.

Remarks:

1. The standard deviation of a statistic is termed as standard error. The standard error of $\bar{X}$, to be written in abbreviated form as $S.E.(\bar{X})$, is equal to $\frac{\sigma}{\sqrt{n}}$, when sampling is with replacement and it is equal to $\frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$, when sampling is without replacement.
2. $S.E.(\bar{X})$ is inversely related to the sample size.
3. The term $\sqrt{\frac{N-n}{N-1}}$ is termed as finite population correction (fpc). We note that fpc tends to become closer and closer to unity as population size becomes larger and larger.
4. As a general rule, fpc may be taken to be equal to unity when sample size is less than 5% of population size, i.e., $n < 0.05N$.

Example 3.8: Construct a sampling distribution of the sample mean for the following population when random samples of size 2 are taken from it (a) with replacement and (b) without replacement. Also find the mean and standard error of the distribution in each case.

Population Unit	:	1	2	3	4
Observation	:	22	24	26	28

Solution: The mean and standard deviation of population are

$$\mu = \frac{22+24+26+28}{4} = 25 \text{ and}$$

$$\sigma = \sqrt{\frac{(22)^2+(24)^2+(26)^2+(28)^2}{4} - (25)^2} = \sqrt{5} = 2.236 \text{ respectively.}$$

(a) When random samples of size 2 are drawn, we have $4^2 = 16$ samples, shown below:

Sample No.	Sample Values	$\bar{X}$
1	22, 22	22
2	22, 24	23
3	22, 26	24
4	22, 28	25
5	24, 22	23
6	24, 24	24
7	24, 26	25
8	24, 28	26
9	26, 22	24
10	26, 24	25
11	26, 26	26
12	26, 28	27
13	28, 22	25
14	28, 24	26
15	28, 26	27
16	28, 28	28

Since all of the above samples are equally likely, therefore, the probability of each value of $\bar{X}$ is $\frac{1}{16}$. Thus, we can write the sampling distribution of $\bar{X}$ as given below:

$\bar{X}$	22	23	24	25	26	27	28	Total
$p(\bar{X})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	1

The mean of $\bar{X}$, i.e.,

$$\mu_{\bar{x}} = E(X) = 22 \times \frac{1}{16} + 23 \times \frac{2}{16} + 24 \times \frac{3}{16} + 25 \times \frac{4}{16} + 26 \times \frac{3}{16} + 27 \times \frac{2}{16} + 28 \times \frac{1}{16} = 25$$

Further, $S.E.(\bar{X}) = \sigma_{\bar{x}} = \sqrt{E(\bar{X}^2) - [E(\bar{X})]^2}$, where

$$E(\bar{X}^2) = \frac{1}{16}(22^2 + 23^2 \times 2 + 24^2 \times 3 + 25^2 \times 4 + 26^2 \times 3 + 27^2 \times 2 + 28^2) = 627.5$$

Thus, $\sigma_{\bar{x}} = \sqrt{627.5 - 25^2} = \sqrt{2.5}$ which is equal to $\frac{\sigma}{\sqrt{n}}$.

(b) When random samples of size 2 are drawn without replacement, we have ${}^4C_2$ samples, shown below:

Sample No.	Sample Values	$\bar{X}$
1	22,24	23
2	22,26	24
3	22,28	25
4	24,26	25
5	24,28	26
6	26,28	27

Since all the samples are equally likely, the probability of each value of $\bar{X}$ is $\frac{1}{6}$. Thus, we can write the sampling distribution of $\bar{X}$ as

$\bar{X}$	23	24	25	26	27	*Total*
$p(\bar{X})$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

Further, $\mu_{\bar{X}} = E(\bar{X}) = \frac{1}{6}[23 + 24 + 25 \times 2 + 26 + 27] = 25.$

To find $S.E.(\bar{X})$, we first find $E(\bar{X}^2)$ given by

$$E(\bar{X}^2) = \frac{1}{6}[23^2 + 24^2 + 2 \times 25^2 + 26^2 + 27^2] = \frac{3760}{6} = 626.67.$$

Thus, $\sigma_{\bar{X}} = \sqrt{626.67 - 25^2} = \sqrt{1.67} = 1.292.$

Alternatively, $\sigma_{\bar{X}} = \sqrt{\frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}} = \sqrt{\frac{4-2}{3} \times \frac{5}{2}} = \sqrt{1.67} = 1.292.$

Standard Error

The standard error of the estimate of regression is given by the positive square root of the variance of e_i values.

The standard error of the estimate of regression of Y on X or simply the standard error of the estimate of Y is given as, $S_{Y.X} = \sigma_Y \sqrt{1 - r^2}$.

Similarly, $S_{Y.X} = \sigma_Y \sqrt{1 - r^2}$ is the standard error of the estimate X.

Remarks: According to the theory of estimation, an unbiased estimate of the variance of e_i values is given by

$$s^2_{Y.X} = \frac{\sum e_i^2}{n-2} = \frac{n}{n-2} \times \frac{\sum e_i^2}{n} = \frac{n}{n-2} \times \sigma_Y^2 (1 - r^2)$$

The standard errors of the estimate of Y and that of X are written as

$s_{Y.X} = \sigma_Y \sqrt{\frac{n}{(n-2)}(1-r^2)}$ and $s_{X.Y} = \sigma_X \sqrt{\frac{n}{(n-2)}(1-r^2)}$ respectively.

Note that difference between these standard errors tend to be equal to the standard errors for large values of n. In practice, the value of $n > 30$ may be treated as large.

Example 3.9: From the following data, compute (i) the coefficient of correlation between X and Y, (ii) the standard error of the estimate of Y:

$$\sum x^2 = 24 \quad \sum y^2 = 42 \quad \sum xy = 30 \quad N = 10$$

Where $x = X - \bar{X}$ and $y = Y - \bar{Y}$

Solution: The coefficient of correlation between X and Y is given by

$$r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}} = \frac{30}{\sqrt{24}\sqrt{42}} = 0.94$$

The standard error of the estimate of Y is given by ($n < 30$)

$$s_{Y.X} = \sqrt{\frac{(1-r^2)\sum y^2}{n-2}} = \sqrt{\frac{(1-0.94^2)\times 42}{8}} = 0.79$$

3.4 SAMPLING FROM NORMAL AND NON-NORMAL POPULATIONS

It can be deduced that when a random sample $X_1, X_2 \ldots\ldots X_n$ is obtained from a normal population with mean μ and standard deviation σ, then each of the X_i's are also distributed normally with mean μ and standard deviation σ.

By the use of additive (or reproductive) property of normal distribution, it follows that the distribution of $\bar{X}$, a linear combination of $X_1, X_2 \ldots\ldots X_n$, will also be normal. As shown in the previous section, the mean and standard error of the distribution would be μ and $\frac{\sigma}{\sqrt{n}}$ respectively.

Remarks: Since normal population is often a large population, the *fpc* is always taken equal to unity.

The nature of the sampling distribution of $\bar{X}$, when parent population is not normal, is provided by Central Limit Theorem. This theorem states that:

If $X_1, X_2 \ldots\ldots X_n$ is a random sample of size n from a non-normal population of size N with mean μ and standard deviation σ, then the sampling distribution of $\bar{X}$ will approach normal distribution with mean μ and standard error $\frac{\sigma}{\sqrt{n}}\left(or \sqrt{\frac{N-n}{N-1}\times\frac{\sigma^2}{n}}\right)$ as n becomes larger and larger.

Remarks: As a general rule, when $n \geq 30$, the sampling distribution of $\bar{X}$ is taken to be normal for practical purposes.

Application of the Sampling Distribution

Decisions by various government and non-government agencies are made on the basis of sample results. For example, a sales manager may take a sample of quantities purchased of its product to predict sales. A government agency may take a sample of residents to assess the effect of a certain welfare program etc. Thus, in order to draw reliable conclusions, we must have a sound knowledge regarding the sample. An extremely common and quite useful knowledge about the sample is given by the sampling distribution of the relevant statistic.

An important application of sampling distribution is to determine the probability of the statistic lying in a given interval.

Sampling Distribution of the Difference between Two Sample Means

Let there be two populations of sizes N_1 and N_2, means μ_1 and μ_2 and standard deviations σ_1 and σ_2 respectively. Let $\bar{X}_1$ be the mean of the random sample of size n_1 obtained from the first population and $\bar{X}_2$ be the mean of the random sample of size n_2 obtained from the second population. Thus, we can regard $\bar{X}_1$ and $\bar{X}_2$ as two independent random variables with means μ_1 and μ_2 and standard errors as $\frac{\sigma_1}{\sqrt{n_1}}\left(or\sqrt{\frac{N_1-n_1}{N_1-1}\times\frac{\sigma_1^2}{n_1}}\right)$ and $\frac{\sigma_2}{\sqrt{n_2}}\left(or\sqrt{\frac{N_2-n_2}{N_2-1}\times\frac{\sigma_2^2}{n_2}}\right)$ respectively.

Further, their difference, $\bar{X}_1 - \bar{X}_2$, will also be a random variable with mean $= E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$ and standard error

$$= \sqrt{Variance(\bar{X}_1 - \bar{X}_2)} = \sqrt{Var(\bar{X}_1) + Var(\bar{X}_2)}$$

$$= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \text{ (when both the samples are drawn using } srswr\text{) or}$$

$$= \sqrt{\frac{N_1-n_1}{N_1-1}\times\frac{\sigma_1^2}{n_1} + \frac{N_2-n_2}{N_2-1}\times\frac{\sigma_2^2}{n_2}} \text{ (when both the samples are drawn using } srswor\text{).}$$

Remarks:

1. When both the populations are normal, then $\bar{X}_1 - \bar{X}_2$ will be distributed normally with mean $\mu_1 - \mu_2$ and standard error $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

2. Using Central Limit Theorem, the above result will also hold for a non-normal population when both n_1 and $n_2 > 30$ and *fpc* is approximately equal to unity, i.e., $n_i < 0.05\ N_i$ (for $i = 1, 2$).

Properties of the Sampling Distribution of Means

If a population is *normally distributed*, then:

1. The mean of the sampling distribution of means equals the population mean.
2. The standard deviation of the sampling distribution of means (or standard error of the mean) is smaller than the population standard deviation.

3.5 CENTRAL LIMIT THEOREM

The central limit theorem states that the sampling distribution of any statistic will be normal or nearly normal, if the sample size is large enough. The central limit theorem is a significant result which depends on sample size. It states that as the size of a sample of independent observations approaches infinity, provided data come from a distribution with finite variance, that the sampling distribution of the sample mean approaches a normal distribution.

A very important and useful concept in statistics is the Central Limit Theorem. There are essentially three things we want to learn about any distribution: (1) The location of its center; (2) its width, (3) and how it is distributed. The central limit theorem helps us approximate all three.

Central Limit Theorem: As sample size increases, the sampling distribution of sample means approaches that of a normal distribution with a mean the same as the population and a standard deviation equal to the standard deviation of the population divided by the square root of n (the sample size).

The central limit theorem explains why many distributions tend to be close to the normal distribution. The key ingredient is that the random variable being observed should be the sum or mean of many independent identically distributed random variables. One version of the theorem is

Central Limit Theorem 1 Let X_1, X_2, ... be independent, identically distributed random variables having mean μ and finite non-zero variance σ^2.

Let $S_n = X_1 + \ldots X_n$. Then

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\right) = \Phi(x)$$

where $\Phi(x)$ is the probability that a standard normal random variable is less than x.

In this pallet, we look at rolling dice again. Let X be the number of spots showing when one die is rolled. The mean value μ for rolling one die is 3.5, and the variance is $\sigma^2 = 35/12$. If S_n is the number of spots showing when n dice are rolled, then if n is "large" the random variable

$$\frac{S_n - n\mu}{\sigma\sqrt{n}}$$

should be approximately standard normal, so S_n itself should be approximately normal with mean $3.5*n$ and variance $35n/12$.

The Central Limit Theorem describes the relation of a sample mean to the population mean. If the population mean doesn't exist, then the CLT doesn't apply and the characteristics of the sample mean, Xbar, are not predictable. Attention to detail is needed here: You can always compute the numerical mean of a finite number of observations from any density (if every observation is finite). But the population mean is defined as an integral, which diverges for the Cauchy, so even though a sample mean is finite, the population mean is not. The Cauchy has another interesting property – the distribution of the sample average is that same as the distribution of an individual observation, so the scatter never diminishes, regardless of sample size.

3.6 DETERMINATION OF SAMPLE SIZE

The data which is needed to consider in sample size determination

- Variance or heterogeneity of population
- The degree of acceptable error (confidence interval)
- Confidence level

Generally, we need to make judgments on all these variables.

How to determine Variance or Heterogeneity of Population in Sample Size

This can be determining through:

- Previous studies? Industry expectations? Pilot study?
- Sequential sampling
- Rule of thumb: the value of standard deviation is expected to be 1/6 of the range.

Certain Formulas for determining sample size

Means	$n = (ZS/E)^2$
Proportions	$n = Z^2pq/E^2$
Percentiles	$n = pc\ (100 - pc)\ Z^2/E^2$

Z at 95% confidence = 1.96

Z at 99% confidence = 2.58

The sample size of a statistical sample is the number of observations that constitute it. It is typically denoted n, a positive integer formula, tables, and power function charts are well known approaches to determine sample. The sample size of a statistical sample is the number of observations that constitute its size.

Typically, all else being equal, a larger sample size leads to increased precision in estimates of various properties of the population. This can be seen in such statistical rules as the law of large numbers and the central limit theorem. Repeated measurements and replication of independent samples are often required in measurement and experiments to reach a desired precision.

The sample size determination formulas come from the formulas for the maximum error of the estimates. The formula is solved for n. Be sure to round the answer obtained up to the next whole number, not off to the nearest whole number. If you round off, then you will exceed your maximum error of the estimate in some cases.

3.7 FINITE POPULATION MULTIPLIER

The central limit theorem and the standard errors of the mean and of the proportion are based on the premise that the samples selected are chosen with replacement. However, in virtually all survey research, sampling is conducted without replacement from populations that are of a finite size N. In these cases, particularly when the sample size n is not small in comparison with the population size N (i.e., more than 5% of the population is sampled) so that $n/N > 0.05$, a **finite population correction factor (fpc)** or **finite population multiplier** is used to define both the standard error of the mean and the standard error of the proportion. The finite population correction factor is expressed as Fpc $= \sqrt{(N - n/N - 1)}$

Where N = Population Size and n = Sample Size

Standard error of the mean for finite populations would be = $\delta / \sqrt{n(\text{FPC})}$ and the standard error of the proportion for finite populations = $\sqrt{(p\ (1 - p)/n}$ * FPC.

The effect of the FPC is that the error becomes zero when the sample size n is equal to the population size N.

3.8 SAMPLING DISTRIBUTION OF NUMBER OF SUCCESSES

Let denote the proportion of successes in population, i.e.,

$$\pi = \frac{\textit{Number of successes in population}}{\textit{Total number of units in population}}$$

Let us take a random sample of n units from this population and let X denote the number of successes in the sample. Thus, X is a random variable with mean and standard error

$$\sqrt{n\pi(1-\pi)} \left(or \sqrt{\frac{N-n}{N-1} \times n\pi(1-\pi)} \right)$$

If sampling is done with replacement, then X is a binomial variate with mean $n\pi$ and standard error $\sqrt{n\pi(1-\pi)}$. Using central limit theorem, we can say that the distribution of the number of successes will approach a normal variate with mean $n\pi$ and standard error $\sqrt{n\pi(1-\pi)}$ or $\sqrt{\frac{N-n}{N-1} \times n\pi(1-\pi)}$ for sufficiently large sample. The sample size is said to be sufficiently large if both $n\pi$ and $n(1-\pi)$ are greater than 5.

Sampling Distribution of Proportion of Successes

Let $p = \frac{X}{n}$ be the proportion of successes in sample. Since X is a random variable, therefore, p is also a random variable with mean

$$E(p) = \frac{E(X)}{n} = \frac{n\pi}{n} = \pi \text{ and standard error}$$

$$= \sqrt{\frac{1}{n^2} Var(X)} = \sqrt{\frac{n\pi(1-\pi)}{n^2}} = \sqrt{\frac{\pi(1-\pi)}{n}}$$

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

As in the previous section, the sampling distribution of p will also be normal if both $n\pi$ and $n(1-\pi)$ are greater than 5.

Example 3.10

There are 500 mangoes in a basket out of which 80 are defective. If obtaining a defective mango is termed as a success, determine the mean and standard error of the proportion of successes in a random sample of 10 mangoes, drawn (a) with replacement and (b) without replacement.

Solution

It is given that $\pi = \frac{80}{500} = \frac{4}{25}$. Therefore, $E(p) = \pi = \frac{4}{25}$ and

(a) S.E.(p) $= \sqrt{n\pi(1-\pi)} = \sqrt{10 \times \frac{4}{25} \times \frac{21}{25}} = 1.159$

(b) S.E.(p) $= \sqrt{\frac{500-80}{499} \times 10 \times \frac{4}{25} \times \frac{21}{25}} = 1.063$

Example 3.11

20% under graduates of a large university are found to be smokers. A sample of 100 students is selected at random. Construct the sampling distribution of the number of smokers. Also find the probability that the number of smokers in the sample is greater than 25.

Solution

It is given that $\pi = \frac{20}{100} = \frac{1}{5}$. Since sample size, n = 100, is large, the number of successes X will be distributed normally with mean $100 \times \frac{1}{5} = 20$ and standard error $\sqrt{100 \times \frac{1}{5} \times \frac{4}{5}} = 4$.

Further, $P(X > 25) = P\left(z > \frac{25-20}{4}\right) = P(z > 1.25) = 0.1056.$

Sampling Distribution of the Difference of Two Proportions

Let p_1 be proportion of successes in a random sample of size n_1 from a population with proportion of successes = π_1 and p_2 be the proportion of successes in a random sample of size n_2 from second population with proportion of successes = π_2. Assuming that the sample sizes are large, we can write

$$p_1 \sim N\left(\pi_1, \sqrt{\frac{\pi_1(1-\pi_1)}{n_1}}\right) \text{ and } p_2 \sim N\left(\pi_2, \sqrt{\frac{\pi_2(1-\pi_2)}{n_2}}\right)$$

Thus, their difference $(p_1 - p_2)$ will be distributed normally with mean = $\pi_1 - \pi_2$ and standard error $\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$.

Note: The above result will hold when we ignore *fpc* and the sample size, n_1 and n_2, is greater than 5 divided by the minimum of $\pi_1, (1-\pi_1), \pi_2$ and $(1-\pi_2)$.

Check Your Progress 2

Fill in the blanks:

1. The finite population correction factor is expressed as .....
2. A.....is the probability distribution, under repeated sampling of the population, of a given statistic
3. The .....and the standard errors of the mean and of the proportion are based on the premise that the samples selected are chosen with replacement.

3.9 SUMMARY

Sample is a representative of population. There are 2 types of sample (a) Probability sampling (b) Non probability sample. Probability sampling includes random sampling, stratified random sampling systematic sampling, cluster sampling, Multistage sampling. Random sampling can be chosen by Lottery method or using random number table. Samples can be chosen either with equal probability or varying probability. Random sampling can be systematic or stratified. In systematic random sampling, only the first number is randomly selected. Then by adding a constant "*K*" remaining numbers are generated. In stratified sampling, random samples are drawn from several strata, which have more or less same characteristics. In multistage sampling, sampling is drawn in several stages.

The formula for the sampling distribution depends on the distribution of the population, the statistic being considered, and the sample size used. A more precise formulation would speak of the distribution of the statistic as that for all possible samples of a given size, not just "under repeated sampling". This brief tour of probability, distributions, and the roots of statistical inferences barely scratches the surface. Many of these ideas will be amplified in future articles of this series.

Statistic	Mean	Standard Error	Range of statistics
$\bar{X}$	μ	$\frac{\sigma}{\sqrt{n}}$ *	$-\infty < \bar{X} < \infty$
p	π	$\sqrt{\frac{\pi(1-\pi)}{n}}$ *	$0 < p < 1$
χ^2	n	$2n$	$0 < \chi^2 < \infty$
t	0	$\frac{v}{v-2}$ **	$-\infty < t < \infty$
F	$\frac{v_2}{v_2-2}$ **	$\left(\frac{v_2}{v_2-2}\right)\sqrt{\frac{2(v_1+v_2-2)}{v_1(v_2-4)}}$ **	$0 < F < \infty$

* multiply this value by fpc.

** $n = n - 1$, $n_1 = n_1 - 1$ and $n_2 = n_2 - 1$.

3.10 KEYWORDS

- Sampling
- Parameter
- A Sampling Distribution
- Random sampling
- Quota sampling
- Population
- Statistic
- The Central Limit Theorem
- Stratified random sampling
- Sequential sampling

3.11 REVIEW QUESTIONS

1. Define population and sampling.
2. What are the different types of sample designs?
3. What are the types of probability sampling techniques?
4. Explain the concept of sampling distribution of a statistics.
5. Write down the two methods used in random sampling.
6. Distinguish between:
   (a) Parameter and Statistic.
   (b) Sampling distribution and Probability distribution.
   (c) Standard deviation and Standard error.
7. 'A sample is a part of target population, which is carefully selected to represent the population". Explain
8. Explain the following:
   (a) Types of stratified sampling
   (b) Reasons for stratified sampling

9. "A good sample must be based on random selection". Discuss.
10. Write short notes on:
    (a) Stratified sampling
    (b) Systematic sampling
    (c) Cluster sampling
11. (a) Suppose that the number of hours spent watching television per week by middle-aged women are normally distributed with a standard deviation = 5 hours.

    How large a sample is needed so that we can say with 99% confidence that the sample mean is off by less than one hour from population mean?

    (b) What does the standard error of a statistics measure? If for a random sample of size 100 the variance of X values is 529, estimate the standard error of X.
12. (a) Explain why a random sample of size 50 is to be preferred to a random sample of size 35 to estimate the mean of a population?

    (b) A population consists of the numbers 1, 3, 5, 7 and 9.

    (i) Enumerate all possible samples of size two which can be drawn from the population without replacement.

    (ii) Show that the mean of the sampling distribution of sample mean is equal to population mean.

    (iii) Compute the variance of the sampling distribution of sample mean and show that it is less than the population variance.
13. A manufacturer produces pins with average length of 10 cms with a standard deviation of 3.5 cms. If only those pins having length between 9.5 and 10.5 cms can be used, how many out of a sample of 100 pins must be thrown away?
14. The life of tyres manufactured by a company A is distributed normally with mean 32,000 kms and standard deviation 8,000 kms and of that manufactured by company B is distributed normally with mean 30,000 kms and standard deviation 5,000 kms. If 100 tyres of company A and 81 tyres of company B are selected at random, determine the sampling distribution of the difference between mean lives of tyres.
15. 10% of machines produced by company A are defective and 5% of those produced by company B are defective. A random sample of 250 machines is taken from company A's output and a random sample of 300 machines is taken from company B's output. What is the probability that the difference in sample proportions of defective machines is greater than or equal to 0.02.

Answers to Check Your Progress

Check Your Progress 1

1. Equal Probability
2. Disproportionate stratified sampling

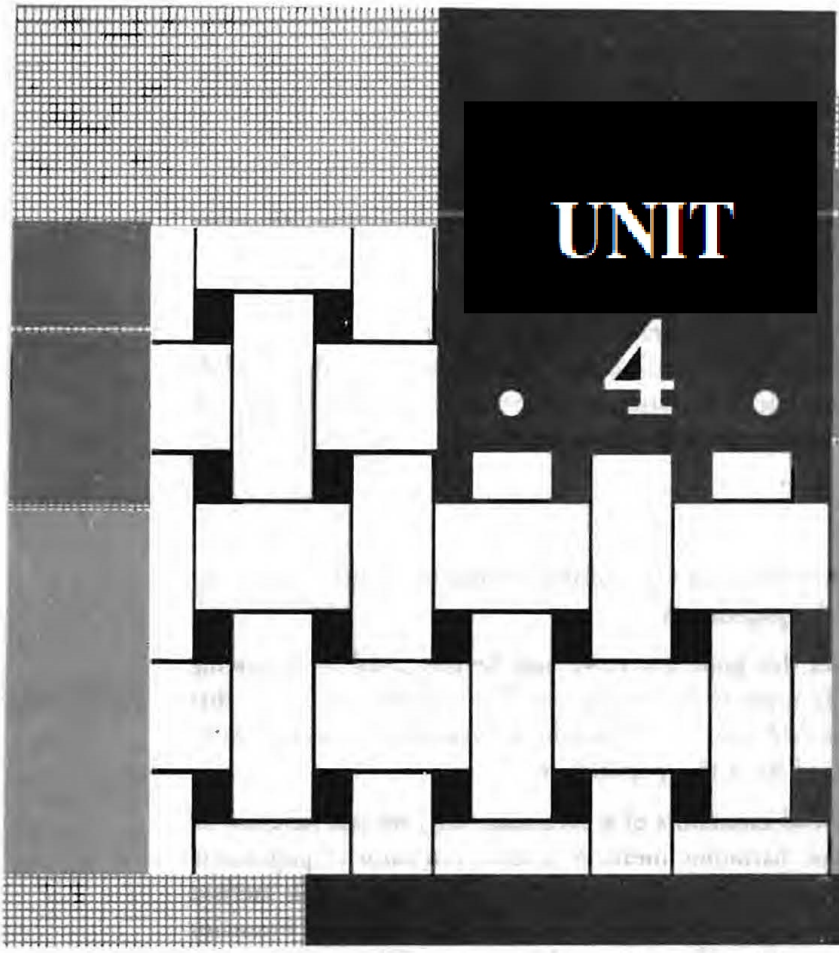
Check Your Progress 2

1. Fpc = $\sqrt{(N - n/N - 1)}$
2. Sampling distribution
3. Central limit theorem

3.12 REFERENCES AND FURTHER READING

• Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2023). Quantitative methods for business (13th ed.). Cengage Learning. ISBN: 9781337909639.
• Newbold, P., Carlson, W. L., & Thorne, B. (2021). Statistics for business and economics (9th ed.). Pearson. ISBN: 9781292315039.
• Vohra, N. D. (2020). Quantitative techniques for management (4th ed.). McGraw-Hill Education. ISBN: 9781259062897.

BLOCK – II



UNIT 4

Estimation

CHAPTER OUTLINE

4.1 INTRODUCTION

It is a procedure by which sample information is used to estimate the numerical magnitude of one or more parameters of the population. A function of sample values is called an estimator (or statistic) while its numerical value is called an estimate. For example is an estimator of population mean m. On the other hand if for a sample, the estimate of population mean is said to be 50.

Theory of Estimation

Let X be a random variable with probability density function (or probability mass function) $f(X; q_1, q_2, ... q_k)$, where $q_1, q_2, ... q_k$ are k parameters of the population.

Given a random sample $X_1, X_2, ... X_n$ from this population, we may be interested in estimating one or more of the k parameters $q_1, q_2, ... q_k$. In order to be specific, let X be a normal variate so that its probability density function can be written as $N(X : m, s)$. We may be interested in estimating m or s or both on the basis of random sample obtained from this population.

It should be noted here that there can be several estimators of a parameter, e.g., we can have any of the sample mean, median, mode, geometric mean, harmonic mean, etc., as an estimator of population mean m. Similarly, we can use either or as an estimator of population standard deviation s. This method of estimation, where single statistic like Mean, Median, Standard deviation, etc., is used as an estimator of population parameter, is known as Point Estimation. Contrary to this it is possible to estimate an interval in which the value of parameter is expected to lie. Such a procedure is known as Interval Estimation. The estimated interval is often termed as Confidence Interval.

4.2 ESTIMATOR OR POINT ESTIMATION

As we know the meaning of estimator, the words "estimator" and "estimate" are sometimes used interchangeably. We can say an estimator or point estimate is a statistic that is used to infer the value of an unknown parameter in a statistical model. Being a function of the data, the estimator is itself a random variable; a particular realization of this random variable is called the estimate.

Point Estimation (Properties of Good Estimators)

As mentioned above, there can be more than one estimators of a population parameter. Therefore, it becomes necessary to determine a good estimator out of a number of available estimators. We may recall that an estimator, a function of random variables $X_1, X_2, ... X_n$, is a random variable. Therefore, we can say that a good estimator is one whose distribution is more concentrated around the population parameter. R. A. Fisher has given the following properties of a good estimators. These are:

(i) Unbiasedness (ii) Consistency (iii) Efficiency (iv) Sufficiency.

Unbiasedness

An estimator $t\ (X_1, X_2, ... X_n)$ is said to be an unbiased estimator of a parameter θ if $E(t) = \theta$.

If $E(t) \neq \theta$, then t is said to be a biased estimator of θ. The magnitude of bias $= E(t) - \theta$. We have seen in § 20.2 that $E(\bar{X}) = \mu$, therefore, $\bar{X}$ is said to be an unbiased estimator of population mean μ.

Further, refer to § 20.4.1, we note that $E(S^2)=\frac{n-1}{n}\cdot\sigma^2$, where $S^2=\frac{1}{n}\sum(X_i-\bar{X})^2$. Therefore, S^2 is a biased estimator of σ^2. The magnitude of bias $=\left(\frac{n-1}{n}-1\right)\sigma^2=-\frac{1}{n}\sigma^2$.

Contrary to this, if we define $s^2=\frac{1}{n-1}\sum(X_i-\bar{X})^2$, we have seen in § 20.4.1 that $E(s^2) = \sigma^2$. Thus, s^2 is an unbiased estimator of σ^2. Also from § 20.3.1 we note that $E(p) = \pi$, therefore, p is an unbiased estimator of π.

Consistency

It is desirable to have an estimator, with a probability distribution, that comes closer and closer to the population parameter as the sample size is increased. An estimator possessing this property is called a consistent estimator. An estimator $t_n(X_1, X_2, \ldots\ldots X_n)$ is said to be consistent if its probability distribution converges to θ as $n \to \infty$.

Symbolically, we can write $P(t_n \to \theta) = 1$ as $n \to \infty$. Alternatively, t_n is said to be a consistent estimator of θ if $E(t_n) \to \theta$ and $\text{Var}(t_n) \to 0$, as $n \to \infty$.

We may note that $\bar{X}$ is a consistent estimator of population mean μ because $E(\bar{X})=\mu$ and $Var(\bar{X})=\frac{\sigma^2}{n}\to 0$ as $n\to\infty$.

Note: An unbiased estimator is necessarily a consistent estimator.

Efficiency

Let t_1 and t_2 be two estimators of a population parameter θ such that both are either unbiased or consistent. To select a good estimator, from t_1 and t_2, we consider another property that is based upon its variance.

If t_1 and t_2 are two estimators of a parameter θ such that both of them are either unbiased or consistent, then t_1 is said to be more efficient than t_2 if $\text{Var}(t_1) < \text{Var}(t_2)$. The efficiency of an estimator is measured by its variance.

For a normal population, we know that both the sample mean and median are unbiased estimator of population mean. However, their respective variances are $\frac{\sigma^2}{n}$ and $\frac{\pi}{2}\cdot\frac{\sigma^2}{n}$, where σ^2 is population variance. Since $\frac{\sigma^2}{n}<\frac{\pi}{2}\cdot\frac{\sigma^2}{n}$, therefore, sample mean is said to be efficient estimator of population mean.

Remarks: The precision of an estimator = 1/ S. E. of estimator.

An estimator having minimum variance among all the estimators of a population parameter is termed as Most Efficient Estimator or Best Estimator. If an estimator is unbiased and best, then it is termed as Best Unbiased Estimator. Further, if the best unbiased estimator is a linear function of the

sample observations, it is termed as Best Linear Unbiased Estimator (BLUE). It may be pointed out here that sample mean is best linear unbiased estimator of population mean.

Cramer Rao Inequality:

This inequality gives the minimum possible value of the variance of an unbiased estimator. If t is an unbiased estimator of parameter θ of a continuous population with probability density function $f(X, \theta)$, then

$$Var(t) \geq \frac{1}{nE\left(\frac{\partial \log f(X,\theta)}{\partial \theta}\right)^2}$$

Sufficiency

An estimator t is said to be a sufficient estimator of parameter θ if it utilises all the information given in the sample about θ. For example, the sample mean $\bar{X}$ is a sufficient estimator of μ because no other estimator of μ can add any further information about μ.

Let $X_1, X_2, \ldots\ldots X_n$ be a random sample of n independent observations from a population with p.d.f. (or p.m.f.) given by $f(X; \theta_1, \theta_2)$, where θ_1 and θ_2 are two parameters. The joint probability distribution of $X_1, X_2, \ldots\ldots X_n$, denoted by $L(X; \theta_1, \theta_2)$ is given by :

$$L(X; \theta_1, \theta_2) = f(X_1; \theta_1, \theta_2) \times f(X_2; \theta_1, \theta_2) \times \ldots\ldots \times f(X_n; \theta_1, \theta_2)$$

An estimator t is said to be sufficient for θ_1 if the conditional p.d.f. (or p.m.f.) of $X_1, X_2, \ldots\ldots X_n$ given t is independent of θ_1, i.e.,

$$\frac{f(X_1;\theta_1,\theta_2) \times f(X_2;\theta_1,\theta_2) \times \ldots\ldots \times f(X_n;\theta_1,\theta_2)}{g(t,\theta_1)} = h(X_1, X_2, \ldots\ldots X_n),$$ where $g(t, \theta_1)$ is p.d.f. (or p.m.f.) of t and h is a function of sample values that is independent of θ_1. We may note that each of the functions $g(t, \theta_1)$ and $h(X_1, X_2, \ldots\ldots X_n)$ may or may not be function of θ_2.

Alternatively, we can write the sufficiency condition as

$f(X_1; \theta_1, \theta_2) \times f(X_2; \theta_1, \theta_2) \times \ldots\ldots \times f(X_n; \theta_1, \theta_2) = g(t, \theta_1) \times h(X_1, X_2, \ldots\ldots X_n)$, which implies that if the joint p.d.f. (or p.m.f.) of $X_1, X_2, \ldots\ldots X_n$ can be written as a function of t and θ_1 multiplied by a function independent of θ_1, then t is sufficient estimator of θ_1.

Sufficient estimators are the most desirable but are not very commonly available. The following points must be noted about sufficient estimators:

1. A sufficient estimator is always consistent.
2. A sufficient estimator is most efficient if an efficient estimator exists.
3. A sufficient estimator may or may not be unbiased.

E.xample 4.1: If $X_1, X_2, \ldots\ldots X_n$ is a sample of n independent observations from a normal population with mean μ and variance σ^2, show that $\bar{X}$ is a sufficient estimator of μ but $S^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$ is not sufficient estimator of σ^2.

Solution: The probability density function of a normal variate is given by

$$f(X;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(X-\mu)^2}$$

Thus, the joint probability density function of X_1, X_2, ... X_n is given by

$$f(X_1;\mu,\sigma)\times f(X_2;\mu,\sigma)\times \ldots \times f(X_n;\mu,\sigma)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2}$$

We can write $X_i-\mu=(X_i-\bar{X})+(\bar{X}-\mu)$.

Squaring both sides and taking sum over n observations, we get

$$\begin{aligned}\sum(X_i-\mu)^2 &= \sum(X_i-\bar{X})^2+\sum(\bar{X}-\mu)^2+2\sum(X_i-\bar{X})(\bar{X}-\mu)\\ &=\sum(X_i-\bar{X})^2+n(\bar{X}-\mu)^2+2(\bar{X}-\mu)\sum(X_i-\bar{X})\\ &=\sum(X_i-\bar{X})^2+n(\bar{X}-\mu)^2 \qquad \text{(last term is zero)}\\ &=nS^2+n(\bar{X}-\mu)^2\end{aligned}$$

Therefore, we can write $-\frac{1}{2\sigma^2}\sum(X_i-\mu)^2=-\frac{n}{2\sigma^2}S^2-\frac{n}{2\sigma^2}(\bar{X}-\mu)^2$.

Hence $f(X_1;\mu,\sigma)\times f(X_2;\mu,\sigma)\times \ldots \times f(X_n;\mu,\sigma)$

$$=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{n}{2\sigma^2}S^2-\frac{n}{2\sigma^2}(\bar{X}-\mu)^2}=e^{-\frac{n}{2\sigma^2}(\bar{X}-\mu)^2}\times\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{n}{2\sigma^2}S^2}$$

$$=g(\bar{X},\mu,\sigma)\times h(S^2,\sigma)$$

Since h is independent of μ, therefore $\bar{X}$ is a sufficient estimator of μ. However, S^2 is not sufficient estimator of s^2 because g is not independent of σ.

Further, if we define $S^2=\frac{1}{n}\sum(X_i-\mu)^2$, then

$$f(X_1;\mu,\sigma)\times f(X_2;\mu,\sigma)\times \ldots \times f(X_n;\mu,\sigma)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{n}{2\sigma^2}S^2}$$

Thus, the newly defined S^2 becomes a sufficient estimator of σ^2. We note that $h(X_1, X_2, \ldots X_n) = 1$ in this case.

The above result suggests that if μ is known, then we should use $S^2 = \frac{1}{n}\sum(X_i - \mu)^2$ rather than $S^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$ because former is better estimator of σ^2.

Methods of Point Estimation

Given various criteria of a good estimator, the next logical step is to obtain an estimator possessing some or all of the above properties.

There are several methods of obtaining a point estimator of the population parameter. For example, we can use the method of maximum likelihood, method of least squares, method of minimum variance, method of minimum χ^2, method of moments, etc. We shall, however, use the most popular method of maximum likelihood.

Method of Maximum Likelihood

Let $X_1, X_2, \ldots\ldots X_n$ be a random sample of n independent observations from a population with probability density function (or p.m.f.) $f(X; \theta)$, where θ is unknown parameter for which we desire to find an estimator.

Since $X_1, X_2, \ldots\ldots X_n$ are independent random variables, their joint probability function or the probability of obtaining the given sample, termed as likelihood function, is given by

$$L = f(X_1; \theta) \cdot f(X_2; \theta) \cdot \ldots\ldots \cdot f(X_n; \theta) = \prod_{i=1}^{n} f(X_i; \theta).$$

We have to find that value of θ for which L is maximum. The conditions for maxima of L are: $\frac{dL}{d\theta} = 0$ and $\frac{d^2L}{d\theta^2} < 0$. The value of θ satisfying these conditions is known as Maximum Likelihood Estimator (MLE).

Generalising the above, if L is a function of k parameters $\theta_1, \theta_2, \ldots\ldots \theta_k$, the first order conditions for maxima of L are: $\frac{\partial L}{\partial \theta_1} = \frac{\partial L}{\partial \theta_1} = \ldots\ldots \frac{\partial L}{\partial \theta_k} = 0$.

This gives k simultaneous equations in k unknowns $\theta_1, \theta_2, \ldots\ldots \theta_k$, and can be solved to get k maximum likelihood estimators.

Sometimes it is convenient to work using logarithm of L. Since log L is a monotonic transformation of L, the maxima of L and maxima of log L occur at the same value.

Properties of Maximum Likelihood Estimators

1. The maximum likelihood estimators are consistent.
2. The maximum likelihood estimators are not necessarily unbiased. If a maximum likelihood estimator is biased, then by slight modifications it can be converted into an unbiased estimator.
3. If a maximum likelihood estimator is unbiased, then it will also be most efficient.

4. A maximum likelihood estimator is sufficient provided sufficient estimator exists.
5. The maximum likelihood estimators are invariant under functional transformation, i.e., if t is a maximum likelihood estimator of θ, then $f(t)$ would be maximum likelihood estimator of $f(\theta)$.

4.3 INTERVAL ESTIMATION

Using point estimation, it is possible to provide a single quantity as an estimator of a parameter. Any point estimator, even if it satisfies all the characteristics of a good estimator, has a limitation that it provides no information about the magnitude of errors due to sampling. This problem is taken care of by the method of interval estimation, that gives a range of the estimator of the parameter.

The method of interval estimation is based upon the sampling distribution of an estimator. The standard error of the estimator is used in the construction of an interval so that the probability of the parameter lying within the interval can be specified.

Given a random sample of n observations $X_1, X_2, \ldots\ldots X_n$, we can find two values l_1 and l_2 such that the probability of population parameter θ lying between l_1 and l_2 is (say) η. Using symbols, we can write $P(l_1 \leq \theta \leq l_2) = \eta$.

Such an interval is termed as a Confidence Interval for θ and the two limits l_1 and l_2 are termed as *Confidential* or *Fiducial Limits*. The percentage probability or confidence is termed as the *Level of Confidence* or *Confidence Coefficient* of the interval. For example, the level of confidence of the above interval is $100\eta\%$. The level of confidence implies that if a large number of random samples are taken from a population and confidence intervals are constructed for each, then $100\eta\%$ of these intervals are expected to contain the population parameter θ. Alternatively, a $100\ \eta\%$ confidence interval implies that we are $100\ \eta\%$ confident that the population parameter θ lies between l_1 and l_2.

As compared to point estimation, the interval estimation is better because it takes into account the variability of the estimator in addition to its single value and thus, provides a range of values. Unlike point estimation, interval estimation indicates that estimation is an uncertain process.

The methods of construction of confidence intervals in various situations are explained through the following examples.

Confidence Interval for Population Mean

Example 4.2: Construct 95% and 99% confidence intervals for mean of a normal population.

Solution: Let $X_1, X_2, \ldots\ldots X_n$ be a random sample of size n from a normal population with mean μ and standard deviation σ.

We know that sampling distribution of $\bar{X}$ is normal with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$. Therefore,

$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ will be a standard normal variate.

From the tables of areas under standard normal curve, we can write

$$P[-1.96 \leq z \leq 1.96] = 0.95 \quad \text{or} \quad P[-1.96 \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq 1.96] = 0.95 \qquad \ldots\ (1)$$

The inequality $-1.96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ can be written as

$$-1.96\frac{\sigma}{\sqrt{n}} \leq \bar{X}-\mu \text{ or } \mu \leq \bar{X}+1.96\frac{\sigma}{\sqrt{n}} \quad \ldots. (2)$$

Similarly, from the inequality $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96$, we can write

$$\mu \geq \bar{X}-1.96\frac{\sigma}{\sqrt{n}} \quad \ldots. (3)$$

Combining (2) and (3), we get

$$\bar{X}-1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}+1.96\frac{\sigma}{\sqrt{n}}$$

Thus, we can write equation (1) as

$$P\left(\bar{X}-1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

This gives us a 95% confidence interval for the parameter μ. The lower limit of μ is $\bar{X}-1.96\frac{\sigma}{\sqrt{n}}$ and the upper limit is $\bar{X}+1.96\frac{\sigma}{\sqrt{n}}$. The probability of μ lying between these limits is 0.95 and therefore, this interval is also termed as 95% confidence interval for μ.

In a similar way, we can construct a 99% confidence interval for μ as

$$P\left(\bar{X}-2.58\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}+2.58\frac{\sigma}{\sqrt{n}}\right) = 0.99$$

Thus, the 99% confidence limits for μ are $\bar{X} \pm 2.58\frac{\sigma}{\sqrt{n}}$.

Remarks: When σ is unknown and $n < 30$, we use t value instead of 1.96 or 2.58 and use S in place of σ.

Confidence Interval

In general, a confidence interval is given by:

[Sample statistic ± Table value like $Z_{\alpha/2}$ * SE], where SE is the standard deviation of the sampling distribution of the statistic.

The methods of construction of confidence intervals in various situations are explained through the following examples.

Confidence Interval for Population Mean

Example 4.3: Construct 95% and 99% confidence intervals for mean of a normal population.

Solution: Let $X_1, X_2, \ldots\ldots X_n$ be a random sample of size n from a normal population with mean μ and standard deviation σ.

We know that sampling distribution of $\bar{X}$ is normal with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$. Therefore, $z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ will be a standard normal variate.

From the tables of areas under standard normal curve, we can write

$$P[-1.96 \le z \le 1.96] = 0.95 \quad \text{or} \quad P\left[-1.96 \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le 1.96\right] = 0.95 \qquad \ldots. (1)$$

The inequality $-1.96 \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ can be written as

$$-1.96\frac{\sigma}{\sqrt{n}} \le \bar{X} - \mu \quad \text{or} \quad \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}} \qquad \ldots. (2)$$

Similarly, from the inequality $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le 1.96$, we can write

$$\mu \ge \bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \qquad \ldots. (3)$$

Combining (2) and (3), we get

$$\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

Thus, we can write equation (1) as

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

This gives us a 95% confidence interval for the parameter μ. The lower limit of μ is $\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}$ and the upper limit is $\bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$. The probability of μ lying between these limits is 0.95 and therefore, this interval is also termed as 95% confidence interval for μ.

In a similar way, we can construct a 99% confidence interval for μ as

$$P\left(\bar{X} - 2.58\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 2.58\frac{\sigma}{\sqrt{n}}\right) = 0.99$$

Thus, the 99% confidence limits for μ are $\bar{X} \pm 2.58\frac{\sigma}{\sqrt{n}}$.

Remarks: When s is unknown and $n < 30$, we use t value instead of 1.96 or 2.58 and use S in place of σ.

Confidence Interval for Population Proportion

Example 4.4: Obtain the 95% confidence limits for the proportion of successes in a binomial population.

Solution: Let the parameter π denote the proportion of successes in population. Further, p denotes the proportion of successes in n (≥ 50) trials. We know that the sampling distribution of p will be approximately normal with mean π and standard error $\sqrt{\frac{\pi(1-\pi)}{n}}$.

Since π is not known, therefore, its estimator p is used in the estimation of standard error of p, i.e.,

$$S.E.(p) = \sqrt{\frac{p(1-p)}{n}}$$

Thus, the 95% confidence interval for p is given by

$$P\left(p - 1.96\sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + 1.96\sqrt{\frac{p(1-p)}{n}}\right) = 0.95$$

This gives the 95% fiducial limits as $p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$.

Example 4.5: In a newspaper article of 1600 words in Hindi, 64% of the words were found to be of Sanskrit origin. Assuming that the simple sampling conditions hold good, estimate the confidence limits of the proportion of Sanskrit words in the writer's vocabulary.

Solution: Let π be the proportion of Sanskrit words in the writer's vocabulary. The corresponding proportion in the sample is given as $p = 0.64$.

$$\therefore \quad S.E.(p) = \sqrt{\frac{0.64 \times 0.36}{1600}} = \frac{0.48}{40} = 0.012$$

We know that almost whole of the distribution lies between 3σ limits. Therefore, the confidence interval is given by

$$P[p - 3S.E.(p) \leq \pi \leq p + 3\ S.E.(p)] = 0.9973$$

Thus, the 99.73% confidence limits of π are 0.604 (= 0.64 − 3 × 0.012) and 0.676 (= 0.64 + 3 × 0.012) respectively.

Hence, the proportion of Sanskrit words in the writer's vocabulary are between 60.4% to 67.6%.

Interval Estimates of Mean and Proportion from Large Samples

If $n > 30$ or if σ is known and the Population being sampled is normal, a $(1 - \alpha)$ Confidence interval for the population mean is given by

$$\bar{\chi} \pm z_{\alpha/2} (\sigma / \sqrt{n})$$

If σ is unknown and $n > 30$, sample standard deviations can be used as an approximation for σ $n = 100$ and $\bar{\chi} = 425$, $\sigma = 900$, 99% Confidence Interval is $425 \pm 2.58 * (900/10)$ or 192.8 to 657.2 or $P(192.8 \leq \mu \leq 657.2) = 0.99$.

By using point estimation, we may not get desired degree of accuracy in estimating a parameter. Therefore, it is better to replace point estimation by interval estimation.

An interval estimate of an unknown parameter is an interval of the form $L_1 \leq \theta \leq L_2$, where the end points L_1 and L_2 depend on the numerical value of the statistic θ for particular sample on the sampling distribution of θ. $100(1 - \alpha)$% Confidence Interval:

A $100(1 - \alpha)$% confidence interval for a parameter θ is an interval of the fprm $[L_1, L_2]$ such that $P(L_1 \leq \alpha \leq L_2) = 1 - \alpha$, 0 & lt; α & lt;1 regardless of the actual value of θ.

The quantities L_1 and L_2 are called upper and lower confidence limits and Degree of confidence (confidence coefficient) is $1 - \alpha$.

Check Your Progress 1

Fill in the blanks:

1. ..... is the process by which sample information is used to estimate the numerical magnitude of one or more parameters of the populations.
2. Using ..... estimation, it is possible to provide a single quantity as an estimator of a parameter.
3. As compared to point estimation, the interval estimation is .....
4. The method of interval estimation is based upon the ..... of an estimator.

Interval Estimation Using Distribution

A typical statistical aim is to demonstrate with 95% certainty that the true value of a parameter is within a distance B of the estimate: B is an error range that decreases with increasing sample size (n). The value of B generated is referred to as the 95% confidence interval.

For example, a simple situation is estimating a proportion in a population. To do so, a statistician will estimate the bounds of a 95% confidence interval for an unknown proportion.

The rule of thumb for (a maximum or 'conservative') B for a proportion derives from the fact the estimator of a proportion, $\hat{p} = X / n$, (where X is the number of 'positive' observations) has a (scaled) binomial distribution and is also a form of sample mean (from a Bernoulli distribution [0,1] which has a maximum variance of 0.25 for parameter $p = 0.5$). So, the sample mean X/n has maximum variance $0.25/n$. For sufficiently large n (usually this means that we need to have observed at least 10 positive and 10 negative responses), this distribution will be closely approximated by a normal distribution with the same mean and variance.

Using this approximation, it can be shown that ~95% of this distribution's probability lies within 2 standard deviations of the mean. Because of this, an interval of the form

$$(\hat{p} - 2\sqrt{0.25/n}, \hat{p} + 2\sqrt{0.25/n}) = (\hat{p} - B, \hat{p} + B)$$

will form a 95% confidence interval for the true proportion.

If we require the sampling error ε to be no larger than some bound B, we can solve the equation

$$\varepsilon \approx B = 2\sqrt{0.25/n} = 1/\sqrt{n}$$

to give us

$$1/\varepsilon^2 \approx 1/B^2 = n$$

So, n = 100 <=> B = 10%, n = 400 <=> B = 5%, n = 1000 <=> B = ~3%, and n = 10000 <=> B = 1%. One sees these numbers quoted often in news reports of opinion polls and other sample surveys.

Sample Size for Estimation

The sample size of a statistical sample is the number of observations that constitute it. It is typically denoted n, a positive integer (natural number).

Typically, all else being equal, a larger sample size leads to increased precision in estimates of various properties of the population, though the results will become less accurate if there is a systematic error in the experiment. This can be seen in such statistical rules as the law of large numbers and the central limit theorem. Repeated measurements and replication of independent samples are often required in measurement and experiments to reach a desired precision.

A typical example would be when a statistician wishes to estimate the arithmetic mean of a quantitative random variable (for example, the height of a person). Assuming that they have a random sample with independent observations, and also that the variability of the population (as measured by the standard deviation σ) is known, then the standard error of the sample mean is given by the formula:

$$\sigma/\sqrt{n}.$$

It is easy to show that as n becomes very large, this variability becomes small. This leads to more sensitive hypothesis tests with greater statistical power and smaller confidence intervals.

Three questions must be answered to determine the sample size:

1. Standard deviation of the population: It is rare that we know exact standard deviation of the population. Typically, the standard deviation of the population is estimated a) from the results of a previous survey, b) from a pilot study, c) from secondary data.
2. Maximum acceptable difference: This is the maximum amount of error that you are willing to accept. That is, it is the maximum difference that the sample mean can deviate from the true population mean before you call the difference significant.
3. Desired confidence level (%): The confidence level is your level of certainty that the sample mean does not differ from the true population mean by more than the maximum acceptable difference. Generally a 95% confidence level is used.

Example 4.6: A bank wants to determine the average number of times the customer visit bank per month. They have decided that their estimate needs to be accurate within plus or minus one-tenth of a visit, and they want to be 95% sure that their estimate does differ from true number of visits by more than one-tenth of a visit. Previous research has shown that the standard deviation is .7 visits. What is the required sample size?

Population standard deviation: .7

Maximum acceptable difference: .1

Desired confidence interval (%): 95

Required sample size = 188

Determination of an Approximate Sample Size for a Given Degree of Accuracy

Let us assume that we want to find the size of a sample to be taken from the population such that the difference between sample mean and the population mean would not exceed a given value, say ϵ, with a given level of confidence. In other words, we want to find n such that

$$P\left(\left|\bar{X}-\mu\right| \le \epsilon\right) = 0.95 \text{ (say)} \qquad \ldots. (1)$$

Assuming that the sampling distribution of $\bar{X}$ is normal with mean μ and $S.E._{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, we can write

$$P\left(-1.96 \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le 1.96\right) = 0.95 \text{ or } P\left(\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| \le 1.96\right) = 0.95$$

$$\text{or } P\left(\left|\bar{X}-\mu\right| \le 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95 \qquad \ldots. (2)$$

Comparing (1) and (2), we get

$$\epsilon = 1.96 \cdot \frac{\sigma}{\sqrt{n}} \text{ or } n = \left(\frac{1.96\sigma}{\epsilon}\right)^2 = \frac{3.84\sigma^2}{\epsilon^2}$$

Remarks:

1. The sample size required with a maximum error of estimation, ϵ and with a given level of confidence is $n = \frac{z^2\sigma^2}{\epsilon^2}$, where z is the value of standard normal variate for a given level of confidence and σ^2 is the variance of population.
2. For a given level of confidence and σ^2, n is inversely related to ϵ^2, the square of the maximum error of estimation. This implies that to reduce ϵ to $\frac{\epsilon}{k}$, the size of the sample must be k^2 times the original sample size.

3. The lesser the magnitude of $\in$, the more precise will be the interval estimate.

Example 4.7: What should be the **sample size for estimating** mean of a **normal population** if the probability that sample mean differs from population mean by not more than 30% of standard deviation is 0.99.

Solution: Let n be the size of the sample. It is given that

$$P\left(\left|\bar{X}-\mu\right| \le 0.30\sigma\right) = 0.99 \quad \text{.... (1)}$$

Assuming that the sampling distribution of $\bar{X}$ is normal with mean μ and $S.E._{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, we can write

$$P\left(\left|\bar{X}-\mu\right| \le 2.58\frac{\sigma}{\sqrt{n}}\right) = 0.99 \text{ (from table of areas)} \quad \text{.... (2)}$$

Comparing (1) and (2), we get

$$0.30\sigma = 2.58\frac{\sigma}{\sqrt{n}} \Rightarrow n = \left(\frac{2.58}{0.30}\right)^2 = 73.96 \text{ or } 74 \text{ (approx.)}$$

Example 4.8: A survey of middle class families of Delhi is proposed to be conducted for the estimation of average monthly consumption (in Rs) per family. What should be the size of the sample so that the average consumption is estimated within a range of Rs 300 with 95% level of confidence. It is known that the standard deviation of the consumption in population is Rs 1,600.

Solution: Let n denote the size of the sample to be drawn. With usual notations, we want to find n such that

$$P\left(\left|\bar{X}-\mu\right| \le 300\right) = 0.95 \quad \text{.... (1)}$$

Assuming that the sampling distribution of $\bar{X}$ is normal with mean μ and $S.E._{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, we can write $P\left(\left|\bar{X}-\mu\right| \le 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$.

or
$$P\left(\left|\bar{X}-\mu\right| \le \frac{1.96\times 1600}{\sqrt{n}}\right) = 0.95 \quad \text{.... (2)}$$

Comparing (1) and (2), we get

$$300 = \frac{1.96\times 1600}{\sqrt{n}} \text{ or } n = \left(\frac{1.96\times 1600}{300}\right)^2 = 109.3$$

Since this value is greater than 109, therefore, the size of the sample should be 110.

Confidence Interval for Population Standard Deviation

Let $S=\sqrt{\frac{1}{n}\sum\left(X_i-\bar{X}\right)^2}$ be the sample standard deviation of a random sample of size n drawn from a normal population with standard deviation σ. It can be shown that the sampling distribution of S is approximately normal, for large values of n, with mean s and standard error $\frac{\sigma}{\sqrt{2n}}$. Thus, $z=\frac{S-\sigma}{\sigma/\sqrt{2n}}$ can be taken as a standard normal variate.

Check Your Progress 2

Find out whether the following is true or false:

1. The width of confidence interval can be controlled in two ways.
2. More is the sample size wider will be the interval.
3. Lower the level of confidence the narrower will be the interval.

4.4 SUMMARY

Estimation is a procedure by which sample information is used to estimate the numerical magnitude of one or more parameters of the population. A function of sample values is called an estimator (or statistic) while its numerical value is called an estimate.

The method of estimation, where single statistic like Mean, Median, Standard deviation, etc. is used as an estimator of population parameter, is known as Point Estimation. Contrary to this it is possible to estimate an interval in which the value of parameter is expected to lie. Such a procedure is known as Interval Estimation. The estimated interval is often termed as Confidence Interval.

4.5 KEYWORDS

- Point Estimation
- Estimator
- Confidence Intervals
- Sample Size
- Interval Estimation
- Estimates
- Large Samples

4.6 REVIEW QUESTIONS

1. A random sample of 400 farms in certain year revealed that the average yield per acre of sugarcane was 925 kgs with a standard deviation of 88 kgs.
   (a) Determine the 95% confidence interval for the population mean.
   (b) What should be the size of the sample if the width of 95% confidence interval estimate of m is not more than 15?

2. A random sample of 100 sale receipts of a firm showed that its average sales per customer are Rs 250 with a standard deviation of Rs 50 (assume that there is one receipt for each customer).
   (a) Determine the 99% confidence interval for the mean sales.
   (b) How does the width of the confidence interval change if sample size is 400 instead?
   (c) How many sale receipts should be included in the sample in order that a 98% confidence interval has a maximum error of estimation equal to Rs 10.
3. A survey revealed that 30% of the persons of a state are suffering from a particular disease. How many persons should be included in the sample so that the maximum width of the 95% confidence interval of proportion of persons suffering from the disease is 0.15 units?
4. A random sample of size 64 has been drawn from a population with standard deviation 20. The mean of the sample is 80. (i) Calculate 95% confidence limits for the population mean. (ii) How does the width of the confidence interval changes if the sample size is 256 instead?
5. In a random sample of 100 articles taken from a large batch of articles, 10 are found to be defective. Obtain a 95% confidence interval for the true proportion of defectives in the batch.
6. A random sample of size 10 from a normal population gives the values 64, 72, 65, 70, 68, 71, 65, 62, 66, 67. If it is known that the standard error of the sample mean is , find 95% confidence limits for the population mean. Also find the population variance.
7. In a sample of 26 items drawn from a normal population, the 90% confidence limits for the population mean were computed as 46.584 and 53.416. Find mean and standard deviation of the sample.
8. A simple random sample of size 66 was drawn in the process of estimating the mean annual income of 950 families of a certain township. The mean and standard deviation of the sample were found to be Rs 4,730 and Rs 7.65 respectively. Find the 95% confidence interval for the population mean.
9. With a sample size of 400, the calculated standard error of mean is 2 with a mean of 120. What sample size would be required so that we could be 95% confident that the population mean is within ± 3.5 of the sample mean?
10. Mr X wants to determine, on the basis of a sample study, the mean time required to complete certain job so that he may be 95% confident that the mean may remain within ± 2 days of the true mean. As per the available record, the population variance is 64. How large should the sample be for his study?
11. A firm wishes to estimate, with a maximum allowable error of 0.05 and a 95% level of confidence, the proportion of consumers who prefer its product. How large a sample will be required to make such an estimate if the preliminary sales reports indicate that 20% of all consumers prefer the firm's product?
12. In a market area, there are 600 shops. A researcher wants to estimate the number of customers visiting these shops per day. The researcher also wants that the sampling error in this estimate is no larger than ± 10 with 95% confidence. The previous studies indicate that the standard deviation of customer arrivals is 85. If the cost per interview is Rs 20 (this includes field work, supervision of interviewers, coding, editing and tabulation of results and report writing, etc.), calculate the total cost of survey involved.

Suppose that the researcher is willing to sacrifice some accuracy in order to reduce cost. If he settles for an estimate with 90% confidence, how much reduction in cost can be achieved?

13. Ellen Harris, a time methods engineer, was accumulating normal times for various tasks on a labour- intensive assembly process. The process included 200 separate jobs stations, each performing the same assembly task. She sampled 5 stations and obtained the following assembly times for each station:

    1.8, 2.4, 2.2, 2.6, 1.6 minutes.

    (i) Calculate the mean assembly time.

    (ii) Estimate the population standard deviation.

    (iii) Construct a 99% confidence interval for the mean assembly time.

14. In a random sample of 81 items taken from a large consignment, some were found to be defective. If the standard deviation of the proportion of defective items is 1/18, find the 95% confidence limits of the percentage of defective items in the consignment.

15. The mean of a sample of size 16 from a normal population is 20. If it is known that variance of the population is 4, find the standard error of the sample mean and 95% confidence interval for population mean.

Answers to Check Your Progress

Check Your Progress 1

1. Estimation
2. Point
3. Better
4. Sampling distribution.

Check Your Progress 2

1. True
2. False
3. True.

4.7 REFERENCES AND FURTHER READING

- Balnaves, M., & Caputi, P. (2023). Introduction to quantitative research methods: An investigative approach. SAGE Publications. ISBN: 9781446209144.
- Groebner, D., Shannon, P., & Fry, P. (2019). Business statistics: A decision-making approach (10th ed.). Pearson. ISBN: 9781292220395.
- Lee, N., & Peters, M. (2021). Applied statistics for business and management using Microsoft Excel. Wiley. ISBN: 9781292243578.
- Arnold, R. (2024). Quantitative methods for management. Oxford University Press. ISBN: 9780198744488.
- Taha, H. A. (2022). Operations research: An introduction (11th ed.). Pearson. ISBN: 9781292266973.

UNIT

5

Testing of Hypotheses

CHAPTER OUTLINE

5.1 BASIC CONCEPT OF HYPOTHESIS

A hypothesis is a preconceived idea about the nature of a population or about the value of its parameters. The statements like the distribution of heights of students of a university is normally distributed, the number of road accidents per day in Delhi is 10, etc., are some examples of a hypothesis.

The test of a hypothesis is a procedure by which we test the validity of a given statement about a population. This is done on the basis of a random sample drawn from it.

The hypothesis to be tested is termed as Null Hypothesis, denoted by H_0. This hypothesis asserts that there is no difference between population and sample in the matter under consideration. For example, if H_0 is that population means $m = m_0$ then we regard the random sample to have been obtained from a population with mean m_0.

Corresponding to any H_0, we always define an Alternative Hypothesis. This hypothesis, denoted by H_a, is alternate to H_0 , i.e., if H_0 is false then H_a is true and vice-versa.

One-tailed and Two-tailed Tests

There are two type of tests: One-tailed and two-tailed tests.

A hypothesis test may be one-tailed or two-tailed. In one-tailed test the test-statistic for rejection of null hypothesis falls only in one-tailed of sampling distribution curve.

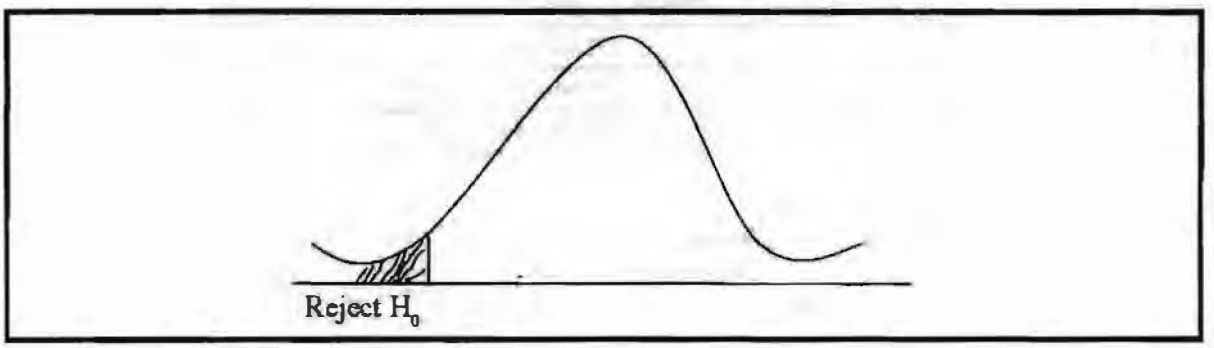


Figure 5.1

One-tailed Test

Example 5.1: In a right side test, the critical region lies entirely in the right tail of the sample distribution. Whether the test is one-sided or two-sided - depends on alternate hypothesis.

Example 5.2: A tyre company claims that mean life of its new tyre is 15,000 km. Now the researcher formulates the hypothesis that tyre life is = 15,000 km.

A two-tailed test is one in which the test statistics leading to rejection of null hypothesis falls on both tails of the sampling distribution curve as shown.

When we should apply a hypothesis test that is one-tailed or two-tailed depends on the nature of the problem. One-tailed test is used when the researcher's interest is primarily on one side of the issue. Example: "Is the current advertisement less effective than the proposed new advertisement"?

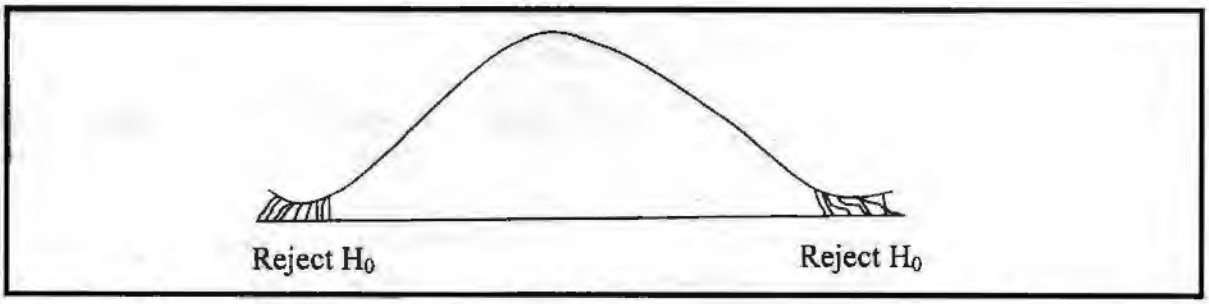


Figure 5.2

Two-tailed Test

A two-tailed test is appropriate, when the researcher has no reason to focus on one side of the issue. Example: "Are the two markets – Mumbai and Delhi different to test market a product?"

Example 5.3: A product is manufactured by a semi-automatic machine. Now, assume that the same product is manufactured by the fully automatic machine. This will be two-sided test, because the null hypothesis is that "the two methods used for manufacturing the product do not differ significantly".

$$\therefore H_0 = \mu_1 = \mu_2$$

Table 5.1: Alternate hypothesis

Sign of alternate hypothesis	Type of test
=	Two-sided
<	One-sided to right
>	One-sided to left

Degree of Freedom

It tells the researcher the number of elements that can be chosen freely. Example: $a + b/2 = 5$. fix a=3, b has to be 7. Therefore, the degree of freedom is 1.

Select test criteria

If the hypothesis pertains to a larger sample (30 or more), the Z-test is used. When the sample is small (less than 30), the T-test is used.

Compute

Carry out computation.

Make Decisions

Accepting or rejecting of the null hypothesis depends on whether the computed value falls in the region of rejection at a given level of significance.

Type I and Type II Errors

There are two kinds of errors that can be made insignificance testing: (1) a true null hypothesis can be incorrectly rejected and (2) a false null hypothesis can fail to be rejected. The former error is called a Type I error and the latter error is called a Type II error. A type I error occurs when one rejects the null

hypothesis when it is true. A Type I error is often referred to as a 'false positive', and is the process of incorrectly rejecting the null hypothesis in favor of the alternative. The probability of a type I error is the level of significance of the test of hypothesis, and is denoted by *alpha* (α). Usually a one-tailed test hypothesis is used when one talks about type I error. A Type II error is the opposite of a Type I error and is the false acceptance of the null hypothesis. A type II error occurs when one rejects the alternative hypothesis (fails to reject the null hypothesis) when the alternative hypothesis is true. The probability of a type II error is denoted by *beta* (β). A Type II error is only an error in the sense that an opportunity to reject the null hypothesis correctly was lost. It is not an error in the sense that an incorrect conclusion was drawn since no conclusion is drawn when the null hypothesis is not rejected.

These two types of errors are defined in the table.

Table 5.2

Statistical Decision	True State of the Null Hypothesis	
	H_0 True	H_0 False
Reject H_0	Type I error	Correct
Do not Reject H_0	Correct	Type II error

The two types of errors are shown by the following figure.

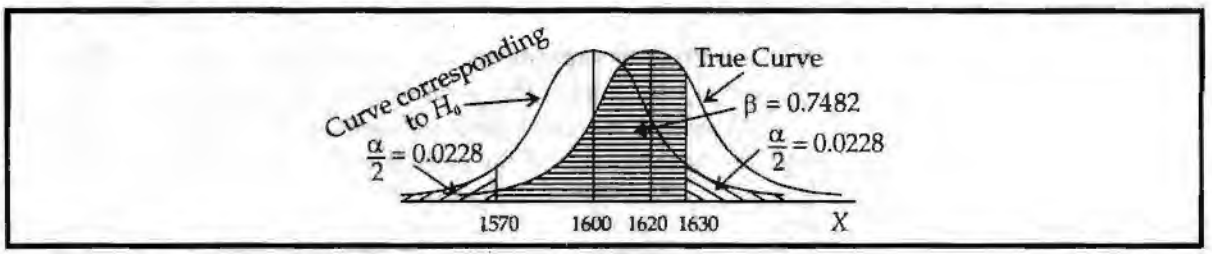


Figure 5.3

Two types of errors

It is obvious, from the above figure, that it is not possible to simultaneously control both types of errors because a decrease in probability of committing one type of error is accompanied by the increase in probability of committing the other type of error. Further, we may note that farther the true value of parameter from the hypothesised value, smaller would be the size of type II error, β. The graph of various values of μ against β is known as the *Operating Characteristic Surve.*

In the procedure of testing a hypothesis, the probability or size of type I error, i.e., α is specified in advance. Usually we take $\alpha = 0.05$ (i.e., 5%) or 0.01 (i.e., 1%).

Critical Region and One Tailed versus Two Tailed Tests

Let $H_0 : \mu = \mu_0$ against $H_a : \mu \neq \mu_0$, where μ_0 denotes some specified value of population mean μ. For example, $\mu_0 = 1600$, in the example considered above.

If we decide to have $\alpha = 0.05$, we know that for a standard normal variate $P[-1.96 \leq z \leq 1.96] = 1 - 0.05 = 0.95$, the procedure of testing of hypothesis can be outlined as:

Reject H_0 if the computed value of z from the sample $\left(i.e.,\ z_{cal} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)$ lies outside the interval (– 1.96, 1.96) and accept it otherwise.

In terms of figure, the portion of z axis covering the interval (– 1.96, 1.96), i.e, A to B is termed as the *Acceptance Region* and its remaining portions, which lie to the left of point A and to the right of point B, are termed as the *Region of Rejection or Critical Region* (C.R.).

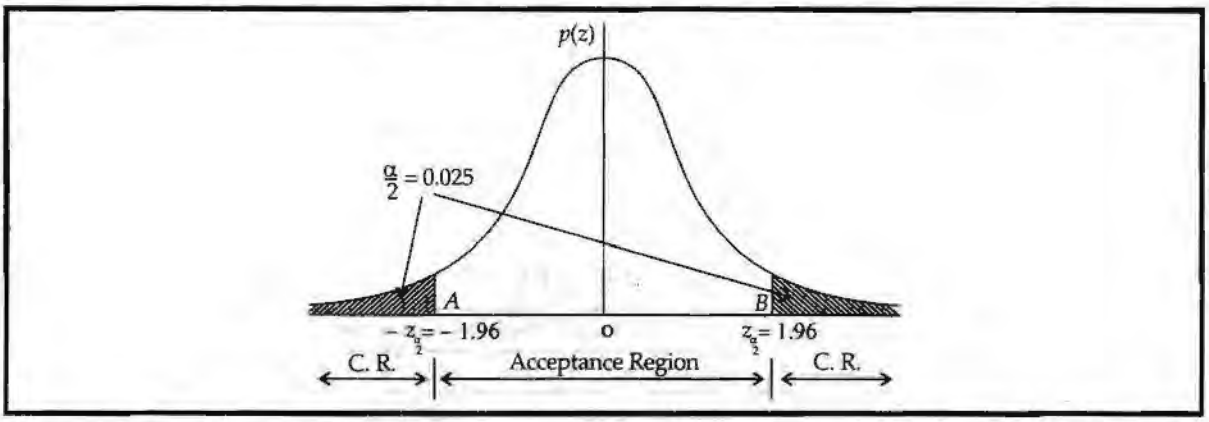


Figure 5.4

Critical Region and One Tailed versus Two Tailed Tests

The specification of the critical region for a test depends upon the nature of the alternative hypothesis and the value of α. For example, $H_a : \mu \neq \mu_0$, this implies that μ may be less or greater than μ_0. Thus, the critical region is to be specified on both tails of the curve with each part corresponding to half of the value of α. A test having critical region at both the tails of the probability curve is termed as a *two tailed test.*

Further, if $H_a : \mu > \mu_0$ or $\mu < \mu_0$, the critical region is to be specified only at one tail of the probability curve and the corresponding test is termed as a *one tailed test.* These situations are shown in the following figures.

The values of the random variable separating the acceptance region from critical region are termed as *critical value(s)*. For example, $z_{\alpha/2}$ and z_α, shown above, are critical values. Similarly, for a normal distribution the critical values for a two tailed test are – 1.96 and 1.96 for $\alpha = 0.05$ or – 2.58 and 2.58 for $\alpha = 0.01$ and the corresponding value for a one tailed test is ± 1.645 or ± 2.33 depending upon whether $\alpha = 0.05$ or 0.01.

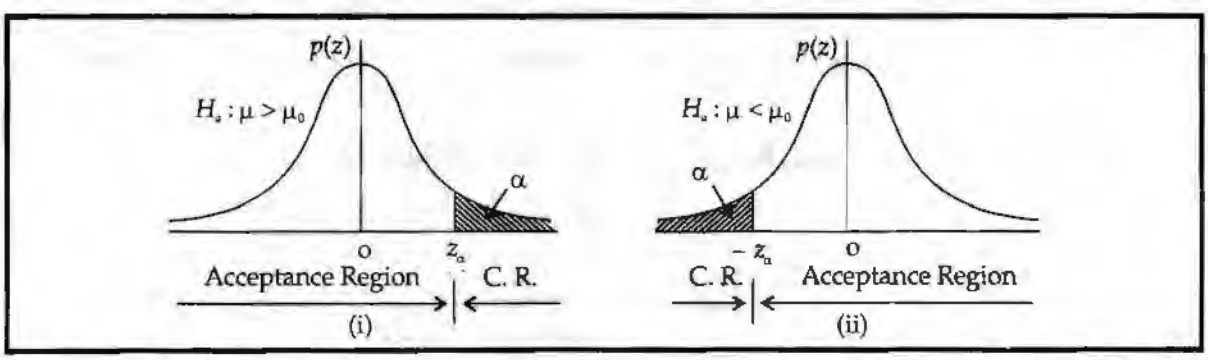


Figure 5.5

Acceptance Region

Remarks:

1. Out of the two types of errors, the type I error is considered to be more serious. Consequently, the probability of type I error is fixed at a low value (often 0.05 or lower). Thus, when the computed value of a statistic falls in the critical region, implying thereby that the probability of H_0 being true is low or equivalently the probability of H_0 being false is high, we reject H_0. However, if the computed value of statistics lies in the acceptance region, it would not be appropriate to say that the probability of H_0 being true is very high because the probability of accepting a false H_0 (the value of β) may also be high. Thus, accepting H_0 only implies that the sample information does not provide any evidence of H_0 being false. Because of this nature of the tests of hypothesis, the conclusion "accept H_0" is often replaced by "do not reject H_0" or "there is no evidence against H_0 on the basis of available sample information", etc.
2. The tests of hypothesis are also known as the *Tests of Significance*. We know that if the sample result is highly unlikely, H_0 is rejected because the sample result is significantly different from the hypothesised value. Alternatively, it implies that the observed difference between the computed and the hypothesised value is not attributable due to *chance or fluctuations of sampling*.

5.2 ONE SAMPLE TESTS

One Sample z Test

One sample tests, in general, examples of "goodness of fit" tests where we are testing whether our data supports predictions regarding the value of the population mean. This is most commonly represented as follows:

$$H_0 : \mu = \mu_0$$

where μ = actual true population mean

μ_0 = hypothesized population mean (under H_0)

In order to calculate a *z*-score, you would need to use the following formula:

$$Z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} \quad \text{where} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

Why do we use the standard error? Z is a test for the sample mean. To test Ho, we want to evaluate the probability or likelihood of getting our result, x, if Ho is true. To do this, we need to see where x falls on its sampling distribution. Now if the population standard deviation is unknown, but N is large ($N >= 30$), you can still use z, but you must substitute the sample standard deviation for the population standard deviation when calculating the standard error of the mean.

One Sample t Test

One sample t-test is a statistical procedure that is used to know the mean difference between the sample and the known value of the population mean. In one sample t-test, we know the population mean. In small samples ($N < 30$), sample standard deviations are biased estimates of their corresponding population standard deviations. In other words, σ does not estimate σ perfectly — it is usually too small. We draw a random sample from the population and then compare the sample mean with the population mean

and make a statistical decision as to whether or not the sample mean is different from the population. In one sample size, sample size should be less than 30. For example, we can use this when we take a sample from the city and we know the mean of the country (population mean). If we want to know whether the city mean differs from the country mean and we want to compare the two means, we will use the statistical test known as the one sample t-test.

Thus, we simply adjust the standard error using something called degrees of freedom (in this case, $df = N - 1$). Substituting these variations into the traditional z formula, we obtain the t test formula:

$$t_{df} = \frac{\bar{x} - \mu_0}{s_{\bar{x}}} \quad \text{where } s_{\bar{x}} = s/\sqrt{N-1}$$

One sample t-test and SPSS

In most of the statistical software, one sample t-test options are available. In SPSS, to perform the one sample t-test, we use the following procedures:

1. Click on the "SPSS 16" icon from the start menu.
2. Click on the "open data" icon and select the "data one sample t-test."
3. Click on the "analysis" option and select the "compare mean" option, from the analysis.
4. Select "one sample t-test" from the compare mean option. As we click on the one sample t-test, the window will appear and this window is called the one sample t-test window. Now, in this window, select the dependent variable and insert them into the test variable box. Type the population mean value in the test value box. Click on "option" and select the "percentage of confidence interval." As we click on the "ok" button, the result table for the one sample t-test will appear in front of us.

Check Your Progress 1

Fill in the blanks:

1. The .....describes how to use sample data to accept or reject the null hypothesis.
2. A type II error occurs when one rejects the .....
3. A ..... error is often referred to as a 'false positive'.

5.3 HYPOTHESES TESTING OF MEANS WHEN POPULATION STANDARD DEVIATION IS KNOWN

Test of Hypothesis Concerning the Equality of Standard Deviation (Small Samples)

We have to test $H_0 : \sigma_1 = \sigma_2$ against $\sigma_1 > \sigma_2$. The statistic $F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$, would become $\frac{s_1^2}{s_2^2}$ under H_0, follows F - distribution with ν_1 (= $n_1 - 1$) and ν_2 (= $n_2 - 1$) degrees of freedom.

Remarks:

1. We can write $s_1^2 = \frac{1}{n_1 - 1}\sum\left(X_{1i} - \bar{X}_1\right)^2 = \frac{n_1}{n_1 - 1}S_1^2 = \frac{1}{n_1 - 1}\left(\sum X_{1i}^2 - \frac{\sum X_{1i}^2}{n_1}\right)$ and

$$s_2^2 = \frac{1}{n_2 - 1}\sum\left(X_{2i} - \bar{X}_2\right)^2 = \frac{n_2}{n_2 - 1}S_2^2 = \frac{1}{n_2 - 1}\left(\sum X_{2i}^2 - \frac{\sum X_{2i}^2}{n_2}\right).$$

2. In the variance ratio $F = \frac{s_1^2}{s_2^2}$, we take, by convention the largest of the two sample variance as σ_1^2. Thus, this test is always a one tailed test with critical region at the right hand tail of the F - curve.

3. The $100(1 - \alpha)\%$ confidence limits for the variance ratio $\frac{\sigma_1^2}{\sigma_2^2}$, are given by

$$P\left[\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\alpha/2}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{1-\alpha/2}}\right] = 1 - \alpha.$$

Example 5.4: Two independent samples of sizes 10 and 12 from two normal populations have their mean square deviations about their respective means equal to 12.8 and 15.2 respectively. Test the equality of variances of the two populations.

Solution: We have to test $H_0 : \sigma_1 = \sigma_2$ against $\sigma_1 > \sigma_2$.

It is given that $S_1^2 = 15.2$, $S_2^2 = 12.8$, $n_1 = 12$ and $n_2 = 10$.

The unbiased estimates of respective population variances are

$$s_1^2 = \frac{12}{11} \times 15.2 = 16.58 \quad \text{and} \quad s_2^2 = \frac{10}{9} \times 12.8 = 14.22.$$

Thus, $F_{cal} = \frac{16.58}{14.22} = 1.166.$

The value of F from tables at 5% level of significance with 11 and 9 *d.f.* is 3.10. Since this value is greater than F_{cal}, there is no evidence against H_0.

Example 5.5: The increase in weight (in 100 gms) due to food A and food B given to two independent samples of children was recorded as follows. Test whether (i) mean weights and (ii) standard deviations of the two samples are equal.

Sample I : 6, 12, 10, 14, 12, 12, 10, 7, 5, 7.

Sample II : 9, 11, 8, 5, 6, 12, 7, 13, 10.

Solution: We shall first test $H_0 : \sigma_1 = \sigma_2$ against $\sigma_1 > \sigma_2$.

The means of the samples are $\bar{X}_1 = \frac{95}{10} = 9.5$ and $\bar{X}_2 = \frac{81}{9} = 9.0$, respectively.

We can write $s_k^2 = \frac{n_k}{n_k - 1}\left(\frac{\sum X_{ki}^2}{n_k} - \bar{X}_k^2\right) = \frac{\sum X_{ki}^2}{n_k - 1} - \frac{n_k}{n_k - 1}\bar{X}_k^2$ $(k = 1, 2)$

Thus, we have $s_1^2 = \frac{987}{9} - \frac{10}{9} \times 9.5^2 = 9.39$ and $s_2^2 = \frac{789}{8} - \frac{9}{8} \times 9^2 = 7.50.$

Further, the test statistic is $F = \frac{9.39}{7.50} = 1.25.$

The critical value of F at 5% level of significance and (9,8) *d.f.* is 3.39, therefore, there is no evidence against H_0. Hence, σ_1 and σ_2 may be treated as equal.

To test $H_0 : \mu_1 = \mu_2$ against $H_a ; \mu_1 \neq \mu_2$, we note that samples are small, t-test is to be used. Since $\sigma_1 = \sigma_2 = \sigma$(say), its unbiased estimate is

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{9 \times 9.39 + 8 \times 7.50}{10 + 9 - 2}} = 2.92.$$

The test statistic is $t_{cal} = \frac{|\bar{X}_1 - \bar{X}_2|}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{|9.5 - 9.0|}{2.92}\sqrt{\frac{10 \times 9}{10 + 9}} = 0.37.$

The critical value of t at 5% level of significance and 17 *d.f.* is 2.11. Since this value is greater than the calculated, there is no evidence against H_0. Thus, we conclude that the two samples may be regarded to have drawn from a population with same means and same standard deviations.

Test of Hypothesis Concerning Equality of Standard Deviations (Large Samples)

It can be shown that when sample sizes are large, i.e., n_1, $n_2 > 30$, the sampling distribution of the statistic $S_1 - S_2$ is approximately normal with mean $\sigma_1 - \sigma_2$ and standard error $\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$. Therefore

$$z = \frac{(S_1 - S_2) - (\sigma_1 - \sigma_2)}{\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}} \sim N(0,1)$$

or $z = \frac{S_1 - S_2}{\sigma\sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}}}$ under $H_0 : \sigma_1 = \sigma_2 = \sigma.$

Very often σ is not known and is estimated on the basis of sample. The pooled estimate of σ is $S = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2}$. Thus, the test statistic becomes

$$z_{cal} = \frac{S_1 - S_2}{S\sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}}} = \frac{S_1 - S_2}{S} \times \sqrt{\frac{2n_1 n_2}{n_1 + n_2}}.$$

Example 5.6: The standard deviation of a random sample of the heights of 500 individuals from country *A* was found to be 2.58 inches and that of 600 individuals from country *B* was found to be 2.35 inches. Do the data indicate that the standard deviation of heights in country *A* is greater than that in country *B*?

Solution: We have to test $H_0 : \sigma_1 = \sigma_2$ against $H_a : \sigma_1 > \sigma_2$.

It is given that $S_1 = 2.58$, $n_1 = 500$, $S_2 = 2.35$ and $n_2 = 600$.

The pooled estimate of σ is $S = \sqrt{\frac{500 \times 2.58^2 + 600 \times 2.35^2}{1100}} = 2.46$

The test statistic is $z_{cal} = \frac{2.58 - 2.35}{2.46} \times \sqrt{\frac{600000}{1100}} = 2.17$

Since this value is greater than 1.645, H_0 is rejected at 5% level of significance. Thus, the sample evidence indicates that the standard deviation of heights in country A is greater.

5.4 HYPOTHESES TESTING OF MEANS WHEN POPULATION STANDARD DEVIATION IS UNKNOWN

When σ is not known, we use its estimate computed from the given sample. Here, the nature of the sampling distribution of $\bar{X}$ would depend upon sample size n. There are the following two possibilities:

(i) If parent population is normal and $n < 30$ (popularly known as small sample case), use t - test. The unbiased estimate of σ in this case is given by $s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$.

Also, like normal test, the hypothesis may be one or two tailed.

(ii) If $n \geq 30$ (large sample case), use standard normal test. The unbiased estimate of σ in this case can be taken as $S = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n}}$, since the difference between n and $n - 1$ is negligible for large values of n. Note that the parent population may or may not be normal in this case.

Example 5.7: The yield of alfalfa from six test plots is 2.75, 5.25, 4.50, 2.50, 4.25 and 3.25 tonnes per hectare. Test at 5% level of significance whether this supports the contention that true average yield for this kind of alfalfa is 3.50 tonnes per hectare.

Solution: We note that σ is not given and $n = 6$ (< 30), $\therefore$ t - test is applicable.

Using sample information we have

$$\bar{X} = \frac{2.75 + 5.25 + 4.50 + 2.50 + 4.25 + 3.25}{6} = 3.75.$$

To calculate s, we define $u_i = \dfrac{X_i - 3.75}{0.25} = (X_i - 3.75) \times 4$

X_i	2.75	5.25	4.50	2.50	4.25	3.25
u_i	−4	6	3	−5	2	−2
u_i^2	16	36	9	25	4	4

From the above table $\sum u_i^2 = 94$. Therefore, $s = 0.25\sqrt{\dfrac{94}{6-1}} = 1.085$

We have to test $H_0 : \mu = 3.50$ against $H_a : \mu \neq 3.50$.

The test statistic $\dfrac{\bar{X} - \mu_0}{s/\sqrt{n}}$ ~ t - distribution with $(n - 1)$ *d.f.*

Thus, $t_{cal} = \left|\dfrac{3.75 - 3.50}{1.085/\sqrt{6}}\right| = 0.564$

Further, the critical value of t, from table at 5% level of significance and with 5 *d.f.* is 2.571. Since t_{cal} is less than this value, there is no evidence against at 5% level of significance.

Example 5.8: Daily sales figures of 40 shopkeepers showed that their average sales and standard deviation were Rs 528 and Rs 600 respectively. Is the assertion that daily sales on the average is Rs 400, contradicted at 5% level of significance by the sample?

Solution: Since $n > 30$, standard normal test is applicable. It is given that $n = 40$, $\bar{X} = 528$ and $S = 600$.

We have to test $H_0 : \mu = 400$ against $H_a : \mu \neq 400$.

$$z_{cal} = \left|\frac{528 - 400}{600/\sqrt{40}}\right| = 1.35.$$

Since this value is less than 1.96, there is no evidence against H_0 at 5% level of significance. Hence, the given assertion is not contradicted by the sample.

5.5 HYPOTHESIS TESTING OF PROPORTIONS FOR LARGE SAMPLES AND DIFFERENCE IN PROPORTIONS

Like the tests concerning sample mean, the null hypothesis to be tested would be either $\pi = \pi_0$, i.e., the proportion of successes in population is π_0 or $\pi_1 = \pi_2$, i.e., two populations have the same proportion of successes. These tests are based upon the sampling distribution of p, the proportion of successes in sample and the sampling distribution of $p_1 - p_2$, the difference between two sample proportions.

Test of Hypothesis that Population Proportion is π_0

The null hypothesis to be tested is $H_0 : \pi = \pi_0$ against $H_a : \pi \neq \pi_0$ for a two tailed test and $\pi >$ or $< \pi_0$ for a one tailed test. The test statistic is

$$z_{cal} = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = (p - \pi_0)\sqrt{\frac{n}{\pi_0(1-\pi_0)}}$$

Remarks: The 100(1 - α)% confidence limits for π are $p \pm z_{\alpha/2}$S.E.(p).

Example 5.9: A wholesaler in apples claims that only 4% of the apples supplied by him are defective. A random sample of 600 apples contained 36 defective apples. Test the claim of the wholesaler.

Solution: We have to test $H_0 : \pi \leq 0.04$ against $H_a : \pi > 0.04$.

It is given that $p = \frac{36}{600} = 0.06$ and n = 600.

$$\therefore \quad z_{cal} = (0.06 - 0.04)\sqrt{\frac{600}{0.04 \times 0.96}} = 2.5$$

This value is highly significant in comparison to 1.645, therefore, H_0 is rejected at 5% level of significance.

Example 5.10: The manufacturer of a spot remover claims that his product removes at least 90% of all spots. What can be concluded about his claim at the level of significance α = 0.05, if the spot remover removed only 174 of the 200 spots chosen at random from the spots on clothes brought to a dry cleaning establishment?

Solution: We have to test $H_0 : \pi \geq 0.9$ against $H_a : \pi < 0.9$.

It is given that $p = \frac{174}{200} = 0.82$ and n = 200.

$$\therefore \quad z_{cal} = (0.82 - 0.90)\sqrt{\frac{200}{0.9 \times 0.1}} = -3.77$$

Since this value is less than - 1.645, H_0 is rejected at 5% level of significance. Thus, the sample evidence does not support the claim of the manufacturer.

Example 5.11: 470 heads were obtained in 1,000 throws of an unbiased coin. Can the difference between the proportion of heads in sample and their proportion in population be regarded as due to fluctuations of sampling?

Solution: We have to test $H_0 : \pi = 0.5$ against $H_a : \pi \neq 0.5$.

It is given that $p = \frac{470}{1000} = 0.47$ and n = 1000.

$$\therefore \quad z_{cal} = |0.47 - 0.50|\sqrt{\frac{1000}{0.5 \times 0.5}} = 1.897.$$

Since this value is less than 1.96, the coin can be regarded as fair and thus, the difference between sample and population proportion of heads are only due to fluctuations of sampling.

Test of Hypothesis Concerning Equality of Proportions

The null hypothesis to be tested is $H_0 : \pi_1 = \pi_2$ against $H_a : \pi_1 \neq \pi_2$ for a two tailed test and $\pi_1 >$ or $< \pi_2$ for a one tailed test.

The test statistic is $z_{cal} = (p_1 - p_2)\sqrt{\frac{n_1 n_2}{\pi(1-\pi)(n_1+n_2)}}$ under the assumption that $\pi_1 = \pi_2 = \pi$, where π is known. Often population proportion π is unknown and it is estimated on the basis of samples. The pooled estimate of p, denoted by p, is given by $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$.

Thus, the test statistic becomes $z_{cal} = (p_1 - p_2)\sqrt{\frac{n_1 n_2}{p(1-p)(n_1+n_2)}}$.

Remarks: $100(1 - \alpha)\%$ confidence limits of $(\pi_1 - \pi_2)$ can be written as

$$(p_1 - p_2) \pm z_{\alpha/2}\,\text{S.E.}(p_1 - p_2)$$

Example 5.12: In a random sample of 600 persons from a large city, 450 are found to be smokers. In another sample of 900 persons from another large city, 450 are smokers. Do the data indicate that the cities are significantly different with respect to the prevalence of smoking? Let the level of significance be 5%.

Solution: We have to test $H_0 : \pi_1 = \pi_2$ against $H_a : \pi_1 \neq \pi_2$.

It is given that $n_1 = 600$, $n_2 = 900$, $X_1 = X_2 = 450$.

$$\therefore \quad p_1 = \frac{X_1}{n_1} = \frac{450}{600} = 0.75 \text{ and } p_2 = \frac{X_2}{n_2} = \frac{450}{900} = 0.50$$

The pooled estimate of π, i.e., $p = \frac{450 + 450}{600 + 900} = 0.6$

Thus, $z_{cal} = |0.75 - 0.50|\sqrt{\frac{600 \times 900}{0.6 \times 0.4 \times 1500}} = 9.682$

This value is highly significant, therefore, H_0 is rejected. Thus, the given samples indicate that the two cities are significantly different with regard to the prevalence of smoking.

Example 5.13: A company is considering two different television advertisements for the promotion of a new product. Management believes that advertisement *A* is more effective than advertisement *B*. Two test market areas with virtually identical consumer characteristics are selected ; advertisement *A* is used in one area and advertisement *B* is used in the other area. In a random sample of 60 customers who saw the advertisement *A*, 18 tried the product. In a random sample of 100 customers who saw advertisement *B*, 22 tried the product. Does this indicate that advertisement *A* is more effective than advertisement *B*, if a 5% level of significance is used?

Solution: We have to test $H_0 : \pi_A \leq \pi_B$ against $H_a : \pi_A > \pi_B$.

It is given that $n_A = 60$, $X_A = 18$, $n_B = 100$ and $X_B = 22$.

Thus, $p_A = \frac{18}{60} = 0.30$ and $p_B = \frac{22}{100} = 0.22$.

Also, the pooled estimate of π, i.e., $p = \frac{18+22}{160} = 0.25$.

$$\therefore \quad z_{cal} = (0.30 - 0.22)\sqrt{\frac{60 \times 100}{0.25 \times 0.75 \times 160}} = 1.131$$

Since this value is less than 1.645, there is no evidence against H_0 at 5% level of significance. Thus, the sample information provides no indication that advertisement A is more effective than advertisement B.

Remarks: As in the variable case, we can also test the hypothesis $\pi_1 = \pi_2 + k$. Since $\pi_1 \neq \pi_2$, pooling of proportions is not allowed for the computations of standard error of $p_1 - p_2$. The standard error in this case is

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

5.6 TWO SAMPLE TESTS FOR EQUALITY OF MEANS FOR LARGE AND SMALL SAMPLES

Equality of Means for Dependent Samples

If random samples are obtained from each of the two normal populations, the sampling distribution of the difference of their means is given by

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right).$$

Case I. If σ_1 and σ_2 are known, use standard normal test.

(a) To test $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$ (two tailed test), the test statistic is

$$z_{cal} = \frac{\left|(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\left|\bar{X}_1 - \bar{X}_2\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \text{ under } H_0.$$

This value is compared with 1.96 (2.58) for 5% (1%) level of significance.

(b) To test $H_0 : \mu_1 \leq \mu_2$ against $H_a : \mu_1 > \mu_2$ (one tailed test), the test statistic is $z_{cal} = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$, and the critical value for 5% (1%) level of significance is 1.645 (2.33).

(c) To test $H_0 : \mu_1 \geq \mu_2$ against $H_a : \mu_1 < \mu_2$ (one tailed test), the test statistic, i.e., z_{cal} is same as in (b) above, however, the critical value for 5% (or 1%) level of significance is – 1.645 (or – 2.33).

Case II. If σ_1 and σ_2 are not known, their estimates based on samples are used. This category of tests can be further divided into two sub-groups.

1. *Small Sample Tests* (when either n_1 or n_2 or both are less than or equal to 30). To test $H_0 : \mu_1 = \mu_2$, we use t - test. The respective estimates of σ_1 and σ_2 are given by

$$s_1 = \sqrt{\frac{\sum\left(X_{1i} - \bar{X}_1\right)^2}{n_1 - 1}} = S_1\sqrt{\frac{n_1}{n_1 - 1}} \text{ and } s_2 = \sqrt{\frac{\sum\left(X_{2i} - \bar{X}_2\right)^2}{n_2 - 1}} = S_2\sqrt{\frac{n_2}{n_2 - 1}}$$

This test is more restrictive because it is based on the assumption that the two samples are drawn from independent normal populations with equal standard deviations, i.e., $\sigma_1 = \sigma_2 = \sigma$ (say). The pooled estimate of σ, denotes by s, is defined as

$$s = \sqrt{\frac{\sum\left(X_{1i} - \bar{X}_1\right)^2 + \sum\left(X_{2i} - \bar{X}_2\right)^2}{n_1 + n_2 - 2}} = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

(a) To test $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$ (two tailed test), the test statistic is

$$t_{cal} = \frac{\left|\bar{X}_1 - \bar{X}_2\right|}{\sqrt{\dfrac{s^2}{n_1} + \dfrac{s^2}{n_2}}} = \frac{\left|\bar{X}_1 - \bar{X}_2\right|}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = \frac{\left|\bar{X}_1 - \bar{X}_2\right|}{s} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$, which follows t - distribution with $(n_1 + n_2 - 2)$ *d.f.*

This value is compared with the value of t from tables, to be denoted as $t_{\alpha/2}(n_1 + n_2 - 2)$, at $100\alpha\%$ level of significance with $(n_1 + n_2 - 2)$ *d.f.*

(b) To test $H_0 : \mu_1 \leq \mu_2$ against $H_a : \mu_1 > \mu_2$ (one tailed test), the test statistic is

$$t_{cal} = \frac{\left(\bar{X}_1 - \bar{X}_2\right)}{s} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$. This value is compared with $t_\alpha(n_1 + n_2 - 2)$ from tables.

(c) To test $H_0 : \mu_1 \geq \mu_2$ against $H_a : \mu_1 < \mu_2$ (one tailed test), the test statistic, i.e., t_{cal} is same as in (b) above. This value is compared with $- t_\alpha(n_1 + n_2 - 2)$.

2. *Large Sample Tests* (when both n_1 and n_2 is greater than 30)

In this case σ_1 and σ_2 are estimated by their respective sample standard deviations S_1 and S_2. The test statistics for two and one tailed tests are $z_{cal} = \dfrac{\left|\bar{X}_1 - \bar{X}_2\right|}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$ and $z_{cal} = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$ respectively.

The remaining procedure is same as in case I above.

Remarks:

1. $100(1-\alpha)\%$ confidence limits for $\mu_1 - \mu_1$ are given by $\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2} S.E_{(\bar{X}_1 - \bar{X}_2)}$.

   If $\bar{X}_1 - \bar{X}_2 \sim t$ - distribution, $z_{\alpha/2}$ is replaced by $t_{\alpha/2}(n_1 + n_2 - 2)$.

2. If the two sample are drawn from populations with same standard deviations, i.e., $\sigma_1 = \sigma_2 = \sigma$ (say), then $S.E_{(\bar{X}_1 - \bar{X}_2)} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ for problems covered under case I and $S.E_{(\bar{X}_1 - X_2)} = S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ for problems covered under case II, large sample tests. S is a pooled estimate of s, is given by

$$S = \sqrt{\frac{\sum(X_{1i} - \bar{X}_1)^2 + \sum(X_{2i} - \bar{X}_2)^2}{n_1 + n_2}} = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2}}$$

Example 5.14: An investigation of the relative merits of two kinds of flashlight batteries showed that a random sample of 100 batteries of brand X lasted on the average 36.5 hours with a standard deviation of 1.8 hours, while a random sample of 80 batteries of brand Y lasted on the average 36.8 hours with a standard deviation of 1.5 hours. Use a level of significance of 1% to test whether the observed difference between average life times is significant.

Solution: Let X and Y denote the life time of flashlight batteries of type X and type Y respectively and let μ_X and μ_Y be their respective population means.

It is given that $\bar{X} = 36.5$, $S_X = 1.8$, $n_X = 100$, $\bar{Y} = 36.8$, $S_Y = 1.5$, $n_Y = 80$.

We have to test $H_0 : \mu_X = \mu_Y$ against $H_a : \mu_X \neq \mu_Y$.

Since sample sizes are large (> 30), it is a large sample case.

The test statistic is $z_{cal} = \dfrac{|36.5 - 36.8|}{\sqrt{\dfrac{1.8^2}{100} + \dfrac{1.5^2}{80}}} = \dfrac{0.3}{0.246} = 1.219$

Since this value is less than 2.58, there is no evidence against H_0 at 1% level of significance and thus, the observed difference between average life times cannot be regarded as significant.

Example 5.15: Measurements performed on random samples of two kinds of cigarettes yielded the following results on their nicotine content (in mgs)

Brand A : 21.4, 23.6, 24.8, 22.4, 26.3

Brand B : 22.4, 27.7, 23.5, 29.1, 25.8

Assuming that the nicotine content is distributed normally, test the hypothesis that brand B has a higher nicotine content than brand A.

Solution: We have to test $H_0 : \mu_A \geq \mu_B$ against $H_a : \mu_A < \mu_B$.

Note that the rejection of H_0 would imply that brand B has a higher nicotine content than brand A.

The means of the two samples are

$$\bar{X}_A = \frac{21.4 + 23.6 + 24.8 + 22.4 + 26.3}{5} = 23.7$$

and $$\bar{X}_B = \frac{22.4 + 27.7 + 23.5 + 29.1 + 25.8}{5} = 25.7.$$

Also $$\sum (X_{Ai} - \bar{X}_A)^2 = 14.96 \text{ and } \sum (X_{Bi} - \bar{X}_B)^2 = 31.30$$

The pooled estimate of σ is $s = \sqrt{\frac{14.96 + 31.30}{5 + 5 - 2}} = 2.40$

Thus, the test statistic is $t_{cal} = \frac{(23.7 - 25.7)}{2.40} \times \sqrt{\frac{5 \times 5}{5 + 5}} = -1.318.$

The critical value of t at 5% level of significance and 8 *d.f.* is – 1.86. Since t_{cal} is greater than this value, it lies in the region of acceptance and hence, there is no evidence against at 5% level of significance. Thus, the nicotine content in brand *B* is not higher than in brand *A*.

Example 5.16: Two salesmen *A* and *B* are working in a certain district. From a sample survey conducted by the head office, the following results were obtained. State whether there is any significant difference in the average sales between the two salesmen:

	A	*B*
No. of Sales	20	18
Average Sales (in Rs)	170	205
Standard deviation (in Rs)	20	25

Solution: Since n_1, $n_2 < 30$, it is a small sample case.

We have to test $H_0 : \mu_A = \mu_B$ against $H_a : \mu_A \neq \mu_B$.

Assuming that the two samples have come from the same population with S.D. σ, we find its pooled estimate as

$$s = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{20 \times 20^2 + 18 \times 25^2}{36}} = 23.12$$

Also $t_{cal} = \frac{|170 - 205|}{23.12} \sqrt{\frac{20 \times 18}{20 + 18}} = 4.66$. This value is highly significant, therefore, H_0 is rejected at 5% level of significance.

Example 5.17: The mean life of a random sample of 10 light bulbs was found to be 1456 hours with a S.D. of 423 hours. A second sample of 17 bulbs chosen at random from a different batch showed a mean life of 1280 hours with S.D. of 398 hours. Is there a significant difference between the mean life of the two batches?

Solution: Note that the two samples have been obtained from the same population with unknown σ.

We have to test $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$.

It is given that $\bar{X}_1 = 1456$, $S_1 = 423$, $n_1 = 10$, $\bar{X}_2 = 1280$, $S_2 = 398$, $n_2 = 17$.

The pooled estimate of σ is $s = \sqrt{\dfrac{10 \times 423^2 + 17 \times 398^2}{10 + 17 - 2}} = 423.42$

Therefore $t_{cal} = \dfrac{|1456 - 1280|}{423.42} \times \sqrt{\dfrac{10 \times 17}{10 + 17}} = 1.04$

The value of t from table at 5% level of significance and with 25 *d.f.* is 2.06. Since t_{cal} is less than this value, there is no evidence against H_0. Hence, the observed difference in mean life of bulbs of the two batches can be regarded as due to fluctuations of sampling.

When the Hypothesized Difference is not Zero

Let H_0: $\mu_1 \leq \mu_2 + k$ against H_a: $\mu_1 > \mu_2 + k$, where k is constant. The above can also be written as.

H_0: $\mu_1 - \mu_2 \leq k$ against H_a: $\mu_1 - \mu_2 > k$

Thus we can write

$$\bar{X}_1 - \bar{X}_2 \sim N\left(k, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

or $Z_{cal} = \dfrac{|\bar{X}_1 - \bar{X}_2 - k|}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$ under H_0.

In a similar way, we can write the expressions for t_{cal} under different situations.

Example 5.18: A sample of 100 electric bulbs of 'Philips' gave a mean life of 1500 hours with a standard deviation of 60 hours. Another sample of 100 electric bulbs of "HMT" gave a mean life of 1615 hours with a standard deviation of 80 hours. Can we conclude that the mean life of 'HMT' bulbs is greater then that of 'Philips' bulbs by 100 hours?

Let $\bar{X}_1 = 1615$, $S_1 = 80$, $n_1 = 100$, $\bar{X}_2 = 1500$, $S_2 = 60$, $n_2 = 100$.

We can write

H_0: $\mu_1 \leq \mu_2 + 100$ against H_a: $\mu_1 > \mu_2 + 100$

$$Z_{cal} = \frac{|1615 - 1500 - 100|}{\sqrt{\dfrac{80^2}{100} + \dfrac{60^2}{100}}} = 1.5$$

Since $Z_{cal} < 1.645$, we accept H_0 at 5% and say that the difference in mean life of 'HMT' bulbs and that of 'Philips' bulbs is less than or equal to 100 hours.

Difference between Proportions for Large Samples

Suppose we have two populations with proportions equal to P_1 and P_2. Suppose further that we take all possible samples of size n_1 and n_2. And finally, suppose that the following assumptions are valid.

- The size of each population is large relative to the sample drawn from the population. That is, N_1 is large relative to n_1, and N_2 is large relative to n_2. (In this context, populations are considered to be large if they are at least 10 times bigger than their sample.)
- The samples from each population are big enough to justify using a normal distribution to model differences between proportions. The sample sizes will be big enough when the following conditions are met: $P_1 \geq 10, n_1(1-P_1) \geq 10, n_2P_2 \geq 10,$ *and* $n_2(1-P_2) \geq 10.$
- The samples are independent; that is, observations in population 1 are not affected by observations in population 2, and vice versa.

Given these assumptions, we know the following.

- The set of differences between sample proportions will be normally distributed. We know this from the central limit theorem.
- The expected value of the difference between all possible sample proportions is equal to the difference between population proportions. Thus, $E(p_1 - p_2) = P_1 - P_2$.
- The standard deviation of the difference between sample proportions (σ_d) is approximately equal to:

$$\sigma_d = \text{sqrt}\{ [P_1(1 - P_1) / n_1] + [P_2(1 - P_2) / n_2] \}$$

It is straightforward to derive the last bullet point, based on material covered in previous lessons. The derivation starts with a recognition that the variance of the difference between independent random variables is equal to the sum of the individual variances. Thus,

$$\sigma_d^2 = \sigma^2_{P1-P2} = \sigma_1^2 + \sigma_2^2$$

If the populations N_1 and N_2 are both large relative to n_1 and n_2, respectively, then

$$\sigma_1^2 = P_1(1 - P_1) / n_1 \qquad \text{And} \qquad \sigma_2^2 = P_2(1 - P_2) / n_2$$

Therefore,

$$\sigma_d^2 = [P_1(1 - P_1) / n_1] + [P_2(1 - P_2) / n_2] \quad \text{And}$$
$$\sigma_d = \text{sqrt}\{[P_1(1 - P_1) / n_1] + [P_2(1 - P_2) / n_2]\}$$

Example 5.19: Suppose the Cartoon Network conducts a nation-wide survey to assess viewer attitudes toward Superman. Using a simple random sample, they select 400 boys and 300 girls to participate in the study. Forty percent of the boys say that Superman is their favorite character, compared to thirty percent of the girls. What is the 90% confidence interval for the true difference in attitudes toward Superman?

(A) 0 to 20 percent more boys prefer Superman

(B) 2 to 18 percent more boys prefer Superman

(C) 4 to 16 percent more boys prefer Superman

(D) 6 to 14 percent more boys prefer Superman

(E) None of the above

Solution

The correct answer is (C). The approach that we used to solve this problem is valid when the following conditions are met.

- The sampling method must be simple random sampling. This condition is satisfied; the problem statement says that we used simple random sampling.
- Both samples should be independent. This condition is satisfied since neither sample was affected by responses of the other sample.
- The sample should include at least 10 successes and 10 failures. Suppose we classify choosing Superman as a success, and any other response as a failure. Then, we have plenty of successes and failures in both samples.
- The sampling distribution should be approximately normally distributed. Because each sample size is large, we know from the central limit theorem that the sampling distribution of the difference between sample proportions will be normal or nearly normal; so this condition is satisfied.

Since the above requirements are satisfied, we can use the following four-step approach to construct a confidence interval.

- Identify a sample statistic. Since we are trying to estimate the difference between population proportions, we choose the difference between sample proportions as the sample statistic. Thus, the sample statistic is $p_{boy} - p_{girl} = 0.40 - 0.30 = 0.10$.
- Select a confidence level. In this analysis, the confidence level is defined for us in the problem. We are working with a 90% confidence level.
- Find the margin of error. Elsewhere on this site, we show how to compute the margin of error when the sampling distribution is approximately normal. The key steps are shown below.
- Find standard deviation or standard error. Since we do not know the population proportions, we cannot compute the standard deviation; instead, we compute the standard error. And since each population is more than 10 times larger than its sample, we can use the following formula to compute the standard error (SE) of the difference between proportions:

  $SE = sqrt\{ [p_1 * (1 - p_1) / n_1] + [p_2 * (1 - p_2) / n_2] \}$

  $SE = sqrt\{ [0.40 * 0.60 / 400] + [0.30 * 0.70 / 300] \}$

  $SE\ sqrt[(0.24 / 400) + (0.21 / 300)] = sqrt(0.0006 + 0.0007) = sqrt(0.0013) = 0.036$

- Find critical value. The critical value is a factor used to compute the margin of error. Because the sampling distribution is approximately normal and the sample sizes are large, we can express the critical value as a z score by following these steps.
  - Compute alpha (α): $\alpha = 1 - (\text{confidence level} / 100) = 1 - (90/100) = 0.10$
  - Find the critical probability (p^*): $p^* = 1 - \alpha/2 = 1 - 0.10/2 = 0.95$
  - The critical value is the z score having a cumulative probability equal to 0.95. From the Normal Distribution Calculator, we find that the critical value is 1.645.

- Compute margin of error (ME): ME = critical value * standard error = 1.645 * 0.036 = 0.06
- Specify the confidence interval. The range of the confidence interval is defined by the sample statistic ± margin of error. And the uncertainty is denoted by the confidence level.

Therefore, the 90% confidence interval is 0.04 to 0.16. That is, we are 90% confident that the true difference between population proportions is in the range defined by 0.10 + 0.06. Since both ends of the confidence interval are positive, we can conclude that more boys than girls choose Superman as their favorite cartoon character.

Check Your Progress 2

Fill in the blanks:

1. If s_1 and s_2 are not known, their ..... based on samples are used.
2. Often population proportion p is unknown and it is estimated on the basis of .....
3. If s_1 and s_2 are known, use .....

5.7 SUMMARY

1. Population Mean
   (i) Compute, where when s is known or when s is unknown but n > 30.
   (a) Reject H_0 if $|z| > z_a/2$ when $H_0 : m = m_0$ against $H_a : m \neq m_0$
   (b) Reject H_0 if $z > z_a$ when $H_0 : m \leq m_0$ against $H_a : m > m_0$
   (c) Reject H_0 if $z < -z_a$ when $H_0 : m \geq m_0$ against $H_a : m < m_0$
   (ii) Compute, when s is unknown but $n \leq 30$.
   (a) Reject H_0 if $|t| > t_a/2, (n-1)$ when $H_0 : m = m_0$ against $H_a : m \neq m_0$
   (b) Reject H_0 if $t > t_a(n-1)$ when $H_0 : m \leq m_0$ against $H_a : m > m_0$
   (c) Reject H_0 if $t < -t_a(n-1)$ when $H_0 : m \geq m_0$ against $H_a : m < m_0$
2. Difference of Means
   (i) Compute, where when s_1 and s_2 are known or when s_1 and s_2 are unknown but n_1, n_2 >30.
   (a) Reject H_0 if $|z| > z_a/2$ when $H_0 : m_1 = m_2$ against $H_a : m_1 \neq m_2$
   (b) Reject H_0 if $z > z_a$ when $H_0 : m_1 \leq m_2$ against $H_a : m_1 > m_2$
   (c) Reject H_0 if $z < -z_a$ when $H_0 : m_1 \geq m_2$ against $H_a : m_1 < m_2$
   (ii) Compute, where and or compute (paired t-test), where, when s_1 and s_2 are unknown but n_1, $n_2 \leq 30$
   (a) Reject H_0 if $|t| > t_a/2, (n-1)$ when $H_0 : m_1 = m_2$ against $H_a : m \neq m_2$
   (b) Reject H_0 if $t > t_a, (n-1)$ when $H_0 : m_1 \leq m_2$ against $H_a : m_1 > m_2$
   (c) Reject H_0 if $t < -t_a, (n-1)$ when $H_0 : m_1 \geq m_2$ against $H_a : m_1 < m_2$

3. Proportion

Compute, where np_0, $n(1 - p_0) > 5$.

(a) Reject H_0 if $|z| > z_a/2$ when $H_0 : p = p_0$ against $H_a : p \neq p_0$

(b) Reject H_0 if $z > z_a$ when $H_0 : p \leq p_0$ against $H_a : p > p_0$

(c) Reject H_0 if $z < -z_a$ when $H_0 : p \geq p_0$ against $H_a : p < p_0$

4. Difference of Proportions

Compute, where, and n_1p_1, n_1q_1, n_2p_2, $n_2q_2 > 5$.

(a) Reject H_0 if $|z| > z_a/2$ when $H_0 : p_1 = p_2$ against $H_a : p_1 \neq p_2$

(b) Reject H_0 if $z > z_a$ when $H_0 : p_1 \leq p_2$ against $H_a : p_1 > p_2$

(c) Reject H_0 if $z < -z_a$ when $H_0 : p_1 \geq p_2$ against $H_a : p_1 < p_2$

5. Standard Deviation

(i) Compute, when $n \leq 30$.

(a) Reject H_0 if $c^2 > c^20.5a$, $(n - 1)$ (or $< c^2(1 - 0.5a)$) when $H_0 : s = s_0$ against $H_a : s \neq s_0$

(b) Reject H_0 if $c^2 > c^2a$, $(n - 1)$ when $H_0 : s \leq s_0$ against $H_a : s > s_0$

(c) Reject H_0 if $c^2 < c^2a$, $(n - 1)$ when $H_0 : s \geq s_0$ against $H_a : s < s_0$

(ii) Compute, when $n > 30$.

(a) Reject H_0 if $|z| > z_a/2$ when $H_0 : s = s_0$ against $H_a : s \neq s_0$

(b) Reject H_0 if $z > z_a$ when $H_0 : s \leq s_0$ against $H_a : s > s_0$

(c) Reject H_0 if $z < -z_a$ when $H_0 : s \geq s_0$ against $H_a : s < s_0$

5.8 KEYWORDS

- Hypothesis
- Two-tailed
- Analysis plan
- Significance level
- Type II error
- One-tailed
- Null hypothesis
- P-value
- Type I error

5.9 REVIEW QUESTIONS

1. A manufacturer of ball-point refills claims that the manufactured refills have a mean life of 40 pages with a S.D. of 2 pages. A purchasing agent selects a sample of 100 pens and puts them for test. The mean writing life for the sample was found to be 39 pages. Should the purchasing agent reject the manufacturer's claim at 5% level of significance?

2. Certain motor oil is packed in tins holding 5 litres each. The filling machine can maintain this but with a S.D. of 0.15 litre. Two samples of 36 tons each are taken from the production line. If the

sample means are 5.20 and 4.95 litres respectively, can we be 99% sure that the sample have come from a population of 5 liters?

3. A company selects 9 salesmen at random and their sales figures for the previous month are recorded. These salesmen then undergo a course devised by a business consultant and their sales figures for the following month are compared as shown in the following table. Has the training course caused an improvement in the salesmen's ability? Let the level of significance be 5%.

4. A random sample of 12 families in one city showed an average monthly food expenditure of Rs 1,380 with a S.D. of Rs 100 and a random sample of 15 families in another city showed an average monthly food expenditure of Rs 1,320 with a S.D. of Rs 120. Test whether the difference between the two means is significant at 5% level of significance?

5. A normal population is supposed to have a mean of 3 cms and a S.D. of 2.31 cms. A sample of 900 members is found to have a mean of 3.24 cms. Can it be regarded as a simple random sample drawn from this population?

6. A radio shop sells, on an average, 200 radios per day with a standard deviation of 50 radios. After an extensive advertising campaign, the management will compute the average sales for the next 25 days to see whether an improvement has occurred. Assume that the daily sales of radios are normally distributed.

   (i) Write down the null and alternate hypothesis.

   (ii) Test the hypothesis at 5% level of significance if = 216.

   (iii) How large must be in order that the null hypothesis is rejected at 5% level of significance?

7. A firm found with the help of a sample survey of a city (sample size 900) that 3/4th of the population consumes things produced by it. The firm then advertised the goods in newspaper and radio. After one year, a sample of size 1,000 reveals that the proportion of consumers of the goods produced by the firm is 4/5. Is this rise significant to indicate that the advertisement was effective?

8. In a year, there are 956 births in town *A*, of which 52.5% were males while in town *A* and *B* combined, this proportion in a total of 1406 births was 0.496. Is there a significant difference in the proportion of male births in the two towns at 5% level of significance?

9. Sky Packets guarantee that 90% of their deliveries are on time. In a recent week, 81 deliveries were made of which 6 were late. Sky Packet's Managing Director says with 95% confidence that there has been a significant improvement in deliveries. Should the Managing Director's statement be accepted?

10. A company has its head office at Calcutta and a branch office at Mumbai. The personal director wanted to know if the workers at the two places would like the introduction of a new plan of work and consequently a survey was conducted for this purpose. Out of a sample of 500 workers at Calcutta, 62% favoured the new plan. At Mumbai, out of a sample of 400 workers, 41% were against the new plan. Is there any significant difference between the two groups in their attitudes towards the new plan at 5% level of significance?

11. A stock broker claims that he can predict with 80% accuracy whether the value of a stock will rise or fall during the coming month. As a test he predicts the outcome of 40 stocks and is correct in 28 of the predictions. Does this evidence support the stock broker's claim?

12. Two independent random samples, one of 12 observations with mean 15 and sum of squares of deviations from mean equal to 135 and another of 16 observations with mean 22 and sum of

squares of deviations from mean equal to 250, were obtained from two normal populations. Test at 5% level of significance whether the two samples can be regarded to have come from the same population?

13. Test the hypothesis that $s = 8$, given that $S = 10$ for a random sample of size 51. Also construct 95% confidence interval for s.
14. The standard deviation of a random sample of 25 observations from a normal population was computed to be 8.
    (i) Test the hypothesis that population standard deviation is 12.
    (ii) Construct a 95% confidence interval for the population variance.
15. A random sample of 20 observations shows a standard deviation of 6.
    (i) Test the hypothesis that population standard deviation is greater than 7.
    (ii) Obtain 95% fiducial limits of population standard deviation.

Answers to Check Your Progress

Check Your Progress 1

1. Analysis plan
2. Alternative hypothesis
3. Type I

Check Your Progress 2

1. Estimates
2. Samples
3. Standard normal test

5.10 REFERENCES AND FURTHER READING

- Frost, J. (2019). Statistics by Jim (3rd ed.). Statistics By Jim Publishing. Retrieved from https://www.statisticsbyjim.com
- Pallant, J. (2020). SPSS survival manual (7th ed.). Open University Press.
- Vogt, W. P., & Johnson, R. (2021). Dictionary of statistics & methodology: A nontechnical guide for the social sciences (4th ed.). Sage Publications.
- Gravetter, F. J., & Wallnau, L. B. (2022). Statistics for The Behavioral Sciences (10th ed.). Cengage Learning.

UNIT

6

Chi Square

CHAPTER OUTLINE

6.1 CHI-SQUARE TEST OF INDEPENDENCE

We know that if X is a random variate distributed normally with mean μ and standard deviation σ, then $z = \frac{X - \mu}{\sigma}$ is a standard normal variate. Square of z, *i.e.*, $z^2 = \frac{(X - \mu)^2}{\sigma^2}$ if distributed as χ^2 - variate with one degree of freedom and is written as χ_1^2. Further, the value of χ_1^2, a squared value, will lie between 0 to ∞, for z lying between $-\infty$ to ∞. Since most of the z-values are close to zero, the probability density of χ^2 will be highest near zero. The χ^2 distribution with one degree of freedom is shown in Figure 6.1.

Generalising the above result, we can say that if $X_1, X_2, \ldots\ldots X_n$ are n independent normal variates each with mean μ_i and standard deviations σ_i, $i = 1, 2, \ldots\ldots n$, respectively, then the sum of squares $\sum z_i^2 = \sum \frac{(X_i - \mu_i)^2}{\sigma_i^2}$ is a χ^2 variate with n degrees of freedom, i.e., χ_n^2. Thus, we can say that χ_n^2 is sum of squares of n independent standard normal variate.

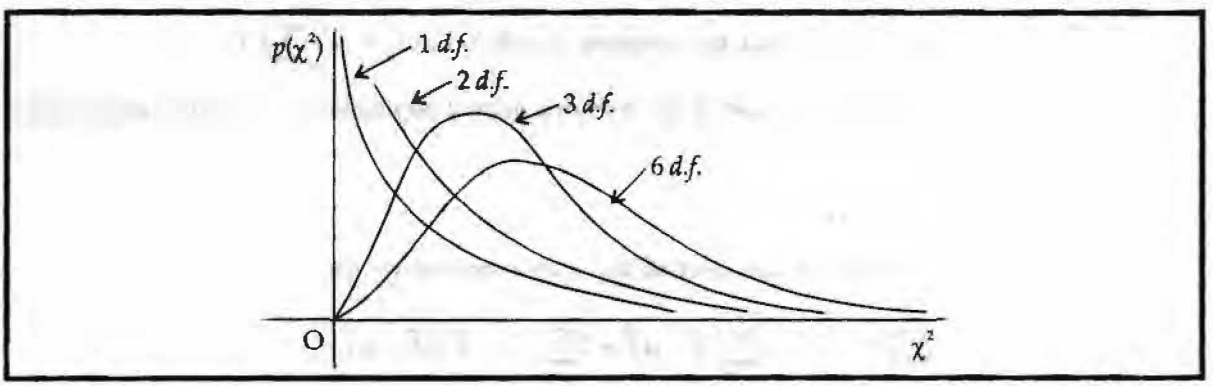


Figure 6.1

Chi-square Test of Independence

Features of χ^2 Distribution

1. The distribution has only one parameter, i.e., number of degrees of freedom or *d.f.* (in abbreviated form) which is a positive integer.
2. We may note that as the *d.f.* increases, the height of the probability density function decreases. The distribution is positively skewed and the skewness decreases as *d.f.* increases. For large values of *d.f.*, the distribution approaches normal distribution. The curves for various *d.f.* are shown in Figure 6.1.
3. The mean of χ_n^2, i.e., $E(\chi_n^2) = n$ and its variance = $2n$, where n = *d.f.*
4. *Additive property*

   The sum of two independent χ^2 variates is also a χ^2 variate with degrees of freedom equal to the sum of their individual degrees of freedom.

If χ_n^2 and χ_m^2 are two independent random variates with n and m degrees of freedom respectively, then $\chi_n^2 + \chi_m^2$ is also a χ^2 variate with $n + m$ degrees of freedom.

Remarks:

1. The degrees of freedom is defined as the number of independent random variables. If n is the number of variables and k is the number of restrictions on them, the degrees of freedom are said to be $n - k$.

2. On the basis of the definition of degrees of freedom, given above, we can say that $\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2$ is a χ^2 variate with $(n - 1)$ degrees of freedom. It may be pointed out here that one degree of freedom is reduced because for a given value of $\bar{X}$, the number of independent variables is $(n - 1)$.

Sampling Distribution of Variance

Using χ^2-distribution, we can construct the sampling distribution of $S^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$.

Let $X_1, X_2 \ldots\ldots X_n$ be a random sample of size n from a normal population with mean μ and variance σ^2. We can write

$$X_i - \mu = (X_i - \bar{X}) + (\bar{X} - \mu)$$

Squaring both sides and taking sum over all the n observations, we get

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(\bar{X} - \mu)^2 + 2\sum_{i=1}^{n}(X_i - \bar{X})(\bar{X} - \mu)$$

$$= \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2 + 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \bar{X})$$

We note that the last term is zero. Therefore, we have

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$$

Dividing both sides by σ^2, we get

$$\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} + \frac{n(\bar{X} - \mu)^2}{\sigma^2}$$

or $$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2} - \frac{(\bar{X} - \mu)^2}{\sigma^2/n} = \chi_n^2 - \chi_1^2 = \chi_{n-1}^2$$

Thus, $\frac{\sum(X_i - \bar{X})^2}{\sigma^2}$ or $\frac{nS^2}{\sigma^2}$ is a χ^2-variate with $(n - 1)$ *d.f.*

Mean and Standard Error of S^2

Since the random variable $\frac{nS^2}{\sigma^2}$ is a χ^2-variate with $(n-1)$ *d.f.*,

therefore $E\left[\frac{nS^2}{\sigma^2}\right]=n-1$ or $\frac{n}{\sigma^2}E\left(S^2\right)=n-1$.

Thus, we have $E\left(S^2\right)=\frac{n-1}{n}\cdot\sigma^2$

Further, if we define $s^2=\frac{1}{n-1}\sum\left(X_i-\bar{X}\right)^2$ so that $s^2=\frac{n}{n-1}\cdot S^2$, we have

$E\left(s^2\right)=\frac{n}{n-1}\cdot E\left(S^2\right)=\frac{n}{n-1}\cdot\frac{n-1}{n}\cdot\sigma^2=\sigma^2$ (See Remarks 2 below).

To find variance of S^2, we make use of the fact that variance of $\frac{nS^2}{\sigma^2}$ is $2(n-1)$. This implies that

$$E\left[\frac{nS^2}{\sigma^2}-(n-1)\right]^2=2(n-1) \text{ or } \frac{n^2}{\sigma^4}E\left(S^2-\frac{n-1}{n}\cdot\sigma^2\right)^2=2(n-1)$$

$$\therefore\ E\left[S^2-E\left(S^2\right)\right]^2=\frac{2(n-1)}{n^2}\cdot\sigma^4 \text{ or } Var\left(S^2\right)=\frac{2(n-1)}{n^2}\cdot\sigma^4$$

Further, variance of s^2 = variance of $\left(\frac{n}{n-1}\cdot S^2\right)$. This gives

$$Var\left(s^2\right)=\frac{n^2}{(n-1)^2}\cdot Var\left(S^2\right)=\frac{n^2}{(n-1)^2}\times\frac{2(n-1)}{n^2}\cdot\sigma^4=\frac{2}{n-1}\cdot\sigma^4$$

Remarks:

1. The distributions of χ^2 and S^2 are based upon the assumption that the parent population is normal. If the parent population is not normal, it is not possible to comment upon the nature of the distribution of the above statistics.
2. It will be discussed in the following chapter that when expected value of a statistic equals the value of parameter, it is said to be an unbiased estimate of the parameter.

6.2 THE STUDENT'S T-DISTRIBUTION

Let X_1, X_2 ..... X_n be n independent random variables from a normal population with mean m and standard deviation s (unknown).

When s is not known, it is estimated by s, the sample standard deviation $\left(s=\sqrt{\frac{1}{n-1}\sum\left(X_i-\bar{X}\right)^2}\right)$.

In such a case we would like to know the exact distribution of the statistic $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ and the answer to this is provided by t - distribution.

W.S. Gosset defined t statistic as $t=\frac{\bar{X}-\mu}{s/\sqrt{n}}$ which follows t - distribution with (n - 1) degrees of freedom.

Features of t- distribution

1. Like χ^2 - distribution, t - distribution also has one parameter ν = n - 1, where n denotes sample size. Hence, this distribution is known if n is known.

2. Mean of the random variable t is zero and standard deviation is $\sqrt{\frac{\nu}{\nu-2}}$, for $\nu > 2$.

3. The probability curve of t - distribution is symmetrical about the ordinate at t = 0. Like a normal variable, the t variable can take any value from - ∞ to ∞.

4. The distribution approaches normal distribution as the number of degrees of freedom become large.

5. The random variate t is defined as the ratio of a standard normal variate to the square root of χ^2 - variate divided by its degrees of freedom.

   To show this we can write $t=\frac{\bar{X}-\mu}{s/\sqrt{n}}=\frac{(\bar{X}-\mu)\sqrt{n}}{s}$

   Dividing numerator and denominator by s, we get

$$t=\frac{\frac{(\bar{X}-\mu)\sqrt{n}}{\sigma}}{\frac{s}{\sigma}}=\frac{\frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}}}{\sqrt{s^2/\sigma^2}}=\frac{\frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}}}{\sqrt{\frac{1}{n-1}\cdot\frac{\sum(X_i-\bar{X})^2}{\sigma^2}}}$$

$$=\frac{\frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}}}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}=\frac{Standard\ Normal\ Variate}{\sqrt{\chi^2\text{-}variate}}$$

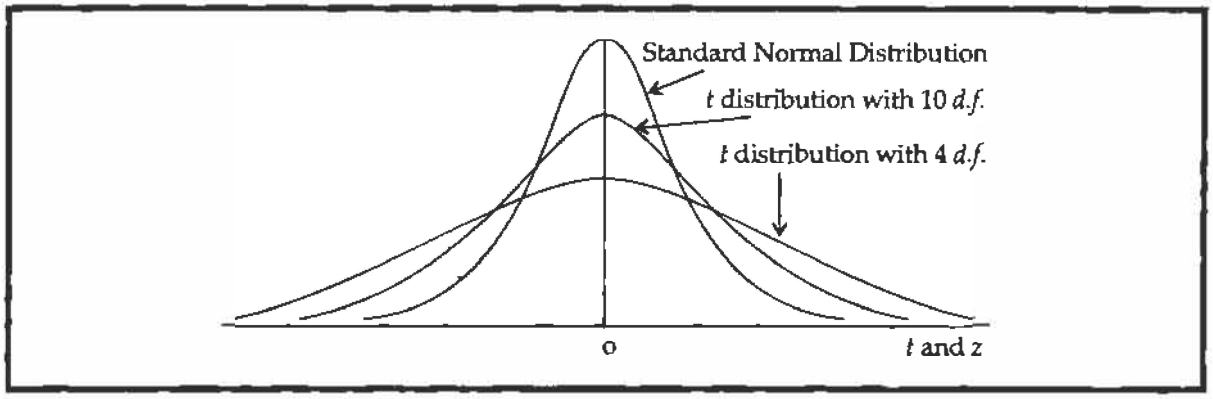


Figure 6.2

t-distribution

6.3 SNEDECOR'S F- DISTRIBUTION

Let there be two independent random samples of sizes n_1 and n_2 from two normal populations with variances σ_1^2 and σ_2^2 respectively. Further, let $s_1^2 = \frac{1}{n_1 - 1}\sum(X_{1i} - \bar{X}_1)^2$ and $s_2^2 = \frac{1}{n_2 - 1}\sum(X_{2i} - \bar{X}_2)^2$ be the variances of the first sample and the second samples respectively. Then F - statistic is defined as the ratio of two χ^2 - variates. Thus, we can write

$$F = \frac{\frac{\chi^2_{n_1-1}}{n_1 - 1}}{\frac{\chi^2_{n_2-1}}{n_2 - 1}} = \frac{\frac{(n_1-1)s_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)s_2^2}{\sigma_2^2}/(n_2-1)} = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}}$$

Check Your Progress 1

Fill in the blanks:

1. The distribution has only one parameter, i.e., number of ..... which is a positive integer.
2. The degrees of freedom are defined as the number of .....variables.

6.4 CHI-SQUARE TEST

Uses of χ^2 Test

In addition to the use of χ^2 in tests of hypothesis concerning the standard deviation, it is used as a test of goodness of fit and as a rest of independence of two attributes. These tests are explained in the following secrions.

χ^2 - test as a Goodness of Fit

χ^2 -test can be used to test, how far the fitted or the expected frequencies are in agreement with the observed frequencies. We know that for large values of n, the sampling distribution of X, the number of successes, is normal with mean np and variance $n\pi(1-\pi)$. Thus, $z = \dfrac{X - n\pi}{\sqrt{n\pi(1-\pi)}} \sim N(0,1)$.

Further, square of z is a χ^2 - variate with one degree of freedom. We can write

$$z^2 = \frac{(X-n\pi)^2}{n\pi(1-\pi)} = (X-n\pi)^2\left[\frac{(1-\pi)+\pi}{n\pi(1-\pi)}\right] \qquad (\because 1 = 1-\pi+\pi)$$

$$= (X-n\pi)^2\left[\frac{1}{n\pi} + \frac{1}{n(1-\pi)}\right] = \frac{(X-n\pi)^2}{n\pi} + \frac{(X-n\pi)^2}{n(1-\pi)} \qquad \text{.....(1)}$$

We can write $\dfrac{(X-n\pi)^2}{n(1-\pi)} = \dfrac{(X-n+n-n\pi)^2}{n(1-\pi)} = \dfrac{\left[(X-n)+n(1-\pi)\right]^2}{n(1-\pi)}$

$$= \frac{\left[(n-X)-n(1-\pi)\right]^2}{n(1-\pi)} = \frac{\left[(n-X)-E(n-X)\right]^2}{E(n-X)}$$

Similarly $\dfrac{(X-n\pi)^2}{n\pi} = \dfrac{\left[X-E(X)\right]^2}{E(X)}$

Thus, equation (1) can be written as $z^2 = \dfrac{\left[X-E(X)\right]^2}{E(X)} + \dfrac{\left[(n-X)-E(n-X)\right]^2}{E(n-X)}$

Here X denotes the observed number of successes and $(n - X)$ the observed number of failures.

Let O_1, E_1 denote the observed and expected number of successes respectively and O_2, E_2 denote the observed and expected number of failures respectively.

$\therefore \quad z^2 = \dfrac{(O_1-E_1)^2}{E_1} + \dfrac{(O_2-E_2)^2}{E_2}$ is a χ^2 - variate with 1 *d.f.*

Also we note that $O_1 + O_2 = E_1 + E_2 = n$.

The above result can be generalised for a manifold classification. If a population is divided into k mutually exclusive classes with observed and expected frequencies as $O_1, O_2, \ldots\ldots O_k$ and $E_1, E_2, \ldots\ldots E_k$ respectively, then $\chi^2 = \sum_{i=1}^{k} \dfrac{(O_i-E_i)^2}{E_i}$ is a χ^2-variate with $(k - 1)$ *d.f.* Here again we have $\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i = N$ (total frequency).

Remarks:

1. The above result would hold, approximately, for any type of population provided E_i's are sufficiently large. The approximation is satisfactory if each $E_i \geq 10$, for $k = 2$ and each $E_i \geq 5$ for $k > 2$. The difficulty of small cell frequencies can also be overcome by merging adjoining classes. The degrees of freedom would depend upon the number of classes after regrouping.
2. When there is perfect agreement between the observed and expected frequencies (to be taken as H_0), i.e., each O_i is equal to E_i, $i = 1, 2, \ldots\ldots k$, the value of $\chi^2 = 0$. This implies that the null hypothesis can be rejected only if the value of χ^2 statistic is significantly large. Thus, the tests of goodness of fit are only one tailed tests with critical region lying in right hand tail of the χ^2-distribution.
3. When the population is not completely specified, i.e., one or more of its parameters are to be estimated from the sample, the degrees of freedom of χ^2 would be further reduced by the number of parameters estimated.
4. When the degrees of freedom are greater than 30, the sampling distribution of $\sqrt{2\chi^2}$ is normal with mean $\sqrt{2n}$ and standard error equal to unity.
5. Alternatively, the χ^2 statistic, given above, can be written as

$$\chi^2 = \sum_{i=1}^{k} \frac{O_i^2 - 2O_iE_i + E_i^2}{E_i} = \sum_{i=1}^{k} \frac{O_i^2}{E_i} - 2\sum_{i=1}^{k} O_i + \sum_{i=1}^{k} E_i = \sum_{i=1}^{k} \frac{O_i^2}{E_i} - N$$

Illustration 6.1: 300 digits were chosen from a table of numbers and the following frequency distribution was obtained:

Digit	:	0	1	2	3	4	5	6	7	8	9
Frequency	:	26	28	33	32	28	37	33	30	30	23

Test the hypothesis that the digits are uniformly distributed over the table.

Solution: When H_0 is true, the expected frequency of each digit would be 30.

$$\therefore \quad \chi^2 = \frac{1}{30}\sum O_i^2 - N$$

$$= \frac{1}{30}\left(26^2 + 28^2 + 33^2 + 32^2 + 28^2 + 37^2 + 33^2 + 30^2 + 30^2 + 23^2\right) - 300 = 4.8$$

The value of χ^2 from table for 5% level of significance and 9 d.f. is 16.92. Since the calculated value is less than tabulated, there is no evidence against H_0. Thus, the distribution of numbers over the table may be treated as uniform.

Illustration 6.2: A sample analysis of examination results of 200 M.B.A.'s was made. It was found that 46 students had failed, 68 secured a third division, 62 secured a second division and the rest were placed in the first division. Are these figures commensurate with the general examination result which

is in the ratio of 2 : 3 : 3 : 2 for the various categories, respectively? (Given: Table value of chi-square for 3 *d.f.* at 5% level of significance is 7.81.)

Solution: H_0: The students in various categories are distributed in the ratio 2 : 3 : 3 : 2.

The expected number of students, under the assumption that H_0 is true, are:

$$\text{expected number of failures} = \frac{2}{(2+3+3+2)} \times 200 = 40$$

$$\text{expected number of third divisioners} = \frac{3}{10} \times 200 = 60,$$

$$\text{expected number of second divisioners} = \frac{3}{10} \times 200 = 60 \text{ and}$$

$$\text{expected number of first divisioners} = \frac{2}{10} \times 200 = 40$$

$$\text{Thus, we have } \chi^2 = \frac{(46-40)^2}{40} + \frac{(68-60)^2}{60} + \frac{(62-60)^2}{60} + \frac{(24-40)^2}{40} = 8.44.$$

Since this value is greater than the tabulated value, 7.81, for 3 *d.f.* and 5% level of significance, H_0 is rejected.

Illustration 6.3: A survey of 320 families with 5 children each revealed the following distribution:

No. of boys	:	5	4	3	2	1	0
No. of girls	:	0	1	2	3	4	5
No. of families	:	14	56	110	88	40	12

Is the result consistent with the hypothesis that male and female births are equally probable?

Solution: Assuming that H_0 (i.e., male and female births are equally probable) is true, the expected number of families having r boys (or equivalently $5 - r$ girls) is given by $E_r = 320 \times {}^5C_r \left(\frac{1}{2}\right)^5 = 10 \times {}^5C_r$.

On substituting r = 5, 4, 3, 2, 1, 0, the respective expected frequencies are 10, 50, 100, 100, 50 and 10.

$$\therefore \chi^2 = \frac{(14-10)^2}{10} + \frac{(56-50)^2}{50} + \frac{(110-100)^2}{100} + \frac{(88-100)^2}{100} + \frac{(40-50)^2}{50} + \frac{(12-10)^2}{10} = 7.16.$$

The value from table for 5 d.f. at 5% level of significance is 11.07, which is greater than the calculated value. Thus, there is no evidence against H_0.

Illustration 6.4: The record for a period of 180 days, showing the number of electricity failures per day in Delhi are shown in the following table:

No. of failures	:	0	1	2	3	4	5	6	7
No. of days	:	12	39	47	40	20	17	3	2

Determine, by using χ^2- test, whether the number of failures can be regarded as a Poisson variate?

Solution: *We have to test* H_0: No. of failures is a Poisson variate against H_1: No. of failures is not a Poisson variate.

The mean of the Poisson distribution is

$$m = \frac{0\times12+1\times39+2\times47+3\times40+4\times20+5\times17+6\times3+7\times2}{180} = 2.5$$

The computations of χ^2 are done in the following table:

No. of families	*Expected freq.* (E_i)	*Observed freq.* (O_i)	$\frac{(O_i-E_i)^2}{E_i}$
0	$180\times e^{-2.5} = 14.76$	12	0.52
1	$E_0\times2.5 = 36.94$	39	0.11
2	$E_1\times2.5/2 = 46.17$	47	0.01
3	$E_2\times2.5/3 = 38.48$	40	0.06
4	$E_3\times2.5/4 = 24.05$	20	0.68
5	$E_4\times2.5/5 = 12.02$	17	2.06
6 *or more*	*by difference* = 7.58	5	0.88
Total	180		$\chi^2 = 4.32$

The value of χ^2 from table at 5% level of significance and 5 *d.f.* is 11.07. Since the calculated value is less than the tabulated value, there is no evidence against H_0.

χ^2 - test as a Test for Independence of Two Attributes

Let us assume that a population is classified into m mutually exclusive classes, $A_1, A_2, \ldots\ldots A_m$, according to an attribute A and each of these m classes are further classified into n mutually exclusive classes, like $A_iB_1, A_iB_2, \ldots\ldots A_iB_n$, etc., according to another attribute B.

If O_{ij} is the observed frequency of A_iB_j, i.e., $(A_iB_j) = O_{ij}$, the above classification can be expressed in form of following table, popularly known as contingency table.

$B\rightarrow$ $A\downarrow$	B_1	B_2	.....	B_n	*Total*
A_1	O_{11}	O_{12}	.....	O_{1n}	(A_1)
A_2	O_{21}	O_{22}	.....	O_{2n}	(A_2)
⋮	⋮	⋮	.....	⋮	
A_m	O_{m1}	O_{m2}	.....	O_{mn}	(A_m)
Total	(B_1)	(B_2)	.....	(B_n)	N

Assuming that A and B are independent, we can compute the expected frequencies of each cell, i.e., $E_{ij} = \frac{(A_i)(B_j)}{N}$. Thus, $\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ would be a χ^2-variate with $(m-1)(n-1)$ d.f.

Remarks: The expected frequencies of some cells may be obtained by the application of the above formula while the remaining cell frequencies can be obtained by subtraction. The minimum number of cell frequencies, that must be computed by the use of the formula, is equal to the degrees of freedom of the χ^2 statistic.

Illustration 6.5: The employees in 4 different firms are distributed in three skill categories shown in the following table. Test the hypothesis that there is no relationship between the firm and the type of labour. Let the level of significance be 5%.

Firm → / Type of labour ↓	A	B	C	D
Skilled	24	24	23	49
Semi-skilled	32	60	37	51
Manual	24	56	40	80

Solution: H_0: There is no relation between the firm and the nature of labour.

Calculation of Expected Frequencies

Firm → / labour ↓	A	B	C	D	Total
Skilled	$\frac{80\times120}{500}$ = 19.2	$\frac{140\times120}{500}$ = 33.6	$\frac{100\times120}{500}$ = 24.0	$\frac{180\times120}{500}$ = 43.2	120
Semi-skilled	$\frac{80\times180}{500}$ = 28.8	$\frac{140\times180}{500}$ = 50.4	$\frac{100\times180}{500}$ = 36.0	$\frac{180\times180}{500}$ = 64.8	180
Manual	$\frac{80\times200}{500}$ = 32.0	$\frac{140\times200}{500}$ = 56.0	$\frac{100\times200}{500}$ = 40.0	$\frac{180\times200}{500}$ = 72.0	200
Total	80	140	100	180	500

We note that the totals of corresponding rows or columns are same for the observed as well as the expected frequencies.

From the observed and the expected frequencies, we get χ^2 = 12.81. Further, the value of χ^2 from the table for (4 – 1)(3 – 1) = 6 d.f. and 5% level of significance is 12.59. Since the calculated value is greater than the tabulated value H_0 is rejected.

Illustration 6.6: Samples of household income were taken from four cities. Test whether the cities are homogeneous with regard to the distribution of income?

Cities → Income(Rs) ↓	A	B	C	D	Total
Under 3000	10	15	15	10	50
3000-5000	5	10	15	10	40
Over 5000	15	15	10	20	60
Total	30	40	40	40	150

Solution: H_0: Various cities are homogeneous with regard to the distribution of income.

Computation of Expected Frequencies

Cities → Income(Rs) ↓	A	B	C	D	Total
Under 3000	10.00	13.33	13.33	13.33	50
3000-5000	8.00	10.67	10.67	10.67	40
Over 5000	12.00	16.00	16.00	16.00	60
Total	30	40	40	40	150

Note that the expected frequencies for city *A*, under various income groups, are computed as $\frac{30\times 50}{150}=10.00, \frac{30\times 40}{150}=8.00$ and $\frac{30\times 60}{150}=12.00$. Other frequencies have also been computed in a similar manner.

Using the observed and expected frequencies, the value of χ^2 = 8.28.

Further, the value of χ^2 from tables for 6 d.f. at 5% level of significance is 12.59. Since the calculated value is less than the tabulated value, there is no evidence against H_0.

The value of χ^2 for a 2 × 2 Contingency table

For a 2 × 2 contingency table,

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	$a+b+c+d=N$

, the value of χ^2 can be directly computed with the use of the following formula:

$$\chi^2=\frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

Yate's Correction for Continuity

We know that χ^2 is a continuous random variate but the frequencies of various cells of a contingency table are integers. When N is large, the distribution of $\sum\frac{(O-E)^2}{E}$ is approximately χ^2. However, the corrections for continuity are required when N is small. Yates has suggested the following corrections for continuity in a 2 × 2 contingency table:

If $ad > bc$, reduce a and d by $\frac{1}{2}$ and increase b and c by $\frac{1}{2}$. Similarly, If $ad < bc$, increase a and d by $\frac{1}{2}$ and decrease b and c by $\frac{1}{2}$. Thus, the contingency tables in the two situations become

$$\begin{array}{|c|c|}\hline a-\frac{1}{2} & b+\frac{1}{2} \\ \hline c+\frac{1}{2} & d-\frac{1}{2} \\ \hline\end{array} \text{ and } \begin{array}{|c|c|}\hline a+\frac{1}{2} & b-\frac{1}{2} \\ \hline c-\frac{1}{2} & d+\frac{1}{2} \\ \hline\end{array} \text{ respectively.}$$

The value of χ^2 can now be obtained as $\chi^2 = \dfrac{N\left(|ad-bc|-\frac{N}{2}\right)^2}{(a+b)(a+c)(b+d)(c+d)}$.

Brand and Snedecor formula for a $2 \times r$ Contingency table

For a $2 \times r$ contingency table,

$A \rightarrow$ / $B \downarrow$	A_1	A_2	$\cdots$	A_r	Total
B_1	a_1	a_2	$\cdots$	a_r	a
B_2	b_1	b_2	$\cdots$	b_r	b
Total	n_1	n_2	$\cdots$	n_r	N

, the value of χ^2 can be directly computed by the use of the following formula:

$$\chi^2 = \frac{N^2}{ab}\left(\sum_{i=1}^{r}\frac{a_i^2}{n_i} - \frac{a^2}{N}\right) \text{ with } (r-1) \text{ d.f.}$$

Illustration 6.7: In a recent diet survey, the following results were obtained in an Indian city:

No. of families	Hindus	Muslims	Total
Tea takers	1236	164	1400
Non-tea takers	564	36	600
Total	1800	200	2000

Discuss whether there is any significant difference between the two communities in the matter of taking tea? Use 5% level of significance.

Solution: The null hypothesis to be tested can be written as H_0: There is no difference between the two communities in the matter of taking tea.

Using the direct formula, we have $\chi^2 = \dfrac{2000(1236 \times 36 - 164 \times 564)^2}{1400 \times 1800 \times 200 \times 600} = 15.24.$

The value of χ^2 from table for 1 d.f. and 5% level of significance is 3.84. Since the calculated value is greater than the tabulated value, H_0 is rejected.

6.5 PRACTICAL IN EXCEL SOLVER SPSS

```
CHI-SQUARED GOODNESS-OF-FIT TEST
NULL HYPOTHESIS H0:          DISTRIBUTION FITS THE DATA
ALTERNATE HYPOTHESIS HA:  DISTRIBUTION DOES NOT FIT THE DATA
DISTRIBUTION:                NORMAL
SAMPLE:
   NUMBER OF OBSERVATIONS      = 1000
   NUMBER OF NON-EMPTY CELLS   = 24
   NUMBER OF PARAMETERS USED   = 0
TEST:
   CHI-SQUARED TEST STATISTIC  = 17.52155
   DEGREES OF FREEDOM          = 23
   CHI-SQUARED CDF VALUE       = 0.217101
       ALPHA LEVEL      CUTOFF       CONCLUSION
           10%          32.00690      ACCEPT H0
           5%           35.17246      ACCEPT H0
           1%           41.63840      ACCEPT H0
CELL NUMBER, BIN MIDPOINT, OBSERVED FREQUENCY,
AND EXPECTED FREQUENCY
WRITTEN TO FILE DPST1F.DAT
************************************************
** Normal chi-square goodness of fit test y2 **
************************************************
CHI-SQUARED GOODNESS-OF-FIT TEST
NULL HYPOTHESIS H0:          DISTRIBUTION FITS THE DATA
ALTERNATE HYPOTHESIS HA:  DISTRIBUTION DOES NOT FIT THE DATA
DISTRIBUTION                 NORMAL
SAMPLE:
   NUMBER OF OBSERVATIONS      = 1000
   NUMBER OF NON-EMPTY CELLS   = 26
   NUMBER OF PARAMETERS USED   = 0
TEST:
   CHI-SQUARED TEST STATISTIC  = 2030.784
   DEGREES OF FREEDOM          = 25
   CHI-SQUARED CDF VALUE       = 1.000000
```

Contd...

ALPHA LEVEL	CUTOFF	CONCLUSION
10%	34.38158	REJECT H0
5%	37.65248	REJECT H0
1%	44.31411	REJECT H0

CELL NUMBER, BIN MIDPOINT, OBSERVED FREQUENCY,
AND EXPECTED FREQUENCY
WRITTEN TO FILE DPST1F.DAT

****

** Normal chi-square goodness of fit test y3 **

****

CHI-SQUARED GOODNESS-OF-FIT TEST
NULL HYPOTHESIS H0: DISTRIBUTION FITS THE DATA
ALTERNATE HYPOTHESIS HA: DISTRIBUTION DOES NOT FIT THE DATA
DISTRIBUTION: NORMAL
SAMPLE:
NUMBER OF OBSERVATIONS = 1000
NUMBER OF NON-EMPTY CELLS = 25
NUMBER OF PARAMETERS USED = 0
TEST:
CHI-SQUARED TEST STATISTIC = 103165.4
DEGREES OF FREEDOM = 24
CHI-SQUARED CDF VALUE = 1.000000

ALPHA LEVEL	CUTOFF	CONCLUSION
10%	33.19624	REJECT H0
5%	36.41503	REJECT H0
1%	42.97982	REJECT H0

CELL NUMBER, BIN MIDPOINT, OBSERVED FREQUENCY,
AND EXPECTED FREQUENCY
WRITTEN TO FILE DPST1F.DAT

****

** Normal chi-square goodness of fit test y4 **

****

CHI-SQUARED GOODNESS-OF-FIT TEST
NULL HYPOTHESIS H0: DISTRIBUTION FITS THE DATA
ALTERNATE HYPOTHESIS HA: DISTRIBUTION DOES NOT FIT THE DATA
DISTRIBUTION: NORMAL

Contd...

SAMPLE:

NUMBER OF OBSERVATIONS = 1000

NUMBER OF NON-EMPTY CELLS = 10

NUMBER OF PARAMETERS USED = 0

TEST:

CHI-SQUARED TEST STATISTIC = 1162098.

DEGREES OF FREEDOM = 9

CHI-SQUARED CDF VALUE = 1.000000

ALPHA LEVEL	CUTOFF	CONCLUSION
10%	14.68366	REJECT H0
5%	16.91898	REJECT H0
1%	21.66600	REJECT H0

CELL NUMBER, BIN MIDPOINT, OBSERVED FREQUENCY,
AND EXPECTED FREQUENCY
WRITTEN TO FILE DPST1F.DAT

Check Your Progress 2

Fill in the blanks:

1. When the population is not completely specified, the degrees of freedom of would be further reduced by the number of .....estimated.
2. When the degrees of freedom are greater than ....., the sampling distribution of is normal with mean and standard error equal to unity.

6.6 SUMMARY

An addition to the use of in tests of hypothesis concerning the standard deviation, it is used as a test of goodness of fit and as a test of independence of two attributes. These tests are explained in the following sections. Based on the size of sample more than 30 or less than 30, appropriate tests are chosen i.e. chi square test. -test can be used to test, how far the fitted or the expected frequencies are in agreement with the observed frequencies. We know that for large values of n, the sampling distribution of X, the number of successes, is normal with mean np and variance. There are two types of statistical test parametric test and parametric test. In parametric test distribution is considered as normal. Non parametric tests are easy to use. In data analysis researcher may wish to analyse one or more variable at a time. Z test, T tests are examples of parametric tests.

6.7 KEYWORDS

- Chi-square
- Ch-square test
- Degree of freedom

6.8 REVIEW QUESTIONS

1. What is meant by Significance level?
2. What is chi square test of independence?
3. What do you understand by Degree of freedom?
4. What is chi square goodness fit?

Answers to Check Your Progress

Check Your Progress 1

1. Degree of freedom
2. Independent random

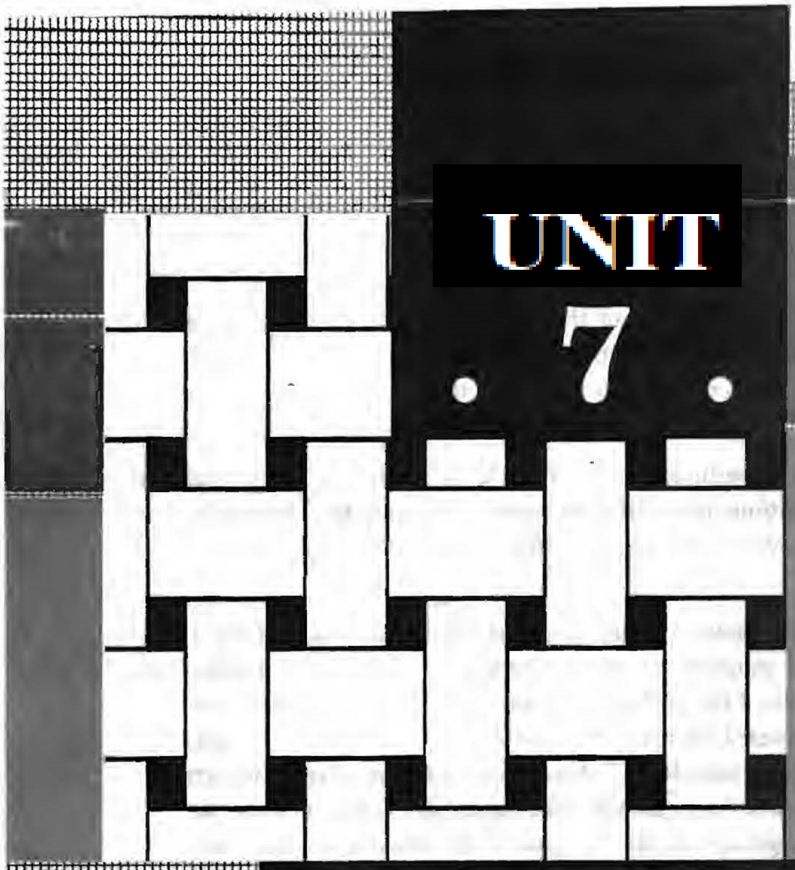
Check Your Progress 2

1. Parameter
2. 30

6.9 REFERENCES AND FURTHER READING

- Creswell, J. W., & Creswell, J. D. (2020). Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.). Sage Publications.
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2021). Multivariate data analysis (8th ed.). Cengage Learning.
- Pallant, J. (2022). SPSS survival manual: A step by step guide to data analysis using SPSS (7th ed.). Open University Press.
- Boslaugh, S. (2023). Statistics in a nutshell (2nd ed.). O'Reilly Media.
- Cohen, L., Manion, L., & Morrison, K. (2024). Research methods in education (9th ed.). Routledge.

BLOCK – III



UNIT

7

Analysis of Variance

CHAPTER OUTLINE

7.1 INTRODUCTION

Analysis of variance is an extension of the test of significance of the difference between two population means to the case involving simultaneous comparison of more than two means. Since this comparison is based on the comparison of variances estimated from different sources, the method is called the analysis of variance (ANOVA).

The technique of analysis of variance was introduced by Sir Ronald A. Fisher. It is essentially a method of partitioning total variation of observations into different sources of variation. This technique was initially used in agricultural research but now it is popular in almost every area of social as well as natural sciences.

Suppose that the manufacturer of a washing powder wants to assess the effectiveness of the three schemes of advertising, say A, B and C. For this purpose, he selects a random sample of 20 distributors which are known to be similar with regard to sales. Out of these he selects, say, 5 distributors at random and assigns the advertising scheme A in areas covered by them. Similarly, we assume that he assigns the advertising scheme B in areas covered by a random sample of 7 distributors and the advertising scheme C in areas covered by the remaining 8 distributors. Since all the distributors are assumed to be similar with regard to sales, the means sales of the three groups would be equal in the absence of advertisement. With different advertisement for different group, their mean sales may or may not remain equal. Thus, the null and the alternative hypotheses can be written as H_0: $\mu_1 = \mu_2 = \mu_3$ and H_a: μ_1, μ_2 and μ_3 are not all equal, respectively. Here μ_1, μ_2 and μ_3 denote the mean sales in populations corresponding to groups 1, 2 and 3 respectively.

The above example can be generalised to the case of k groups (or methods of advertisement) each consisting of n_i distributors, i = 1, 2, ... k. If X_{ij} denotes the sale of the j th distributor in the i th group, the data on sale from all the distributors can be summarised in the form of following table.

Group	*Observations*	*Total*	*Mean*
1	$X_{11}, X_{12}, \cdots\cdots X_{1n_1}$	T_1	$\bar{X}_1$
2	$X_{21}, X_{22}, \cdots\cdots X_{2n_2}$	T_2	$\bar{X}_2$
⋮	⋮	⋮	⋮
k	$X_{k1}, X_{k2}, \cdots\cdots X_{kn_k}$	T_k	$\bar{X}_k$

Also $T = T_1 + T_2 + \ldots\ldots T_k$ and $\bar{X} = \frac{T}{n}$, where $n = n_1 + n_2 + \ldots\ldots n_k$.

In the terminology of the analysis of variance, a scheme of advertising is called a treatment and a distributor is called an experimental unit. The planning of an experiment for the purpose of testing one or more hypotheses is termed as The Design of an Experiment. A basic tool in most of the design of experiments is known as The Analysis of Variance. In the above example, the experimental units are randomly assigned to various treatments and therefore, the experimental design is also termed as completely randomised design and the associated procedure for testing of hypothesis is termed as One-Way Analysis of Variance.

In order to develop the technique of analysis of variance, we assume that the observations with in the i th treatment group are distributed normally with mean μ_i and standard deviation σ (i = 1, 2, ... k). This assumption implies that although a treatment may affect the mean of the group, the dispersion

of observations remain unaffected. Under the assumption that null hypothesis is true, i.e., $H_0 : \mu_1 = \mu_2 = \ldots\ldots = \mu_k = \mu$ (say), each sample observation X_{ij} is a normal random variable with mean m and standard deviation s. Note that the alternate to H_0 is that not all means are equal.

Decomposition of Total Variation

We can write $X_{ij} - \bar{X} = X_{ij} - \bar{X}_i + \bar{X}_i - \bar{X} = (X_{ij} - \bar{X}_i) + (\bar{X}_i - \bar{X})$

Squaring both sides and taking sum over all the observations, we have

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_i - \bar{X})^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)(\bar{X}_i - \bar{X})$$

We note that $2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)(\bar{X}_i - \bar{X}) = 2\sum_{i=1}^{k}(\bar{X}_i - \bar{X})\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i) = 0$

Since $\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i) = 0$, the sum of deviation of values from their mean.

Thus, we can write,

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_i - \bar{X})^2$$

or $$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 + \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$$

i.e., $$\textit{Total variation} = \begin{bmatrix}\textit{Variation within groups} \\ \textit{or unexplained variation}\end{bmatrix} + \begin{bmatrix}\textit{Variation between groups} \\ \textit{or explained variation}\end{bmatrix}$$

7.2 NATURE OF THE TEST STATISTIC

We may note that it is possible to obtain three unbiased estimators of population variance σ^2, based upon the three types of variations given above. For example, the unbiased estimator of σ^2 based on total variation is given by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2 \qquad \ldots\ (1)$$

Similarly, the unbiased estimator of σ^2, using variation within groups, is

$$s_r^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 \qquad \ldots\ (2)$$

It may be pointed out that it is possible to estimate σ^2 even if there is only one sample from the population. The k groups, can be regarded as k samples from a population with variance equal to σ^2. Thus, σ^2 can be estimated in k different ways. For example, the estimator of σ^2 from i th group is given by $s_i^2 = \frac{1}{n_i - 1}\sum_{i=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)^2$, i = 1, 2, ... k. Thus, equation (2) gives a pooled estimator of σ^2 based on all the groups. Further, irrespective of whether $H_0 : \mu_1 = \mu_2 = \mu_3$ is true or not, the value of each s_i^2 (i = 1, 2, ... k) remains unaffected.

A third way of estimating σ^2 is by using the relation

$$Var\left(\bar{X}\right) = \frac{\sigma^2}{n} \text{ or } \sigma^2 = nVar\left(\bar{X}\right).$$

Given k sample means $\bar{X}_1, \bar{X}_2, \cdots\cdots \bar{X}_k$, an unbiased estimator of σ^2, is given by

$$s_t^2 = \frac{1}{k-1}\left[n_1 Var\left(\bar{X}_1\right) + n_2 Var\left(\bar{X}_2\right) + \cdots\cdots + n_k Var\left(\bar{X}_k\right)\right]$$

$$= \frac{1}{k-1}\left[n_1\left(\bar{X}_1 - \bar{X}\right)^2 + n_2\left(\bar{X}_2 - \bar{X}\right)^2 + \cdots\cdots + n_k\left(\bar{X}_k - \bar{X}\right)^2\right]$$

$$= \frac{1}{k-1}\sum_{i=1}^{k} n_i\left(\bar{X}_i - \bar{X}\right)^2 \qquad \ldots\ (3)$$

We note that, when H_0 is true, the difference between s_t^2 and s_e^2 (both estimators of σ^2) would be insignificant. When H_0 is not true, the magnitude of s_t^2 would tend to be greater than s_e^2. The significance of the difference between s_t^2 and s_e^2 can be tested by F - statistic, given by $F = \frac{s_t^2}{s_e^2}$, with (k - 1), (n - k) d.f.

The different sources of variation can also be written in the form of the following table, often known as Analysis of Variance Table.

Analysis of Variance Table

Source of Variation	*Sum of Squares*	*d.f.*	*Mean S.S.*
Between groups (or Treatments)	$TRSS = \sum_{i=1}^{k} n_i\left(\bar{X}_i - \bar{X}\right)^2$	$k-1$	$s_t^2 = \frac{TRSS}{k-1}$
Within groups (or Error)	$ESS = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)^2$	$n-k$	$s_e^2 = \frac{ESS}{n-k}$
Total	$TSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}\right)^2$	$n-1$	

Remarks:

1. The computations of various sum of squares can be done more easily by the use of these alternative expressions.

   TSS (Total Sum of Squares) $= \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \frac{T^2}{n}$

   TRSS (Treatment Sum of Squares) $= \sum_{i=1}^{k}\frac{T_i^2}{n_i} - \frac{T^2}{n}$

   ESS (Error Sum of Squares) = TSS - TRSS $= \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \sum_{i=1}^{k}\frac{T_i^2}{n_i}$

2. The computations can be simplified further by coding. The F - statistic involves the ratio of two variances and therefore, remains invariant to the effect of change of scale and origin.

3. There may be situations in which $F = \frac{s_t^2}{s_e^2}$ is less than unity due to fluctuations of sampling. H_0 should always be accepted in such cases.

4. This test is also known as the test of homogeneity of several population means.

5. Since various experimental units are randomly assigned to various treatments, the one-way ANOVA is also known as Completely Randomised Design.

Example 7.1

A company wants to test whether its three salesmen A, B and C have the same selling ability. Their records of sales (in Rs '000) during various weeks of the last month are given in the following table :

Sales men	*1st week*	*2nd week*	*3rd week*	*4th week*
A	16	21	18	25
B	22	20	15	26
C	25	24	16	20

Prepare an analysis of variance table and test the hypothesis that the mean sales per week of all the salesmen are equal.

Solution

The null hypothesis to be tested is $H_0 : \mu_1 = \mu_2 = \mu_3$, where μ_1, μ_2 and μ_3 denote the mean sales per week by the salesman A, B and C respectively.

(a) **Direct Method**

The sum of squares of observations $\sum\sum X_{ij}^2 = 5288$.

$$\frac{T_1^2}{n_1}=\frac{80^2}{4}=1600.00,\frac{T_2^2}{n_2}=\frac{83^2}{4}=1722.25,\frac{T_3^2}{n_3}=\frac{85^2}{4}=1806.25$$

$$\frac{T^2}{n}=\frac{(80+83+85)^2}{12}=5125.33$$

$\therefore$ $TSS=5288-5125.33=162.67$

$TRSS=1600.00+1722.25+1806.25-5125.33=3.17$

$ESS=TSS-TRSS=162.67-3.17=159.50$

Analysis of Variance Table

Source of Variation	*Sum of Squares*	*d.f.*	*Mean S.S.*
Between Salesmen	$TRSS=3.17$	2	$s_t^2=\frac{3.17}{2}=1.585$
Within Salesmen	$ESS=159.50$	9	$s_e^2=\frac{159.50}{9}=17.72$
Total	$TSS=162.67$	11	

From the table, we have $F=\frac{1.585}{17.72}<1$, \ there is no evidence against H_0.

(b) Coding Method

On subtracting 20 from each observation, we can write

Salesmen	*1st Week*	*2nd Week*	*3rd Week*	*4th Week*	*Total*
A	−4	1	−2	5	0
B	2	0	−5	6	3
C	5	4	−4	0	5

From the above table, we have

$$\sum\sum X_{ij}^2=168,\frac{T_1^2}{n_1}=\frac{0}{4}=0,\frac{T_2^2}{n_2}=\frac{9}{4},\frac{T_3^2}{n_3}=\frac{25}{4}\text{ and }\frac{T^2}{n}=\frac{64}{12}=\frac{16}{3}$$

Further, $$TSS=168-\frac{16}{3}=\frac{488}{3}=162.67$$

and $$TRSS=0+\frac{9}{4}+\frac{25}{4}-\frac{16}{3}=\frac{27+75-64}{12}=\frac{38}{12}=3.17.$$

The rest of the procedure is same as in the direct method.

Example 7.2

The following table gives the yield of a hybrid variety of wheat, in quintals per acre, from 17 trial plots of land treated with four types of fertilisers.

Treatment with fertilizer			
A	*B*	*C*	*D*
24	31	39	38
39	25	41	32
35	26	33	35
	21	40	34
		45	26

Test whether there is any significant difference in the mean yield of wheat due to difference in fertiliser application.

Solution

We have to test $H_0 : \mu_A = \mu_B = \mu_C = \mu_D$, where μ_A, μ_B, μ_C and μ_d denote the mean yield per acre due to fertiliser A, B, C and D respectively.

On subtracting 33 from every observation, the given table can be rewritten as

Treatment with fertiliser			
A	*B*	*C*	*D*
−9	−2	6	5
6	−8	8	−1
2	−7	0	2
	−12	7	1
		12	−7
$T_1 = -1$	$T_2 = -29$	$T_3 = 33$	$T_4 = 0$

From the above table, we can write

$$\sum\sum X_{ij}^2 = 755, \frac{T_1^2}{n_1} = \frac{1}{3}, \frac{T_2^2}{n_2} = \frac{29^2}{4} = \frac{841}{4}, \frac{T_3^2}{n_3} = \frac{1089}{5}, \frac{T_4^2}{n_4} = 0, \frac{T^2}{n} = \frac{9}{17}$$

Thus, we have

$$TSS = 755 - \frac{9}{17} = 754.47$$

$$TRSS = \frac{1}{3} + \frac{841}{4} + \frac{1089}{5} - \frac{9}{17} = 427.85$$

$$ESS = TSS - TRSS = 754.47 - 427.85 = 326.62$$

Analysis of Variance Table

Source of Variation	*S.S.*	*d.f.*	*M.S.S.*
Between Treatment	427.85	3	$s_t^2 = \frac{427.85}{3} = 142.62$
Within Treatment	326.62	13	$s_e^2 = \frac{326.62}{13} = 25.12$
Total	754.47		

From the above table, we have

$$F = \frac{142.62}{25.12} = 5.68 > 3.41$$

The critical value of F for 3,13 $d.f$.and $\alpha = 0.05$. Thus, H_0 is rejected at 5% level of significance.

7.3 TESTING THE SIGNIFICANCE OF REGRESSION USING ANALYSIS OF VARIANCE

The analysis of variance technique can also be used to test the significance of regression coefficient. If the fitted regression line is $Y_{c_i} = a + bX_i$, we can write

$$\Sigma\left(Y_i - \bar{Y}\right)^2 = \Sigma\left(Y_i - Y_{c_i}\right)^2 + \Sigma\left(Y_{c_i} - \bar{Y}\right)^2$$

$$\begin{pmatrix}Total\ SS\end{pmatrix} \quad \begin{pmatrix}Unexplained\ SS \\ or\ Error\ SS\end{pmatrix} \quad \begin{pmatrix}Explained\ SS\ or \\ due\ to\ Regression\end{pmatrix}$$

Thus, the analysis of variance table can be written as

Source of Variation	*Sum of Squares*	*d.f.*	*Mean SS*
Due to Regression	$RSS = \sum\left(Y_{c_i} - \bar{Y}\right)^2$	1	$s_R^2 = \frac{RSS}{1} = RSS$
Error	$ESS = TSS - RSS$	$n-2$	$s_e^2 = \frac{ESS}{n-2}$
Total	$TSS = \sum\left(Y_i - \bar{Y}\right)^2$	$n-1$	

From the above table, we have $F = \frac{s_R^2}{s_e^2}$ with 1, (n - 2) d.f.

Remarks: To simplify calculations, we can write $TSS = \sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n}$ and

$ESS = \sum Y_i^2 - a\sum Y_i - b\sum X_i Y_i$. Then, RSS = TSS - ESS.

Example 7.3

The following intermediate results were obtained for the following model :

$C = \alpha + \beta X + U$, where C = total cost (in Rs '000), X = quantity produced (in '000 units) and U = random error.

n = 10, $\Sigma X = 777$, $\Sigma C = 1{,}657$, $\Sigma CX = 132{,}938$, $\Sigma X^2 = 70{,}903$ and $\Sigma C^2 = 2{,}77{,}119$.

(i) Estimate the parameters of the cost function.

(ii) Compute the coefficient of correlation between total cost and total output and test its significance at 1% level of significance.

(iii) Test the significance of linear regression by analysis of variance.

Solution

(i) The estimate of b is given by

$$b = \frac{\sum CX - \frac{\sum C \sum X}{n}}{\sum X^2 - \frac{\left(\sum X\right)^2}{n}} = \frac{132938 - \frac{777 \times 1657}{10}}{70903 - \frac{777^2}{10}} = 0.398$$

The estimate of a is given by $a = \bar{C} - b\bar{X} = \frac{1657}{10} - 0.398 \times \frac{777}{10} = 134.76$

(ii) $$r_{CX} = \frac{\sum CX - \frac{\sum C \sum X}{n}}{\sqrt{\sum C^2 - \frac{\left(\sum C\right)^2}{n}}\sqrt{\sum X^2 - \frac{\left(\sum X\right)^2}{n}}} = \frac{132938 - \frac{777 \times 1657}{10}}{\sqrt{277119 - \frac{1657^2}{10}}\sqrt{70903 - \frac{777^2}{10}}}$$

$= 0.808.$

To test $H_0 : \rho = 0$, the test statistic is $t = 0.808\sqrt{\frac{8}{\left(1 - 0.808^2\right)}} = 3.88$

Since this value is greater than 2.355, the tabulated value at 8 d.f. and $\alpha = 0.05$, H_0 is rejected. Hence, the correlation in population is significant.

(iii) For testing the significance of regression, we compute

$$TSS = \sum C^2 - \frac{\left(\sum C\right)^2}{n} = 277119 - \frac{1657^2}{10} = 2554.1$$

$$ESS = \sum C^2 - a\sum C - b\sum CX$$

$= 277119 - 134.76 \times 1657 - 0.398 \times 132938 = 912.36$

$\therefore$ RSS = TSS - ESS = 2554.1 - 912.36 = 1641.74

Analysis of Variance Table

Source of Variation	*Sum of Squares*	*d.f.*	*Mean SS*
Due to Regression	$RSS = 1641.74$	1	$s_R^2 = 1641.74$
Error	$ESS = 912.36$	8	$s_e^2 = \frac{912.36}{8} = 114.05$
Total	$TSS = 2554.1$	9	

Thus, $F = \frac{1641.74}{114.05} = 14.39$. The value of F from tables for 1, 8 d.f. at 5% level of significance is 5.32. Therefore, $H_0 : \beta = 0$ is rejected at 5% level of significance.

An important technique for analyzing the effect of categorical factors on a response is to perform an Analysis of Variance. An ANOVA decomposes the variability in the response variable amongst the different factors. Depending upon the type of analysis, it may be important to determine: (a) which factors have a significant effect on the response, and/or (b) how much of the variability in the response variable is attributable to each factor.

Analysis of variance is an extension of the test of significance of the difference between two population means to the case involving simultaneous comparison of more than two means. Since this comparison is based on the comparison of variances estimated from different sources, the method is called the analysis of variance (ANOVA).

The technique of analysis of variance was introduced by Sir Ronald A. Fisher. It is essentially a method of partitioning total variation of observations into different sources of variation. This technique was initially used in agricultural research but now it is popular in almost every area of social as well as natural sciences.

7.4 TEST FOR DIFFERENCE AMONG MORE THAN TWO SAMPLES

Kruskal-Wallis H Test

This extension of the Mann –Whitney U test to multiple samples is a nonparametric alternative to one-way analysis of variance. It tests the null hypothesis that the samples do not differ in mean rank for the criterion variable. Because it takes rank size into account to a certain extent than just the above-below dichotomy of the median test, discussed below, it is more influential and preferable when its assumptions are met. The H test will make known if there are rank differences among multiple groups. For pairwise comparisons, the Mann-Whitney test is suitable.

Median Test

Median Test also called the Westenberg-Mood median test; this is a more extensive but less powerful alternative to the Kruskal-Wallis H test for testing if hardly any independent samples come from the similar population. It tests whether two or more independent samples differ in their median values for a criterion variable.

Jonckheere-Terpstra Test

The Jonckheere-Terpstra tests for ordered differences among groups assumed to be prearranged ordinally. As such it is a test for differences among a number of independent samples which is more powerful than the Kruskal-Wallis H or median tests. It requires that the independent samples (the grouping variable) be ordinally prearranged as well as that the test variable be ordinal. The J-T test tests the hypothesis that as one moves from samples low on the ordering criterion to samples high on the criterion, the within-sample degree of the test variable increases or decreases systematically. The **null** hypothesis in the Jonckheere-Terpstra test is that the test variable does not differ between groups.

Check Your Progress 1

Fill in the blanks:

1. Data are entered as two separate variables, one for the dependent variable and one for the .....
2. Analysis of variance is an extension of the .....of the difference between two population means to the case involving simultaneous comparison of more than two means.

7.5 INFERENCE ABOUT A POPULATION VARIANCE

Most types of statistical inference focus on population means, as they are the simplest way of characterizing a single population or comparing two or more populations. Inference on population variances is also needed, to Place confidence intervals on or test for the size of the population variance and to compare variances from two populations.

Inferences for a Single Population Variance

Recall: For data values $y_1, \ldots, y_n$, a point estimator of σ^2 is given by: $s^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2$

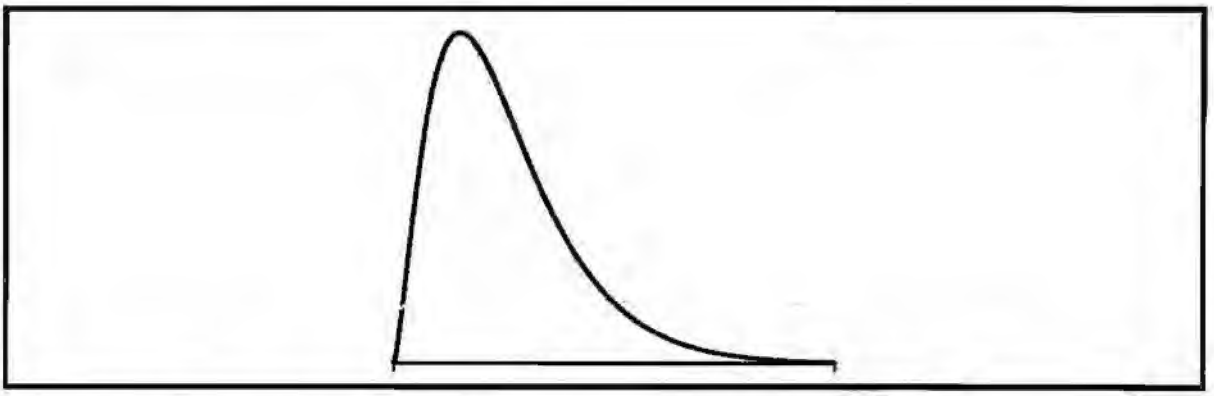
Result: If $Y_1, \ldots, Y_n$ are a random sample from a $N(\mu, \sigma)$ population, then the test statistic:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

- This statistic is said to have chi-squared distribution with $n - 1$ degrees of freedom.

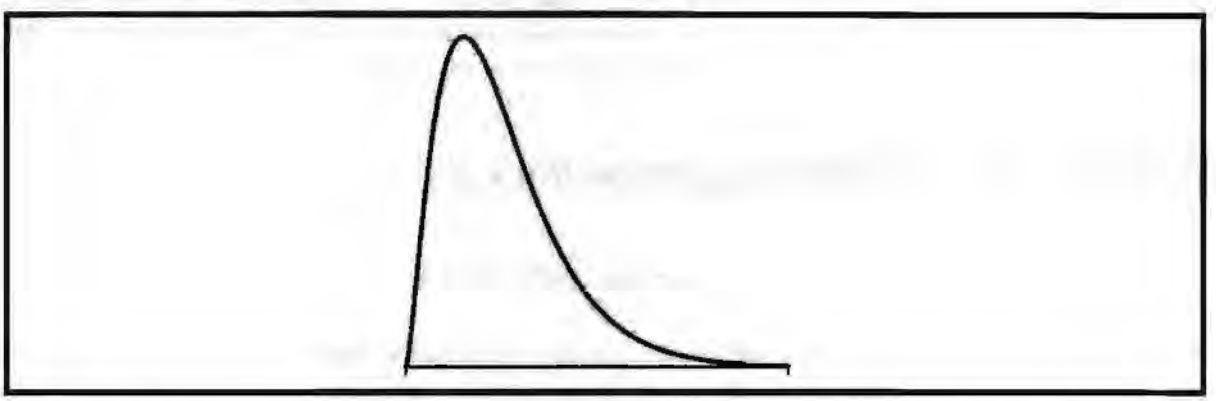
Properties of a $\chi^2(n-1)$ Distribution

1. The distribution is skewed to the right.
2. The degrees of freedom completely specify which χ^2-distribution to use.
3. The mean of a $\chi^2(n-1)$ is $n - 1$ (the degrees of freedom), with variance $2(n-1)$.
4. χ^2 random variables only take non-negative values.



How to Compute χ^2-Tail Values

- Suppose the sample size is $n = 20 \Rightarrow n - 1 = 19$ d.f.
- We want to find values χ^2_L and χ^2_U such that 95% of the $\chi^2(19)$-distribution falls between them.
- We can use Table to find these χ^2 critical values.
- In this case, we have: $\chi^2_L =$ _____, and $\chi^2_U =$ _____.
- Why do we care about all this? This is the basis for finding a confidence interval for σ^2 or σ. How? Consider what is known:



Definition: For $y_1, \ldots, y_n$ a random sample from a $N(\mu, \sigma)$ population, a $100(1 - \alpha)\%$ confidence interval (CI) for σ^2 has the form:

$$\left[\frac{(n-1)S^2}{\chi^2_U}, \frac{(n-1)S^2}{\chi^2_L}\right]$$

where χ^2_U, χ^2_L are the upper and lower $\alpha/2$ tail probabilities of a $\chi^2(n - 1)$-distribution.

Example 7.4: Baseballs vary in terms of their "rebounding coefficient (RC)" (varying from a dead ball) $\Rightarrow$ rabbit ball). A purchaser of baseballs requires a mean RC of 85 with a standard deviation σ no larger than 2.

Suppose we take a random sample of $n = 81$ baseballs and find that $\bar{y} = 84{:}91$ and $s^2 = 4{:}84$ ($s = 2{:}2$). Give a 95% confidence interval for σ.

Assuming a normal population for the rebounding coefficient, we have $n - 1 = 80$ d.f.

$\Rightarrow \quad \chi^2_L =$ _____, $\chi^2_U =$ _____.

We are 95% confident that the interval above contains the true population variance σ^2.

Note: This interval is not symmetric (i.e.: the sample standard deviation $s^2 = 4{:}84$ is not in the center).

This gives us a confidence interval for σ^2. What about σ? Taking square roots, a 95% CI for σ is given by:

Hypothesis Testing

To test $H_0 : \sigma^2 = \sigma^2_0$ versus H_a, the test statistic is:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2_0}$$

For the three alternative hypotheses, the *P*-values and rejection regions are given by:

Alternative Hypothesis	*P*-value	Rejection Region
$H_a : \sigma^2 < \sigma^2_0$	$P(X^2 < \chi^2)$	Reject H_0 if $\chi^2 < \chi^2_L$
$H_a : \sigma^2 > \sigma^2_0$	$P(X^2 > \chi^2)$	Reject H_0 if $\chi^2 > \chi^2_U$
$H_a : \sigma^2 6= \sigma^2_0$	$2 \min[P(X^2 > \chi^2); P(X^2 < \chi^2)]$	Reject H_0 if $\chi^2 < \chi^2_L$ or $\chi^2 > \chi^2_U$

where X^2 is a $\chi^2(n - 1)$ random variable.

- The term σ^2 0 is known as the **hypothesized variance.**

Back to the Baseball Example: We wished to test (for $n - 1 = 80$ d.f.; $\alpha = .05$):

$H_0 : \sigma \leq 2$ (the purchaser's requirements are met)

vs. $H_a : \sigma > 2$ (there is more variability than required):

The test statistic here is: $\chi^2 = \dfrac{80(4.84)}{2^2} = 96.8$

The rejection region is given by: Reject H_0 if $\chi^2 > 101.88$.

Conclusion: Since the *p*-value > α (or since $\chi^2 = 96.8 < 101.88$), then we fail to reject H_0 at an $\alpha = .05$ level, and conclude we do not have enough evidence that $\sigma > 2$.

Some Notes on the Assumptions for χ^2-Inferences

- When working with the population mean, we saw that if the sample size n is large ($n \geq 30$), then even if the population distribution is not normal, the sampling distribution of the *sample mean* $\bar{y}$ is approximately normal allowing the use of a *t*-procedure.

- Unfortunately, this does not carry over to inference on the population variance. Even if n is large, if the population distribution is not normal, the chi-square inferences can be very poor and should not be used. How can we compare two variances when the population distributions are nonnormal?

7.6 INFERENCES FOR COMPARING TWO POPULATION VARIANCES

The primary reason for wanting to compare two population variances is to test the **homogeneous (equal) variance assumption** (i.e.: $\sigma_1^2 = \sigma_2^2$), to decide which type of t-procedure to use when performing inferences on two population means.

This is less of an issue with two means as we can use the separate-variance t-test. This will be an important issue when comparing more than two means (ANOVA).

Result: If $X_1 \sim \chi^2(n_1 - 1)$ and $X_2 \sim \chi^2(n_2 - 1)$, and X_1 & X_2 are independent, then the random variable:

$$\frac{X_1/(n_1-1)}{X_2/(n_2-1)} \sim F(n_1 - 1,\ n_2 - 1),$$

and we say that this random variable has an F-distribution with $df_1 = n_1 - 1$ numerator degrees of freedom and $df_2 = n_2 - 1$ denominator degrees of freedom.

Why is this result important? How will it help in performing inferences on two population variances?

Recall that for two sample problems, we generally assume that:

(i) The two populations are normal.

(ii) The samples taken from these populations are independent.

Both of these assumptions are required for performing inferences on two population variances using the result given.

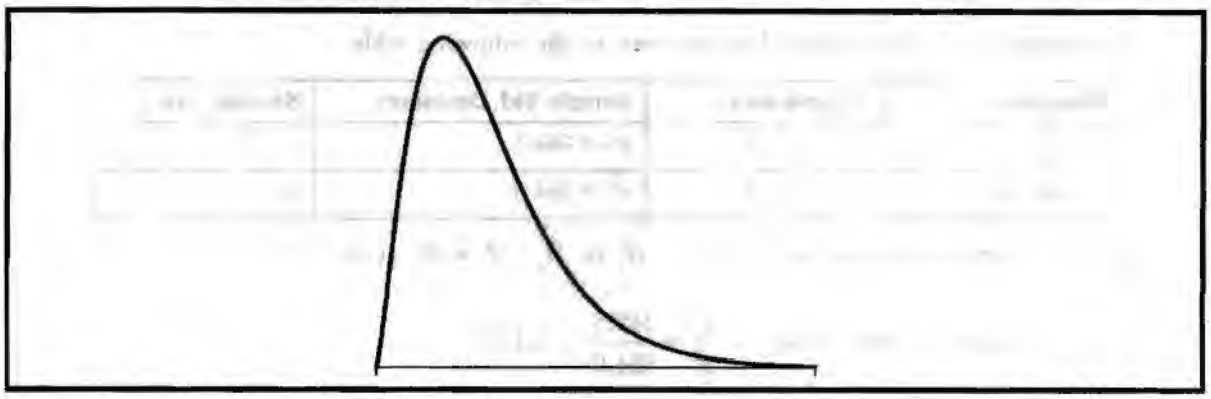
Consider the following. We know from an earlier result in this handout:

$$\left.\begin{aligned}\frac{(n_1-1)s_1^2}{\sigma_1^2} &\sim \chi^2(n_1-1)\\ \frac{(n_2-1)s_2^2}{\sigma_2^2} &\sim \chi^2(n_2-1)\end{aligned}\right\}\ \text{independent } \chi^2 \text{ random variables (since the samples are independent)}$$

$$\Rightarrow \quad \frac{\frac{(n_1-1)s_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)s_2^2}{\sigma_2^2}/(n_2-1)} = \boxed{\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F(\text{df } 1 = n_1 - 1,\ \text{df } 2 = n_2 - 1)} \quad (\text{by the Result})$$

Properties of an $F(n_1 - 1, n_2 - 1)$ Distribution

1. The distribution is skewed to the right.
2. The numerator and denominator degrees of freedom (df_1, df_2) completely specify which F-distribution to use.
3. F random variables only take nonnegative values.
4. Only upper tail values are given.



Examples for Computing F Critical Values

(a) Find the 5% upper tail value for $n_1 = 10$, $n_2 = 6$.

(b) Find the 5% lower tail value for $n_1 = 10$, $n_2 = 6$. How?

In general, to find lower tail critical F-values: $F_L(df1; df2) = \dfrac{1}{F_U(df2, df1)}$.

Hypothesis Test for $\sigma^2_1 = \sigma^2_2$ (Normal Populations)

To test $H_0 : \sigma^2_1 = \sigma^2_2$ versus Ha; the test statistic is:

$$f = \frac{s_1^2/\sigma^2_{10}}{s_2^2/\sigma^2_{20}} = \frac{s_1^2}{s_2^2} \text{ (since } \sigma^2_{10} = \sigma^2_{20} \text{ under } H_0\text{):}$$

Note: Large or small values (different from 1) give evidence against H_0:

For the two alternative hypotheses, the P-values and rejection regions are given by:

Alternative Hypothesis	*P*-value	Rejection Region
$H_a : \sigma^2_1 > \sigma^2_2$	$P(F > f)$	Reject H_0 if $f > F_U = F_{df1,df2}$
$H_a : \sigma^2_1 \neq \sigma^2_2$	$2 \min[P(F > f); P(F > (1 = f))]$	Reject H_0 if $f > F_U = F_{df1,df2}$ or if $f < F_L = 1 = F_{df2,df1}$

where F is an $F(n_1 - 1, n_2 - 1)$ random variable.

Note: We do not consider the alternative hypothesis $H_a : \sigma^2_1 < \sigma^2_2$ in the above inference outline because we can always label the two populations such that $\sigma^2_1 > \sigma^2_2$.

Example 7.5: Suppose a study is conducted on the effect of a particular drug in reducing hypertension. For this study, a sample of rats was split into two groups:

(i) A control group (representing population 1).

(ii) An experimental group (representing population 2) which received the new drug.

- Hypertension was induced by exposing the rats to a cold environment.
- The response variable was blood pressure (y).
- Assume both population distributions are normal. Exactly what is normal here?

The summary statistics for the data are given in the following table:

Population	Sample Mean	Sample Std. Deviation	Sample Size
1 (Control)	$\bar{y}_1 = 167.6$	$s^2_1 = 249.3$	$n_1 = 5$
2 (Expmt'l)	$\bar{y}_2 = 129.4$	$s^2_2 = 584.0$	$n_2 = 7$

We wish to test the hypotheses: $H_0 : \sigma^2_1 = \sigma^2_2$ vs. $H_a : \sigma^2_1 \neq \sigma^2_2$, at $\alpha = .05$.

The test statistic for this test is: $f = \dfrac{s^2_1}{s^2_2} = \dfrac{249.3}{584.0} = 0.427$

The critical values for this test are:

$F_U =$

$F_L =$

Hence the nonrejection region is given by: _____.

Conclusion: Since $f \in (.11, 6.23)$, then we fail to reject H_0 and conclude ($\alpha = .05$) that we do not have enough evidence to say that $\sigma^2_1 \neq \sigma^2_2$.

100(1 – α)% Confidence Interval for σ^2_1/σ^2_2

Since $P\left(F_L < \dfrac{s^2_2/\sigma^2_2}{s^2_1/\sigma^2_1} < F_U\right) = 1 - \alpha$ (for $F_U = F_{df2,df1}(\alpha/2)$, $F_L = 1/F_{df1,df2}(\alpha/2)$), then:

$$P\left(\frac{s^2_1}{s^2_2}F_L < \frac{\sigma^2_1}{\sigma^2_2} < \frac{s^2_1}{s^2_2}F_U\right) = 1-\alpha \left[P\left(\frac{s_1}{s_2}\sqrt{F_L} < \frac{\sigma_1}{\sigma_2} < \frac{s_1}{s_2}\sqrt{F_U}\right) = 1-\alpha\right].$$

Note: In F_U above, df_2 appears as the numerator degrees of freedom and df_1 as the denominator degrees of freedom. This occurs because s^2_2 was used in the numerator and s^2_1 in the denominator of the first probability expression above.

Hence, a 100(1 – α)% CI for $\dfrac{\sigma^2_1}{\sigma^2_2}$ is: $\left(\dfrac{s^2_1}{s^2_2}F_L, \dfrac{s^2_1}{s^2_2}F_U\right)$, where $F_L = 1/F_{df1,df2}(\alpha/2)$, $F_U = F_{df2,df1}(\alpha/2)$.

What does a CI for $\sigma_1 = \sigma_2$ look like?

Example 7.6: In the hypertension drug example, we had:

$$\left.\begin{array}{l} n_1 = 5 \quad s_1^2 = 249.3 \\ n_2 = 7 \quad s_2^2 = 584.0 \end{array}\right\} \text{From normal populations.}$$

⇒ We are 90% confident that this interval contains the true ratio (s^2_1 / s^2_2) of standard deviations for these populations.

Importance of Normality: For inferences on population means, the t-procedures used are robust to mild departures from normality due to the Central Limit Theorem.

- Unfortunately, the assumption of normal populations is critical for inferences on population variances via χ^2-procedures or *F*-procedures.
- These χ^2-& F-procedures are very sensitive to even mild skewness or outlying values.

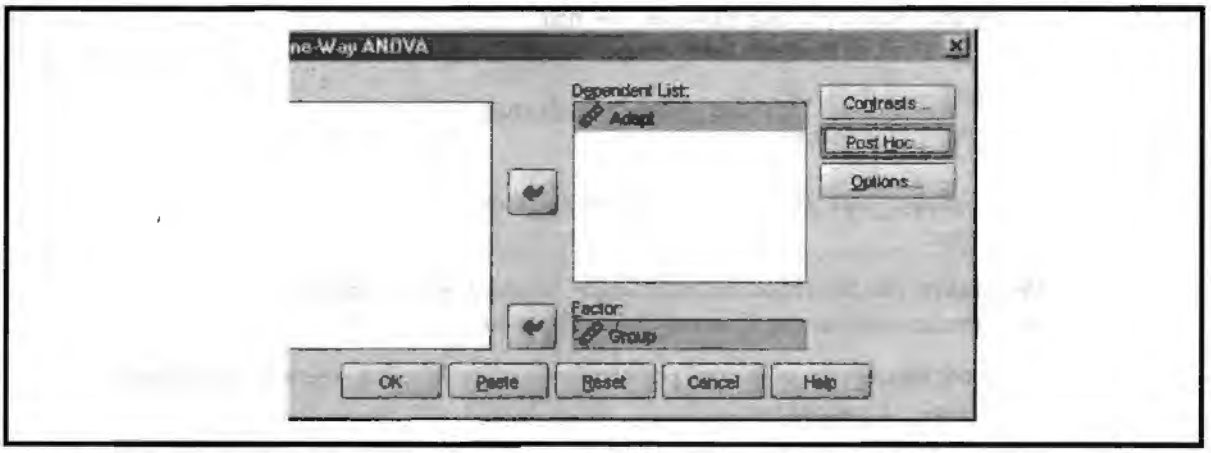
The two tests for homogeneity of more than 2 population variances known as Hartley's test and Levene's test. Hartley's test is based on the *F*-test outlined for two population variances above, but is so highly sensitive to departures from nonnormality that it is rarely used in practice. Levene's test is now the standard test performed in an ANOVA for testing whether the population variances for the different treatment groups are the same. This test will be presented in Stat 452 when the basics of analysis of variance (ANOVA) are introduced.

7.7 ONE WAY ANALYSIS OF VARIANCE PRACTICAL IN EXCEL

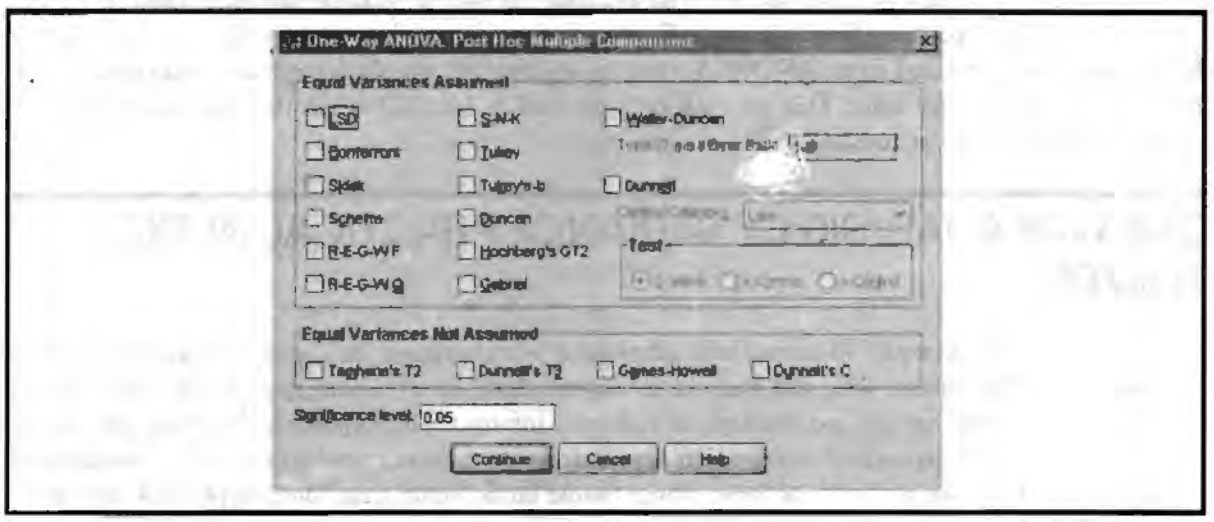
SOLVED

Let's begin with an example: Maternal role adaptation was compared in a group of mothers of low birth-weight (LBW) infants who had been in an experimental intervention, mothers of LBW infants who were in a control group, and mothers of full-term infants. The hypothesis was that mothers of LBW infants in the experimental intervention would adapt to their maternal role as well as mothers of healthy full-term infants, and each of these groups would adapt better than mothers of LBW infants in the control group.

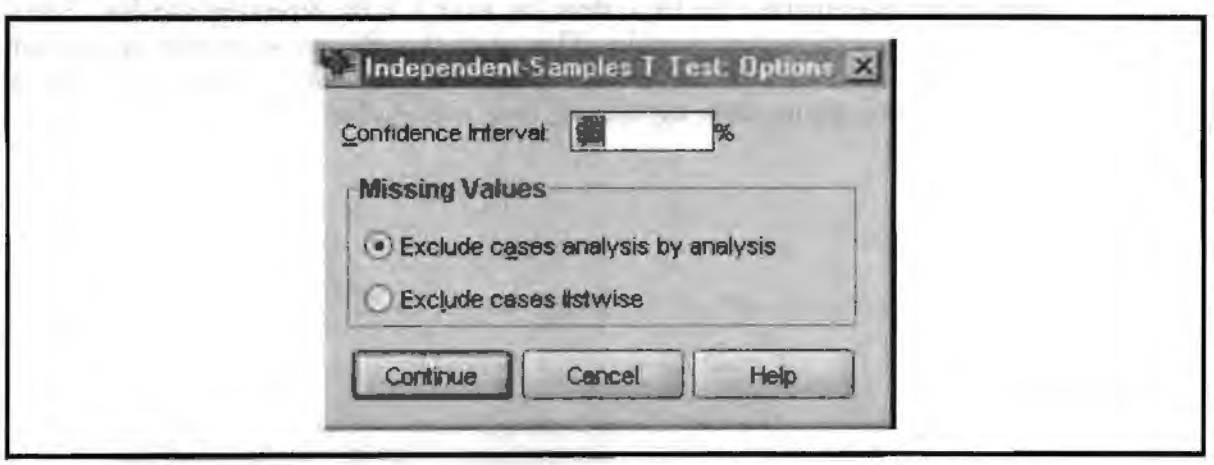
- Open maternal role adaptation.sav.
- Select Analyze/Compare Means/One-Way ANOVA.
- Select maternal role adaptation for the Dependent List since it is the dependent variable. Select group as the Factor or independent variable. Then click Post Hoc to see various options for calculating multiple comparisons. If the ANOVA is significant, we can use the post hoc tests to determine which specific groups differ significantly from one another.



- As you can see, there are many options. Let's select LSD under Equal Variances Assumed since it is Fisher's Least Significant Difference Test which is calculated in the text, except that SPSS will test the differences even if the overall F is not significant.



- Note that .05 is the default under Significance level. After consulting with SPSS technical support, it is clear that this is the experiment-wise or family-wise significance level. So any comparison flagged by SPSS as significant is based on a Bonferroni Type Correction. You do not need to adjust the significance level yourself.



- Click Options. In the next dialog box, select Descriptives under Statistics, and select Means plot so SPSS will create a graph of the group means for us. The default under Missing Values is Exclude cases analysis by analysis. Let's leave this as is. Click Continue and then Ok. The output follows.

Descriptives

Adapt

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
LBW-Exp	29	14.97	4.844	.899	13.12	16.81	10	29
LBW-Control	27	18.33	5.166	.994	16.29	20.38	10	29
Full-Term	37	14.94	3.708	.610	13.60	16.07	10	25
Total	93	15.89	4.747	.492	14.91	16.87	10	29

ANOVA

Adapt

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	226.932	2	113.466	5.532	.005
Within Groups	1845.993	90	20.511		
Total	2072.925	92			

Post Hoc Tests

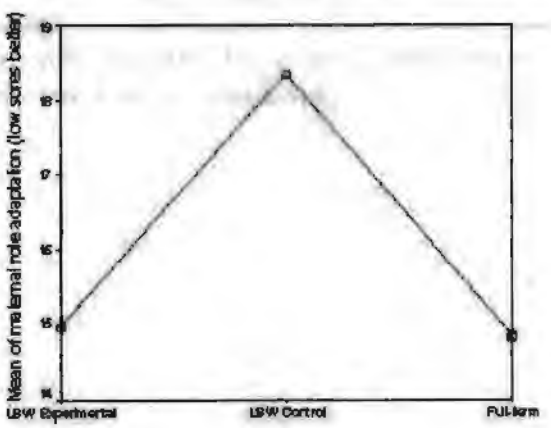
Multiple Comparisons

Adapt
LSD

(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
LBW-Exp	LBW-Control	-3.368*	1.211	.007	-5.77	-.96
	Full-Term	.128	1.123	.910	-2.10	2.36
LBW-Control	LBW-Exp	3.368*	1.211	.007	.96	5.77
	Full-Term	3.495*	1.146	.003	1.22	5.77
Full-Term	LBW-Exp	-.128	1.123	.910	-2.36	2.10
	LBW-Control	-3.495*	1.146	.003	-5.77	-1.22

*. The mean difference is significant at the 0.05 level.

Means Plots



Compare this output to the results presented in the text.

We can see the descriptive statistics and the F value are the same. It is harder to compare the post hoc comparisons because SPSS does not display the t values. They simply report the mean difference and the significance level. The important thing to note is that the conclusions we can draw based on each of these approaches are the same.

The plot that SPSS created is an effective way to illustrate the mean differences. You may want to edit the graph using what you learned in Chapter 3 to make it more elegant. Some people would prefer a bar chart since these are independent groups and a line suggests they are related. You could create a bar chart of these group means yourself.

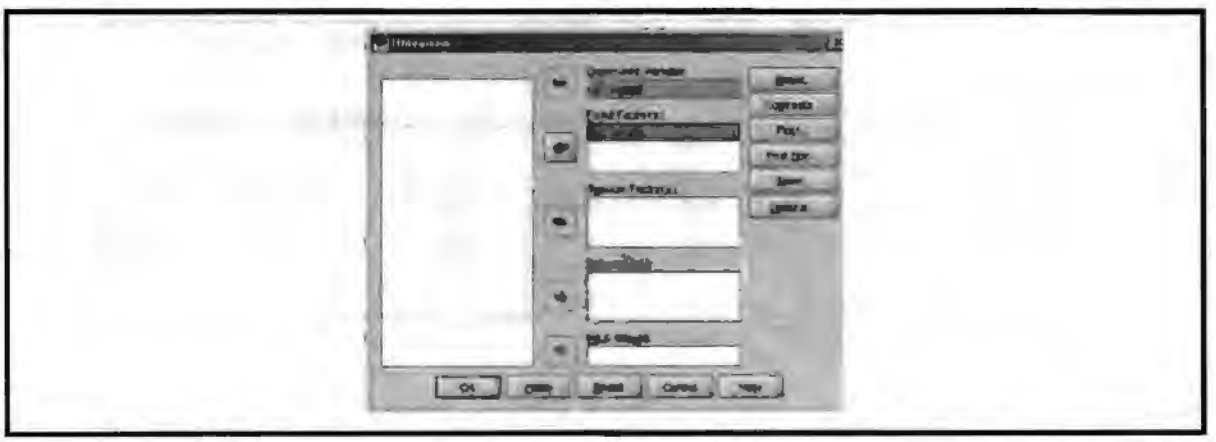
Let's re-run the same analysis using the General Linear Model (GLM) and see how they are similar and different.

General Linear Model to Calculate One-Way ANOVAs

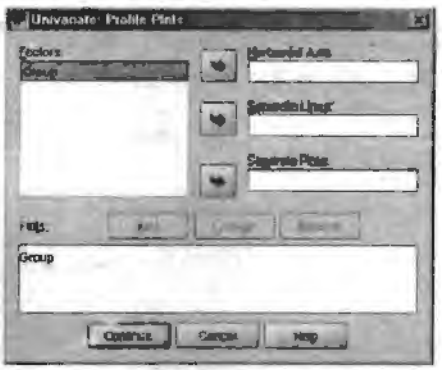
The Univariate General Linear Model is really intended to test models in which there is one dependent variable and multiple independent variables. We can use it to run a simple one-way ANOVA like the one above. One advantage of doing so is that we can estimate effect size from this menu, but we could not from the One-Way ANOVA menus. Let's try it.

Select Analyze/General Linear Model/Univariate.

- As you can see by this dialog box, there are many more opti .. than the One- Way ANOVA. This is because the GLM is a powerful technique that can examine complex designs. We'll just focus on what is relevant to us. As before, select maternal role adaptation as the Dependent Variable and group as the Fixed Factor or independent variable. Then, click Plots.

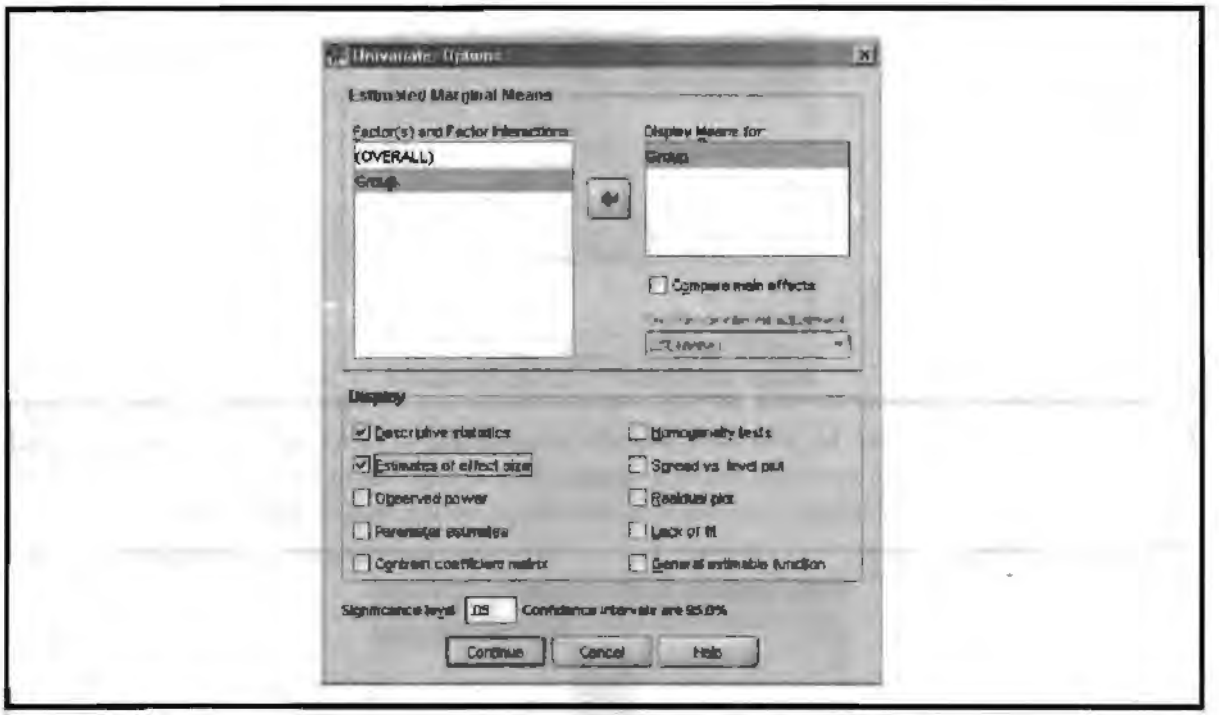


- Select group for the Horizontal Axis (X axis), and click add. Since there is only one dependent variable, SPSS knows that maternal role adaptation is on the Y axis without us needing the specify this. Click Continue.



- Since this procedure can be used with multiple independent variables, we need to specify which ones to run post hoc comparisons for even though there is only one in our design. Select group for Post Hoc Tests for. This time, let's select Bonferroni to see if it makes a difference.

- Under Display, select Descriptive statistics and Estimates of effect size. Then click Continue. In the main dialog box, click Ok. The output follows.



Between-Subjects Factors

		Value Label	N
Group	1	LBW-Exp	29
	2	LBW-Control	27
	3	Full-Term	37

Descriptive Statistics

Dependent Variable:Adapt

Group	Mean	Std. Deviation	N
LBW-Exp	14.97	4.844	29
LBW-Control	18.33	5.166	27
Full-Term	14.84	3.709	37
Total	15.89	4.747	93

Tests of Between-Subjects Effects

Dependent Variable:Adapt

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	226.932[a]	2	113.466	5.532	.005	.109
Intercept	23513.095	1	23513.095	1146.364	.000	.927
Group	226.932	2	113.466	5.532	.005	.109
Error	1845.993	90	20.511			
Total	25562.000	93				
Corrected Total	2072.925	92				

a. R Squared = .109 (Adjusted R Squared = .090)

Post Hoc Tests

Group

Multiple Comparisons

Adapt
Bonferroni

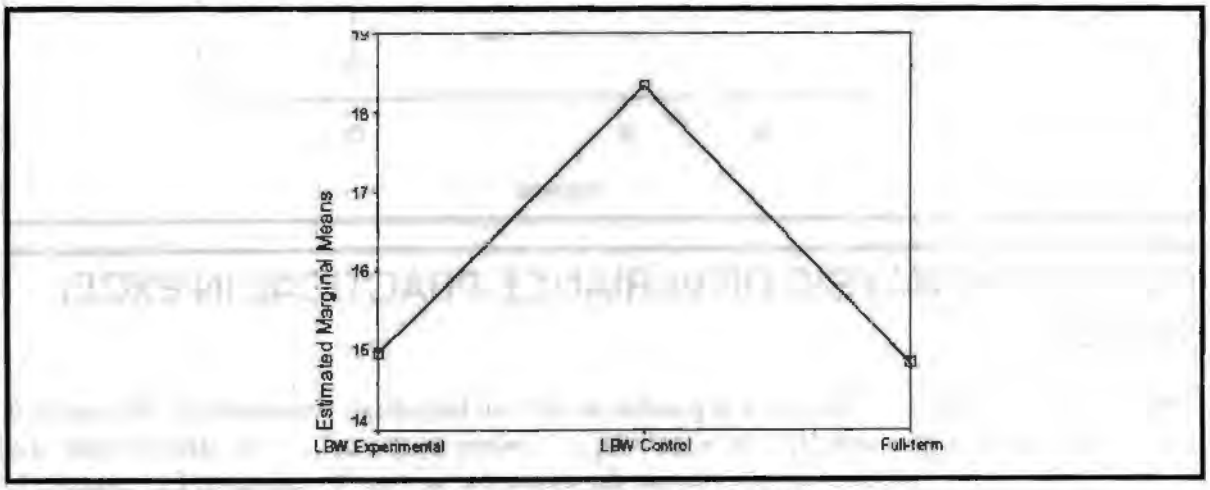
(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
LBW-Exp	LBW-Control	-3.37*	1.211	.020	-6.32	-.41
	Full-Term	.13	1.123	1.000	-2.61	2.87
LBW-Control	LBW-Exp	3.37*	1.211	.020	.41	6.32
	Full-Term	3.50*	1.146	.009	.70	6.29
Full-Term	LBW-Exp	-.13	1.123	1.000	-2.87	2.61
	LBW-Control	-3.50*	1.146	.009	-6.29	-.70

Based on observed means.
The error term is Mean Square(Error) = 20.511.

*. The mean difference is significant at the .05 level.

Profile Plots

Estimated Marginal Means of maternal role

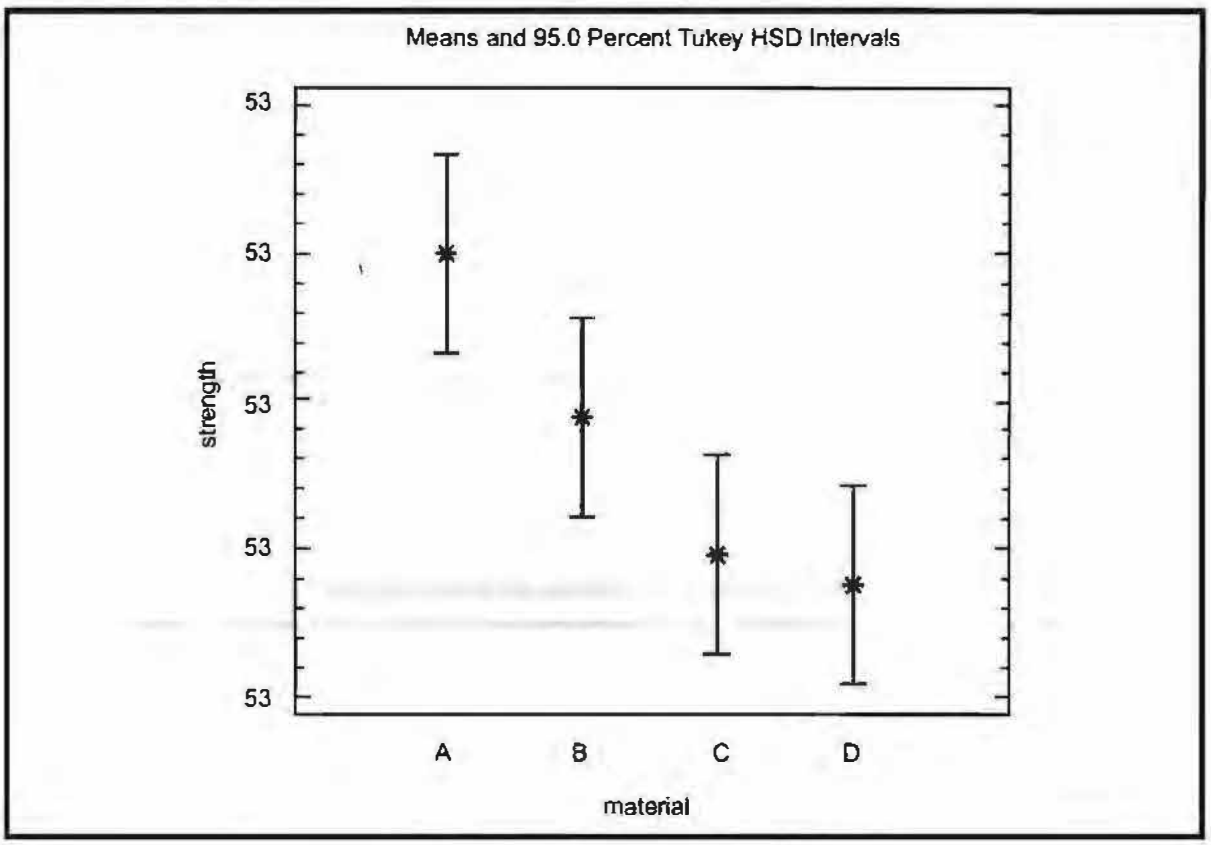


Compare this output to the output from the One-Way ANOVA.

One difference is the appearance of the ANOVA summary table. Now, there is a row labeled intercept and another labeled adjusted. You can ignore these. The F value for Group is still the same, and that is what we are interested in. Notice the eta squared column. What does it say for group? Does this value agree with the text? Unfortunately SPSS does not calculate Omega squared, so you would have to do this by hand. (Unfortunately, it also does not calculate any of the more useful effect size measures, such as d. Did the Bonferroni and the previous LSD multiple comparisons yield the same results?

A one-way analysis of variance is used when the data are divided into groups according to only one factor. The questions of interest are usually: (a) Is there a significant difference between the groups?, and

(b) If so, which groups are significantly different from which others? Statistical tests are provided to compare group means, group medians, and group standard deviations. When comparing means, multiple range tests are used, the most popular of which is Tukey's HSD procedure. For equal size samples, significant group differences can be determined by examining the means plot and identifying those intervals that do not overlap.



7.8 TWO WAY ANALYSIS OF VARIANCE PRACTICAL IN EXCEL SOLVER

Using a two-way analysis of variance, it is possible to test two hypotheses simultaneously. We can now test H_{01} : $\mu_1 = \mu_2 = \mu_3$ ..... and H_{02} : $m_1 = m_2 = m_3$ ....., where μ_1, μ_2, μ_3 ..... *etc.*, denote means due to type I treatments and m_1, m_2, m_3 ..... *etc.*, are the means due to type II treatments. For example, in addition to the testing of the hypothesis that mean sales due to different schemes of advertising are equal, we can also test, simultaneously, another hypothesis, say, the mean sales due to different types of after sales service are equal.

Let the number of type I treatments be k and the number of type II treatments be l. Thus, we can form $k \times l$ distinct combinations (or cells) of the two types of treatments. Each of these cell is assigned an experimental unit selected at random from $k \times l$ homogeneous units (the case of one observation per cell). These cells can be placed in the form of a table having k rows and l columns where its rows denote different levels of type I treatments and its columns denote different levels of type II treatment. For convenience, we shall term them as row and column treatments.

Columns → / *Rows* ↓	1	2	.....	l	*Total*
1	X_{11}	X_{12}	.....	X_{1l}	$T_{1\cdot}$
2	X_{21}	X_{22}	.....	X_{2l}	$T_{2\cdot}$
⋮	⋮	⋮	.....	⋮	⋮
k	X_{k1}	X_{k2}	.....	X_{kl}	$T_{k\cdot}$
Total	$T_{\cdot 1}$	$T_{\cdot 2}$	.....	$T_{\cdot l}$	T

We note that $\bar{X}_{i\cdot} = \frac{T_{i\cdot}}{l}$, $i = 1, 2, \ldots\ldots k$ and $\bar{X}_{\cdot j} = \frac{T_{\cdot j}}{k}$, $j = 1,2, \ldots\ldots l$.

Also $T = \sum_{i=1}^{k} T_{i\cdot} = \sum_{j=1}^{l} T_{\cdot j}$ and $\bar{X} = \frac{T}{kl}$, where $n = k \times l$.

Decomposition of Total Variation

We can write,

$$X_{ij} - \bar{X} = \left(X_{ij} - \bar{X}_{\cdot j} - \bar{X}_{i\cdot} + \bar{X}\right) + \left(\bar{X}_{\cdot j} - \bar{X}\right) + \left(\bar{X}_{i\cdot} - \bar{X}\right)$$

Squaring and taking sum over all the observations, we get

$$\sum_{i=1}^{k}\sum_{j=1}^{l}\left(X_{ij} - \bar{X}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{l}\left(X_{ij} - \bar{X}_{\cdot j} - \bar{X}_{i\cdot} + \bar{X}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{l}\left(\bar{X}_{\cdot j} - \bar{X}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{l}\left(\bar{X}_{i\cdot} - \bar{X}\right)^2$$

(various product terms can be shown to be equal to zero.)

$$\sum_{i=1}^{k}\sum_{j=1}^{l}\left(X_{ij} - \bar{X}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{l}\left(X_{ij} - \bar{X}_{\cdot j} - \bar{X}_{i\cdot} + \bar{X}\right)^2 + k\sum_{j=1}^{l}\left(\bar{X}_{\cdot j} - \bar{X}\right)^2 + l\sum_{i=1}^{k}\left(\bar{X}_{i\cdot} - \bar{X}\right)^2$$

(Total *SS*) = (Error *SS*) + (Column *SS*) + (Row *SS*)

To facilitate computational work, the expressions for various sums of squares can be transformed as:

$$TSS = \sum\sum X_{ij}^2 - \frac{T^2}{n}, \quad CSS = \frac{\sum T_{\cdot j}^2}{k} - \frac{T^2}{n} \quad \text{and} \quad RSS = \frac{\sum T_{i\cdot}^2}{l} - \frac{T^2}{n}$$

Analysis of Variance Table

Source of Variation	*Sum of Squares*	*d.f.*	*Mean SS*	*F Value*
Between Columns	$CSS = \frac{\sum T_{.j}^2}{k} - \frac{T^2}{n}$	$l-1$	$s_c^2 = \frac{CSS}{l-1}$	$F_c = \frac{s_c^2}{s_e^2}$
Between Rows	$RSS = \frac{\sum T_{i.}^2}{l} - \frac{T^2}{n}$	$k-1$	$s_r^2 = \frac{RSS}{k-1}$	$F_r = \frac{s_r^2}{s_e^2}$
Error	$ESS = TSS - CSS - RSS$	$(l-1)(k-1)$	$s_e^2 = \frac{ESS}{(l-1)(k-1)}$	
Total	$TSS = \sum\sum X_{ij}^2 - \frac{T^2}{n}$	$n-1$		

We note that the degrees of freedom of F_c are $(l-1)$, $(l-1)(k-1)$ and that for F_r are $(k-1)$, $(l-1)(k-1)$.

Remarks: The two-way ANOVA is also known as the *Randomised Block Design.* This technique removes the effect of difference in blocks (columns) on the treatments (rows) and vice-versa. Thus, RSS is free from the variations due to columns and CSS are free from the variations due to rows.

Example 7.7: The following data represent the sale (Rs '000) per month of three brands of a toilet soap allocated among three cities:

Cities → / *Brands ↓*	*A*	*B*	C
I	12	48	30
II	42	54	57
III	9	42	21

Test whether (i) the mean sales of the three brands are equal and (ii) the mean sales of the toilet soap in each city are equal.

Solution: We have to test $H_0 : \mu_1 = \mu_2 = \mu_3$, where μ_i denotes mean sale of brand i, and $H_0 : m_1 = m_2 = m_3$, where m_j denotes mean sale of soap in city j.

$$T_{.1} = 12 + 42 + 9 = 63, T_{.2} = 48 + 54 + 42 = 144, T_{.3} = 30 + 57 + 21 = 108$$

$$T_{1.} = 12 + 48 + 30 = 90, T_{2.} = 42 + 54 + 57 = 153, T_{3.} = 9 + 42 + 21 = 72$$

$$T = 63 + 144 + 108 = 315$$

$$TSS = 12^2 + 48^2 + 30^2 + 42^2 + 54^2 + 57^2 + 9^2 + 42^2 + 21^2 - \frac{315^2}{9} = 2538$$

$$(\textit{Between cities})\ CSS = \frac{63^2 + 144^2 + 108^2}{3} - \frac{315^2}{9} = 12123 - 11025 = 1098$$

$$(\textit{Between brands})\ RSS = \frac{90^2 + 153^2 + 72^2}{3} - \frac{315^2}{9} = 12231 - 11025 = 1206$$

$ESS = 2538 - 1098 - 1206 = 234$

Analysis of Variance Table

Source of Variation	*Sum of Squares*	*d.f.*	*Mean S.S.*	*F Value*
Between Cities	$CSS = 1098$	2	549	$F_c = \frac{549}{58.5} = 9.4$
Between Brands	$RSS = 1206$	2	603	$F_r = \frac{603}{58.5} = 10.3$
Error	$ESS = 234$	4	58.5	
Total	$TSS = 2538$	8		

The value of F from table for 2, 4 d.f. at 5% level of significance is 6.94. Thus, both the null hypotheses are rejected.

Check Your Progress 2

Fill in the blanks:

1. A ..... of variance is used when the data are divided into groups according to only one factor.
2. The primary reason for comparing two population variances is to test the ..... assumption to decide which type of t-procedure to use when performing inferences on two population means.

7.9 SUMMARY

Analysis of variance is an extension of the test of significance of the difference between two population means to the case involving simultaneous comparison of more than two means. Inference on population variances is also needed, to Place confidence intervals on or test for the size of the population variance and to compare variances from two populations.

A one-way analysis of variance is used when the data are divided into groups according to only one factor. The questions of interest are usually: (a) Is there a significant difference between the groups?, and (b) If so, which groups are significantly different from which others? Statistical tests are provided to compare group means, group medians, and group standard deviations.

7.10 KEYWORDS

- ANOVA
- Inference
- Two-way analysis of variance
- Population variance
- One-way analysis of variance

7.11 REVIEW QUESTIONS

1. Explain the meaning of the analysis of variance. Write down the one-way analysis of variance table for testing the homogeneity of k groups.
2. What type of null hypothesis is tested using analysis of variance? State basic assumptions of this analysis.
3. Explain the significance of the analysis of variance. What is the role of F distribution in this analysis?
4. Explain the nature of a two-way analysis of variance. Write down a general ANOVA table for a two-way classification.
5. Four varieties of rice were grown on each of the four beds of three identical plots. The output (in quintals) of different varieties is given in the following table: Test whether the mean yield of different varieties of rice is significantly different.
6. The data regarding life (in '00 hours) of three types of bulbs manufactured by a company are given in the following table. Test the hypothesis that mean life of bulbs of different types are same.

Type of Bulbs					
A	16	18	19		
B	14	13	15	20	
C	18	17	19	21	21

7. The Amrit Merchandising Company wishes to test whether its three salesmen *A*, *B* and *C*, tend to make sales of the same size or whether they differ in their selling ability as measured by the average size of their sales. During the last week there have been 14 sale calls, *A* made 5 calls, B made *4* calls and *C* made 5 calls. The weekly sales (in Rs) recorded for the three salesmen are given below. Perform the analysis and give your conclusions.

   A : 300, 400, 300, 500, 0 ; *B* : 600, 300, 300, 400 ; *C* : 700, 300, 400, 600, 500.

8. In an experiment conducted on a farm in a certain village of Rajasthan, the following information was collected regarding the yield in quintals per acre of 6 plots of wheat. Three out of six plots produced Sharbati wheat and the remaining three produced Australian wheat. Set up an analysis of variance table and find out if the variety differences are significant. Test at 5% level of significance.

	Yield in quintals		
Sharbati	10	15	11
Australian	13	12	17

9. The three samples given below have been obtained from three normal populations with equal variance. Test the hypothesis that the population means are equal at 5% level of significance.

   Sample I: 6 8 5 12 9,

   Sample II: 5 3 8 7 7,

   Sample III 10 7 11 10 12.

   (The value of F0.05 with 2, 12 d.f. = 3.89.)

10. The following table illustrates the sample psychological health ratings of corporate executives in the field of Banking, Manufacturing and Fashion retailing:

Banking	:	41	53	54	55	43
Manufacturing	:	45	51	48	43	39
Fashion Retailing	:	34	44	46	45	51

Can we consider the psychological health of corporate executives in the given three fields to be equal at 5% level of significance?

11. A sample of seven observations corresponding to the regression model, $Y = a + bX + \in$, where Y = profit (percent of turnover) and X = capital employed (Rs crores), gave the following data :

    (a) Estimate the regression line of profits (Y) on capital employed (X).

    (b) Prepare an analysis of variance table and use it to test the significance of regression at 5% level of significance.

    (c) Find r2 and interpret its value.

12. The following data represent the number of units of a commodity produced by 3 different workers using 3 different types of machines:

	Machines		
Workers	*A*	*B*	*C*
X	16	64	40
Y	56	72	56
Z	12	56	28

Test: (i) Whether mean productivity is same for the different types of machines, and (ii) whether the three workers differ with respect to mean productivity.

13. In a certain factory, production can be accomplished by four different workers on five different types of machines. A sample study, in the context of a two-way design without repeated values, is being made with two fold objective of examining whether the four workers differ with respect to mean productivity and whether the mean productivity is same for the five machines. The researcher involved in this study reports while analyzing the gathered data as under:

    (i) Sum of squares for variance between machines = 35.2

    (ii) Sum of squares for variance between workers = 53.8

    (iii) Sum of squares for total variance = 174.2

    Set up an ANOVA table for the given information and draw the inferences about the variances in mean productivities at 5% level of significance.

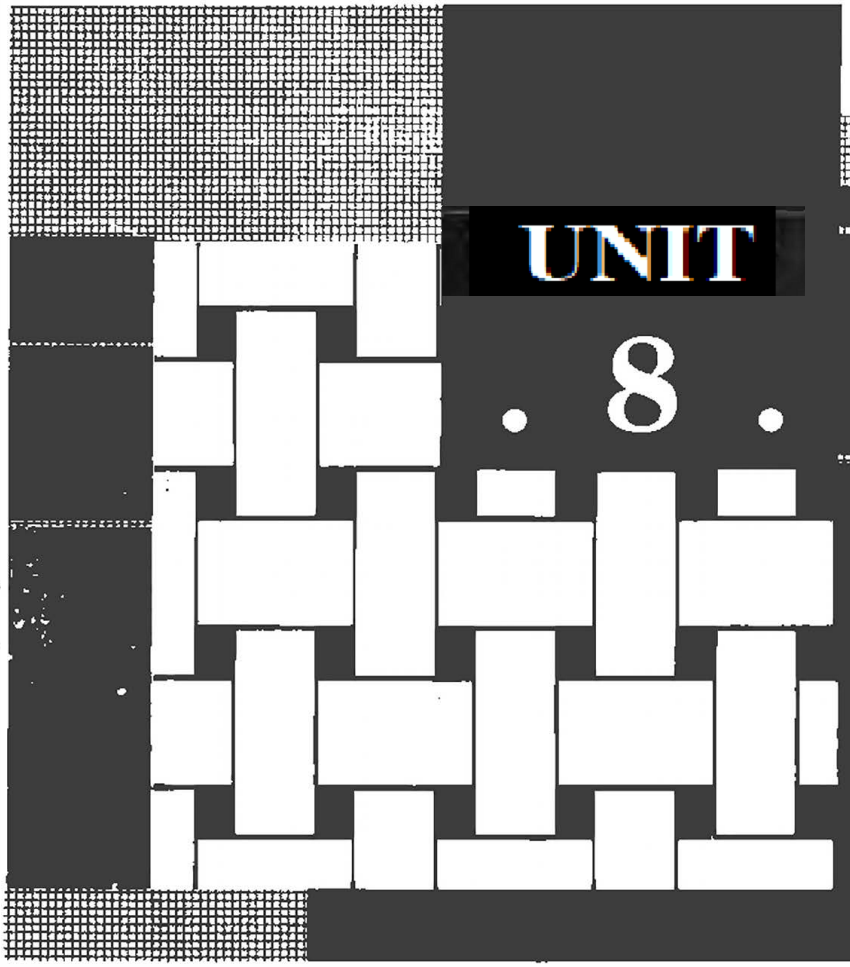
Answers to Check Your Progress

Check Your Progress 1

1. Group Variable.
2. Test of significance

7.12 REFERENCES AND FURTHER READING

- Oakshott, L. (2021). Essential quantitative methods: For business, management, and finance (6th ed.). Macmillan. ISBN: 9781137610890.
- Bell, M. L. (2020). Research methods and quantitative techniques (2nd ed.). Routledge. ISBN: 9781138473876.



UNIT 8

CHAPTER OUTLINE

8.1 INTRODUCTION

The tests of hypothesis, discussed so far, are known as parametric tests because these are based on the assumption that the concerned sample has been obtained from a population with known values of its one or more parameters. For example, the use of t-distribution to test the $H_0 : \mu_1 = \mu_2$ requires that the two samples are drawn from normal populations with equal variances. Similarly, the use of F-test in analysis of variance requires that various samples are obtained from normal populations with equal variances, etc. It should be pointed out that the validity of the results of a parametric test depends upon the appropriateness of these assumptions. Thus, when these assumptions are not met, the parametric tests are no longer applicable.

In contrast to parametric tests, non-parametric tests do not require any assumptions about the parameters or about the nature of population. It is because of this that these methods are sometimes referred to as the distribution free methods. Most of these methods, however, are based upon the weaker assumptions that observations are independent and that the variable under study is continuous with approximately symmetrical distribution. In addition to this, these methods do not require measurements as strong as that required by parametric methods. Most of the non-parametric tests are applicable to data measured in an ordinal or nominal scale. As opposed to this, the parametric tests are based on data measured at least in an interval scale. The measurements obtained on interval and ratio scale are also known as high level measurements.

Level of Measurement

(i) ***Nominal Scale:*** This scale uses numbers or other symbols to identify the groups or classes to which various objects belong. These numbers or symbols constitute a nominal or classifying scale. For example, classification of individuals on the basis of sex (male, female) or on the basis of level of education (matric, senior secondary, graduate, post graduate), etc. This scale is the weakest of all the measurements.

(ii) ***Ordinal Scale:*** This scale uses numbers to represent some kind of ordering or ranking of objects. However, the differences of numbers, used for ranking, don't have any meaning. For example, the top 4 students of class can be ranked as 1, 2, 3, 4, according to their marks in an examination.

(iii) ***Interval Scale:*** This scale also uses numbers such that these can be ordered and their differences have a meaningful interpretation.

(iv) ***Ratio Scale:*** A scale possessing all the properties of an interval scale along with a true zero point is called a ratio scale. It may be pointed out that a zero point in an interval scale is arbitrary. For example, freezing point of water is defined at 0° Celsius or 320 Fahrenheit, implying thereby that the zero on either scale is arbitrary and doesn't represent total absence of heat. In contrast to this, the measurement of distance, say in metres, is done on a ratio scale. The term ratio is used here because ratio comparisons are meaningful. For example, 100 kms of distance is four times larger than a distance of 25 kms while 1000 F may not mean that it is twice as hot as 50° F.

It should be noted here that a test that can be performed on high level measurements can always be performed on ordinal or nominal measurements but not vice-versa. However, if along with the high level measurements the conditions of a parametric test are also met, the parametric test should invariably be used because this test is most powerful in the given circumstances.

From the above, we conclude that a non-parametric test should be used when either the conditions about the parent population are not met or the level of measurements is inadequate for a parametric test.

Advantages

The non-parametric tests have gained popularity in recent years because of their usefulness in certain circumstances. Some advantages of non-parametric tests are mentioned below:

1. Non-parametric tests require less restrictive assumptions vis-a-vis a comparable parametric test.
2. These tests often require very few arithmetic computations.
3. There is no alternative to using a non-parametric test if the data are available in ordinal or nominal scale.
4. None of the parametric tests can handle data made up of samples from several populations without making unrealistic assumptions. However, there are suitable non-parametric tests available to handle such data.

Disadvantages

1. It is often said that non-parametric tests are less efficient than the parametric tests because they tend to ignore a greater part of the information contained in the sample. Inspite of this, it is argued that although the non-parametric tests are less efficient, a researcher using them has more confidence in using his methodology than he does if he must adhere to the unsubstantial assumptions inherent in parametric tests.
2. The non-parametric tests and their accompanying tables of significant values are widely scattered in various publications. As a result of this, the choice of most suitable method, in a given situation, may become a difficult task.

8.2 THE MATCHED-PAIRS SIGN TEST

When the same individual is simultaneously observed with regard to two characteristics, we get a matched pair. For example, a survey of obtaining the opinions of persons regarding two brands of detergents, say X and Y. Given a sample of matched pairs, we can test which of the two detergents is preferable.

The Test Statistic

Let there be n matched pairs (X_i, Y_i), $i = 1, 2, \ldots\ldots n$, in the sample, where X_i and Y_i are the ranks of the respective item X and Y by the i th individual. We associate a plus sign to a pair if $X_i > Y_i$, a minus sign if $X_i < Y_i$ and discard it if $X_i = Y_i$. Let p be the proportion of plus signs, i.e., $p = \dfrac{\textit{Number of plus signs}}{\textit{Total number of matched pairs}}$

Let π be the proportion of plus signs in the population, i.e., the proportion of individuals having preference for X. We note that if more individuals have preference for X, then $\pi > \frac{1}{2}$. Similarly, the indifference of the individuals is indicated by $\pi = \frac{1}{2}$. In general, we can test $H_0 : \pi =$ or $<$ or $> \pi_0$, where π_0 denotes proportion of individuals having preference for X.

It can be shown that p is a random variable which approximately follows a normal distribution (when $n \geq 25$) with mean π and standard error $\sqrt{\dfrac{\pi(1-\pi)}{n}}$.

Example 8.1: A random sample of 40 persons was selected to determine their preference regarding the two brands of a new tooth paste, X and Y. Each person used the two brands for one month and then was asked to rank them by using arbitrary numbers. Their rankings, X and Y, were recorded as follows:

Person	:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
X	:	4	15	2	4	15	3	8	6	10	8	1	5	12	6	2	7	5	1	3	8
Y	:	2	15	1	5	16	2	6	5	15	8	2	7	10	4	2	3	10	2	4	9
Person	:	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
X	:	12	20	3	1	2	1	4	6	28	100	15	4	5	30	9	8	5	10	12	25
Y	:	15	40	2	2	1	4	4	7	25	200	20	3	6	40	15	15	3	2	18	30

Test the hypothesis that more than half the population prefer brand X to Y.

Solution: We have to test $H_0 : \pi \leq \frac{1}{2}$ against $H_a : \pi > \frac{1}{2}$

We assign a plus sign to a pair if $X > Y$, a minus sign if $X < Y$ and discard the pair if $X = Y$. Proceeding in this manner, we find that the number of positive signs in the given sample is 14. Also the total number of pairs, after discarding the cases where $X = Y$, is $n = 36$. Thus, we have

$$p = \frac{14}{36} \text{ and the standard error of } p, \text{ i.e., } \sigma_p = \sqrt{\frac{1}{2} \times \frac{1}{2} \times \frac{1}{36}} = 0.083$$

Now $z = \dfrac{\frac{16}{36} - \frac{1}{2}}{0.083} = -1.339$. Since this value is greater than – 1.645, there is no evidence against H_0 at 5% level of significance. Thus, the sample evidence does not support the view that more than half of the population prefer X to Y.

8.3 WILCOXON MATCHED-PAIRS SIGNED RANK-SUM TEST

This test is a variant of the Wilcoxon Signed Rank Test and can be used to test the hypothesis regarding the equality of medians in paired samples.

The Test Statistics

Various steps for the calculation of test statistics are as follows:

(i) For each pair, calculate the difference $d_i = X_i - Y_i$.

(ii) Omit all observation(s) with equal values and reduce the sample size accordingly.

(iii) Rank these differences in ascending order without regard to their signs.

(iv) The cases of tied ranks are assigned ranks by the average method.

(v) Find T_+ and T_-, where T_+ is the sum of ranks with positive d_i and T_- is the sum of ranks with negative d_i.

(vi) The test statistics, to be denoted by T, is given by T_+ or T_- or minimum of T_+ and T_- for the respective $H_0 : M >$ or $<$ or $= M_d$.

It can be shown that the distribution of T is approximately normal (when $n > 25$) with mean $\mu_T = \frac{n(n+1)}{4}$ and standard error $\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$.

Example 8.2: Two models of a machine are under consideration for purchase. An organisation has one of each type for trial and each operator, out of the team of 25 operators, uses each machine for a fixed length of time. Their outputs are:

Operator No.	:	1	2	3	4	5	6	7	8	9	10	11	12	13
Output from Machine I	:	82	68	53	75	78	86	64	54	62	70	51	80	64
Output from Machine II	:	80	71	46	58	60	72	38	60	65	64	38	79	37
Operator No.	:	14	15	16	17	18	19	20	21	22	23	24	25	
Output from Machine I	:	65	70	55	75	64	72	55	70	45	64	58	65	
Output from Machine II	:	60	73	48	58	60	76	60	50	30	70	55	60	

Is there any significant difference between the output capacities of the two machines? Test at 5% level of significance.

Solution:

O.No.	M_1	M_2	d_i	*Ranks*	*O.No.*	M_1	M_2	d_i	*Ranks*
1	82	80	2	2	14	65	60	5	10
2	68	71	−3	4.5	15	70	73	−3	4.5
3	53	46	7	15.5	16	55	48	7	15.5
4	75	58	17	20.5	17	75	58	17	20.5
5	78	60	18	22	18	64	60	4	7.5
6	86	72	14	18	19	72	76	−4	7.5
7	64	38	26	24	20	55	60	−5	10
8	54	60	−6	13	21	70	50	20	23
9	62	65	−3	4.5	22	45	30	15	19
10	70	64	6	13	23	64	70	−6	13
11	51	38	13	17	24	58	55	3	4.5
12	80	79	1	1	25	65	60	5	10
13	64	37	27	25					

Note: M_1 and M_2, in the above table, denote the outputs of the machine I and II respectively.

From the above table, we have $T_+ = 268$ and $T_- = 57$. Thus, $T = 57$.

We have to test H_0 : Output capacities of the two machines are same against H_a: Output capacities are not same.

The mean and standard error of the statistic are $\mu_T = \frac{25 \times 26}{4} = 162.5$ and $\sigma_T = \sqrt{\frac{25 \times 26 \times 51}{24}} = 37.2$

$\therefore \quad z = \frac{|57 - 162.5|}{37.2} = 2.84 > 1.96$. Thus, H_0 is rejected at 5% level of significance.

Check Your Progress 1

Fill in the blanks:

1. Most of the non-parametric tests are applicable to data measured in a .....
2. .....scale also uses numbers such that these can be ordered and their differences have a meaningful interpretation.
3. It is often said that non-parametric tests are less efficient than the .....

8.4 MANN WHITNEY WILCOXON TEST

This test is used to test the hypothesis that two independent samples have come from two populations with equal means or medians. The test can be a one or a two tailed test. The basic assumption of the test is that the distributions of the two populations are continuous with equal standard deviations.

The Test Statistic

Let n_1 and n_2 be the sizes of the samples taken from populations I and II respectively. Various steps for obtaining the test statistic are as follows:

(i) Rank all the $n_1 + n_2$ observations, arrange in ascending order.

(ii) Find R_1 and R_2, where R_i denotes the sum of ranks of the i th sample (i = 1, 2).

It can be shown that when each n_1 or n_2 is at least 10, the distribution of R_1 (or R_2) will be approximately normal with mean

$$\mu_1 = \frac{n_1(n_1 + n_2 + 1)}{2} \left(or\ \mu_2 = \frac{n_2(n_1 + n_2 + 1)}{2} \right)$$

and standard error $\sigma = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$. (Note : *S.E.* is same in both cases)

Example 8.3: For a random sample of 10 pigs, fed on diet *A*, the increase in weights, in pounds, in a certain period were : 10, 6, 16, 17, 13, 12, 8, 14, 15, 9.

For another sample of 12 pigs, fed on diet *B*, the increase in weights, in pounds, were : 7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 17.

Test the hypothesis that mean increase in weight is more for the pigs fed on diet *B*.

Solution: It is given that n_1 = 10 and n_2 = 12.

We have to test $H_0 : \mu_A \geq \mu_B$ against $\mu_A < \mu_B$.

The ranking of the sample observations is done in the following table.

Sample I	10	6	16	17	13	12	8	14	15	9			
Ranks	6.5	1	16	17.5	10.5	8.5	3.5	12.5	14.5	5			95.5
Sample II	7	13	22	15	12	14	18	8	21	23	10	17	
Ranks	2	10.5	21	14.5	8.5	12.5	19	3.5	20	22	6.5	17.5	

From the above table we get $R_1 = 95.5$.

Also $\mu_1 = \frac{10 \times 23}{2} = 115$ and $\sigma = \sqrt{\frac{10 \times 12 \times 23}{12}} = 15.2$.

$\therefore \quad z = \frac{95.5 - 115}{15.2} = -1.28 > -1.645.$

Thus, there is no evidence against H_0 at 5% level of significance.

8.5 THE KRUSKAL-WALLIS TEST

This test is used to determine whether k independent samples can be regarded to have been obtained from identical populations with respect to their means. The Kruskal-Wallis Test is the non-parametric counter part of the one-way analysis of variance. The assumption of the *F*-test, used in analysis of variance, was that each of the k populations should be normal with equal variance. In contrast to this, the Kruskal-Wallis test only assumes that the k populations are continuous and have the same pattern (symmetrical or skewed) of distribution. The null and the alternative hypotheses of the Kruskal-Wallis test are:

$H_0 : \mu_1 = \mu_2 = \ldots\ldots = \mu_k$ (*i.e.*, means of the k populations are equal)

H_a : Not all μ_i's are equal.

The Test Statistic

The computation of the test statistic follows a procedure that is very similar to the Mann-Whitney Wilcoxon test.

(i) Rank all the $n_1 + n_2 + \ldots\ldots + n_k = n$ observations, arrayed in ascending order.

(ii) Find $R_1, R_2, \ldots\ldots R_k$, where R_i is the sum of ranks of the i th sample.

The test statistic, denoted by H, is given by

$$H = \frac{12}{n(n+1)}\left(\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \ldots\ldots + \frac{R_k^2}{n_k}\right) - 3(n+1)$$

It can be shown that the distribution of H is χ^2 with $k - 1$ *d.f.*, when size of each sample is at least 5. Thus, if $H > \chi^2_{k-1}$, H_0 is rejected.

Example 8.4: To compare the effectiveness of three types of weight-reducing diets, a homogeneous groups of 22 women was divided into three sub-groups and each sub-group followed one of these diet plans for a period of two months. The weight reductions, in kgs, were noted as given below:

Diet Plans	*I*	4.3	3.2	2.7	6.2	5.0	3.9		
	II	5.3	7.4	8.3	5.5	6.7	7.2	8.5	
	III	1.4	2.1	2.7	3.1	1.5	0.7	4.3	3.5	0.3

Use the Kruskal-Wallis test to test the hypothesis that the effectiveness of the three weight reducing diet plans are same at 5% level of significance.

Solution: It is given that $n_1 = 6$, $n_2 = 7$ and $n_3 = 9$.

The total number observations is $6 + 7 + 9 = 22$. These are ranked in their ascending order as given below:

Diet Plans	*I*	12.5	9	6.5	17	14	11			70
	II	15	20	21	16	18	19	22		131
	III	3	5	6.5	8	4	2	12.5	10	1	52

From the above table, we get $R_1 = 70$, $R_2 = 131$ and $R_3 = 52$.

$$\therefore \quad H = \frac{12}{22 \times 23}\left(\frac{70^2}{6} + \frac{131^2}{7} + \frac{52^2}{9}\right) - 3 \times 23 = 15.63.$$

The tabulated value of χ^2 at 2 *d.f.* and 5% level of significance is 5.99. Since H is greater than this value, H_0 is rejected at 5% level of significance.

8.6 THE RUNS TEST FOR RANDOMNESS

The basic assumption in all statistical tests is that the sample obtained from a population be random. The runs test can be used to test whether a given sample is random or not.

A run is defined as a sequence of identical symbols which are preceded and followed by different or no symbols at all. For example, suppose that a sequence of two symbols, *A* and *B*, occurred as follows:

A B AA B A BBB AAA BB A

The number of runs in the above sequence are 9.

It may be pointed out here that when there are n observations, where each is denoted by either symbol *A* or by *B*, the number of possible runs would lie between and including 2 to n. A sample having unusually small or large number of runs is considered as an extreme case and thus, cannot be regarded as random.

Thus, to the test of randomness of a sample , we test H_0 : The sample is random against H_a : The sample is not random. These hypotheses indicate that when the number of runs is significantly large or small, H_0 is rejected.

The Test Statistic

Let n_1 be the number of symbols of one type and n_2 be the number of symbols of other type such that $n = n_1 + n_2$ is the total number of observations in the sequence. Further, let r be the number of runs in the sequence.

Using algebra, it can be shown that r is a random variable with mean $\mu_r = \frac{2n_1n_2}{n} + 1$ and standard error $\sigma_r = \sqrt{\frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}}$. These results can be verified by taking certain specific cases. For example if $n_1 = 1$ and $n_2 = 2$, we can write all possible sequences as *ABB*, *BAB* and *BBA*. The associated values of r are 2, 3, and 2 respectively. Thus, we have

$$\mu_r = \frac{2+3+2}{3} = \frac{7}{3} \text{ and } \sigma_r = \sqrt{\frac{4+9+4}{3} - \frac{49}{9}} = \sqrt{\frac{2}{9}}$$

We note the same result can be obtained by substituting $n_1 = 1$ and $n_2 = 2$ in the formulae for μ_r and σ_r.

When both n_1 and n_2 are at least 10, the distribution of r can be approximated by a normal distribution. Thus, the decision rule is as follows :

$$\text{If } z = \frac{|r - \mu_r|}{\sigma_r} > 1.96, \text{ reject } H_0 \text{ at 5\% level of significance.}$$

Example 8.5: The weights (gms) of 31 apples picked from a consignment are as follows :

106, 107, 76, 82, 106, 107, 115, 93, 187, 95, 123, 125, 111, 92, 86, 70, 127, 68, 130, 129, 139, 119, 115, 128, 100, 186, 84, 99, 113, 204, 111.

Test the hypothesis that the sample is random.

Solution: We have to test $H_0 : r = \mu_r$ against $H_a : r \neq \mu_r$.

Let us denote the increase in the successive observation by a plus (+) sign and the decrease of successive observation by a minus (–) sign. From the given observations, we can write a sequence of plus and minus signs, as given below :

$$+ - + + + + - + - + + - - - + - + - + - - + - + - + + + -$$

From the above sequence, we have $n_1 = 16$ (the number of plus signs), $n_2 = 14$ (the number minus signs) and $r = 20$ (the number of runs). Also $n = 16 + 14 = 30$.

Thus, $$\mu_r = \frac{2 \times 16 \times 14}{30} + 1 = 15.93$$

and $$\sigma_r = \sqrt{\frac{2 \times 16 \times 14(2 \times 16 \times 14 - 30)}{900 \times 29}} = 2.68$$

Further, $$z = \frac{|20 - 15.93|}{2.68} = 1.52$$

Since this value is less than 1.96, there is no evidence against H_0 at 5% level of significance. Thus, the given sample may be treated as random.

Check Your Progress 2

Fill in the blanks:

1. The basic assumption in all statistical tests is that the sample obtained from a population be .....

2. .....test is used to test the hypothesis that two independent samples have come from two populations with equal means or medians.

8.7 SUMMARY

1. *The Runs Test for Randomness:*

   If n_1 is the number of symbols of one type and n_2 is the number of symbols of the other type such that each of them is at least 10 and $n = n_1 + n_2$, then the distribution of r, the number of runs in the sample, is approximately normal with $\mu_r = \frac{2n_1n_2}{n} + 1$ and $\sigma_r = \sqrt{\frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}}$. The test of significance is a two tailed test.

2. *The Wilcoxon Signed Rank Test:*

   This test is used to test whether the given sample can be regarded to have come from a population with a specified value of median. When $n > 25$, the statistic T is a normal variate with $\mu_T = \frac{n(n+1)}{4}$ and $\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$.

3. *The Matched-Pairs Sign Test:* This test is used to test the hypothesis that a proportion in population is greater, less or equal to a given value. When $n \geq 25$, the distribution of the sample proportion of rankings is a normal variate with mean = π and $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$.

4. *The Wilcoxon-Matched-Pairs Signed Rank-Sum Test:*

   This test is used to test a one or a two tailed hypothesis regarding equality of medians. When $n > 25$, the statistic T is a normal variate with $\mu_T = \frac{n(n+1)}{4}$ and $\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$.

5. *The Mann-Whitney Wilcoxon Test:*

   This test is used to test a one or a two tailed hypothesis regarding equality of means or medians when the two samples are independent.

   When both n_1 and n_2 are at least 10, the rank sum R_1 or (R_2) is a normal variate with

   $$\mu_1 = \frac{n_1(n_1+n_2+1)}{2} \left(\text{or } \mu_2 = \frac{n_2(n_1+n_2+1)}{2} \right) \text{ and } \sigma = \sqrt{\frac{n_1n_2(n_1+n_2+1)}{12}}.$$

6. *The Kruskal Wallis Test:*

This test is a non-parametric counterpart of the one way analysis of variance. The test statistic is $H = \frac{12}{n(n+1)}\left(\frac{R_1^2}{n_1}+\frac{R_2^2}{n_2}+\cdots\cdots+\frac{R_k^2}{n_k}\right)-3(n+1)$, a χ^2 variate with $(k-1)$ *d.f.* when each sample size is at least 5.

7. *The Spearman's Rank Correlation Test:*

The test statistic $z = r_s\sqrt{n-1}$ is a standard normal variate ($n \geq 10$).

8. *The Kendalls Test of Concordance:*

The test statistic is $W = \frac{12\sum R^2 - 3n\left[k(n+1)\right]^2}{kn(n+1)}$, a χ^2 variate with $(n-1)$ *d.f.* when each sample size is at least 8.

8.8 KEYWORDS

- Nominal Scale
- Ordinal Scale
- Interval Scale
- Ratio Scale

8.9 REVIEW QUESTIONS

1. Distinguish between parametric and non-parametric methods for testing a statistical hypothesis.
2. What are the advantages and disadvantages of non-parametric methods as compared to parametric methods in statistics?
3. The following figures are a sample of 35 observations. Find median of the data and mark each measurement as *A*, if greater than it, or *B*, if less than it. Use the runs test to find whether the sample is random, at 5% level of significance.

   37, 46, 33, 39, 59, 41, 49, 44, 51, 35, 41, 55, 27, 19, 35, 41, 49, 21, 35, 37, 53, 29, 49, 48, 35, 47, 31, 41, 29, 27, 49, 63, 37, 13, 20.

4. The following are the number of units produced by a group of workers for 25 days. Use runs test and median test to test whether the data can be regarded as a random sample?

   210, 180, 170, 240, 150, 215, 198, 181, 237, 209, 165, 176, 224, 201, 181, 252, 219, 154, 197, 235, 182, 167, 214, 221, 243.

5. A random sample of 50 consumers rated brand *X* and brand *Y* of soft drinks on a scale of 1 to 4, where 4 means the highest preference. The sample data pairs had 33 plus signs, 15 minus signs and 2 zeros. Perform the matched pairs sign test at 5% level of significance to test the hypothesis that more than 60% consumers rate brand *X* higher to brand *Y*.
6. To compare the price of a certain commodity in two towns, ten shops were selected at random in each town. The prices were recorded as follows:

Town A	:	61	60	56	63	56	63	59	56	44	62
Town B	:	55	54	47	59	51	61	57	54	64	58

Use Mann Whitney Wilcoxon method to test the hypothesis that average price is same in the two towns.

7. To compare the average weekly power costs of two factories, independent samples of sizes 12 and 10 are taken from the records of last year. The observations are given below:

Factory I	:	201	225	209	192	190	210	229	223	207	215	198	212
Factory II	:	182	167	240	190	182	200	185	165	187	184		

Use Mann Whitney Wilcoxon method to test the assertion that weekly power costs are higher in factory I.

8. A cooperative store is interested in knowing whether there is any significant difference between the buying habits of male and female shoppers. Samples of 14 males and 16 female shoppers gave the following information:

Male	:	62	38	43	79	77	23	11	52	33	41	70	49	69	43	
Female	:	931	01	72	1181	00	45	68	72	47	83	921	06	63	66	85	81

Use median test to verify whether there is any reason to suppose that the two populations are different.

9. Three different methods of advertising a commodity were used and the respective samples of sizes 9, 10 and 10 identical outlets were taken. The increased sales (in Rs '000) were recorded as follows:

Sample I	:	92	79	77	93	99	93	71	87	98	
Sample II	:	95	76	84	85	89	90	72	82	68	83
Sample III	:	81	91	75	80	78	94	100	86	88	69

Use Kruskal Walli's method to test the hypothesis that mean increase in sales due to the three methods of advertising is same.

10. Random samples of three models of a scooter were tested for the petrol mileage (the number of kilometers per litre). Use Kruskal Walli test to determine if the average mileage of the three models is same.

Model A	:	60	54	76	48	66	52	62	56
Model B	:	62	58	52	48	70			
Model C	:	42	64	36	65	42	60	82	

Answers to Check Your Progress

Check Your Progress 1

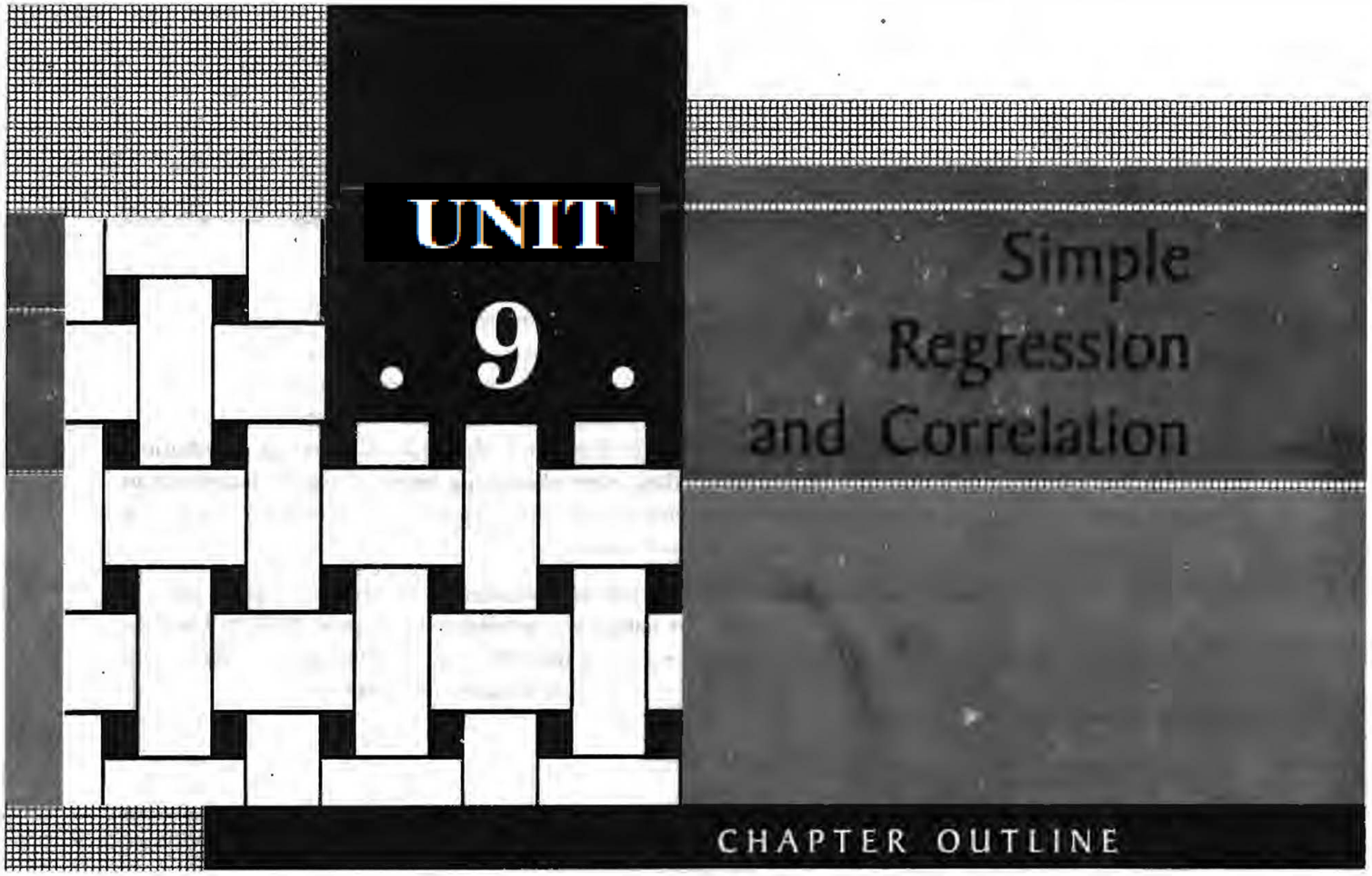
1. Ordinal or nominal scale
2. Interval
3. Parametric test

Check Your Progress 2

1. Random
2. Mann Whitney Wilcoxon Test

8.10 REFERENCES AND FURTHER READING

- Lee, N., & Peters, M. (2021). Applied statistics for business and management using Microsoft Excel. Wiley. ISBN: 9781292243578.
- Arnold, R. (2024). Quantitative methods for management. Oxford University Press. ISBN: 9780198744488.
- Taha, H. A. (2022). Operations research: An introduction (11th ed.). Pearson. ISBN: 9781292266973.
- Francis, A. (2020). Mathematics and statistics for business (7th ed.). Cengage Learning. ISBN: 9781408093829.



CHAPTER OUTLINE

9.1 INTRODUCTION

So far we have considered distributions relating to single characteristics. Such distributions are known as Univariate Distribution. When various units under consideration are observed simultaneously, with regard to two characteristics, we get a Bivariate Distribution. For example, the simultaneous study of the heights and weights of students of a college. For such data also, we can compute mean, variance, skewness etc. for each individual characteristics. In addition to this, in the study of a bivariate distribution, we are also interested in knowing whether there exists some relationship between two characteristics or in other words, how far the two variables, corresponding to two characteristics, tend to move together in same or opposite directions i.e. how far they are associated.

The regression equations are useful for predicting the value of dependent variable for given value of the independent variable. As pointed out earlier, the nature of a regression equation is different from the nature of a mathematical equation, e.g., if $Y = 10 + 2X$ is a mathematical equation then it implies that Y is exactly equal to 20 when $X = 5$. However, if $Y = 10 + 2X$ is a regression equation, then $Y = 20$ is an average value of Y when

$$X = 5.$$

9.2 TYPES OF RELATIONSHIPS

For a bivariate data (X_i, Y_i), $i = 1, 2, \ldots\ldots n$, we can have either X or Y as independent variable. If X is independent variable then we can estimate the average values of Y for a given value of X. The relation used for such estimation is called regression of Y on X. If on the other hand Y is used for estimating the average values of X, the relation will be called regression of X on Y. For a bivariate data, there will always be two lines of regression. It will be shown later that these two lines are different, i.e., one cannot be derived from the other by mere transfer of terms, because the derivation of each line is dependent on a different set of assumptions.

Line of Regression of Y on X

The general form of the line of regression of Y on X is $YC_i = a + bX_i$, where YC_i denotes the average or predicted or calculated value of Y for a given value of $X = X_i$. This line has two constants, a and b. The constant a is defined as the average value of Y when $X = 0$. Geometrically, it is the intercept of the line on Y-axis. Further, the constant b, gives the average rate of change of Y per unit change in X, is known as the regression coefficient.

The above line is known if the values of a and b are known. These values are estimated from the observed data (X_i, Y_i), $i = 1, 2, \ldots\ldots n$.

Note: It is important to distinguish between YC_i and Y_i. Whereas Y_i is the observed value, YC_i is a value calculated from the regression equation.

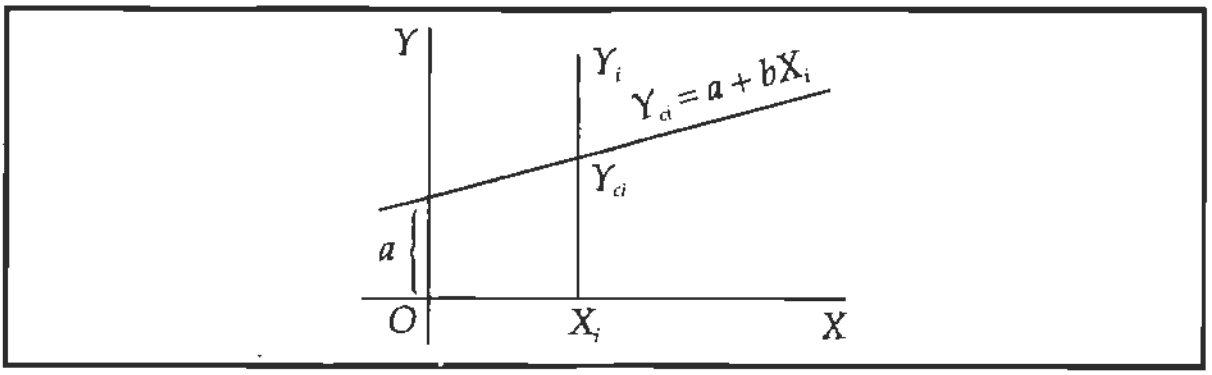


Figure 9.1

Lines of Regression

Using the regression $YC_i = a + bX_i$, we can obtain Y_{C1}, Y_{C2},..... Y_{Cn} corresponding to the X values X_1, X_2, ... X_n respectively. The difference between the observed and calculated value for a particular value of X say X_i is called error in estimation of the i th observation on the assumption of a particular line of regression. There will be similar type of errors for all the n observations. We denote by $e_i = Y_i - Y_{Ci}$ (i = 1, 2, ... n), the error in estimation of the i th observation. As is obvious from figure 9.1, e_i will be positive if the observed point lies above the line and will be negative if the observed point lies below the line. Therefore, in order to obtain a figure of total error, e_i's are squared and added. Let S denote the sum of squares of these errors, i.e.,

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - Y_{Ci})^2$$

Line of Regression of X on Y

The general form of the line of regression of X on Y is $X_{Ci} = c + dY_i$, where X_{Ci} denotes the predicted or calculated or estimated value of X for a given value of $Y = Y_i$ and c and d are constants. d is known as the regression coefficient of regression of X on Y.

In this case, we have to calculate the value of c and d so that

$S = \Sigma(X_i - X_{Ci})^2$ is minimised.

As in the previous section, the normal equations for the estimation of c and d are

$$\Sigma X_i = nc + d\Sigma Y_i \quad \text{.... (1)}$$

and

$$\Sigma X_i Y_i = c\Sigma Y_i + d\Sigma Y_i^2 \quad \text{.... (2)}$$

Dividing both sides of equation (1) by n, we have $\bar{X} = c + d\bar{Y}$.

This shows that the line of regression also passes through the point $(\bar{X}, \bar{Y})$. Since both the lines of regression passes through the point $(\bar{X}, \bar{Y})$, therefore $(\bar{X}, \bar{Y})$ is their point of intersection.

We can write $c = \bar{X} - d\bar{Y}$ .... (3)

As before, the various expressions for d can be directly written, as given below:

$$d = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum Y_i^2 - n\bar{Y}^2} \quad \text{.... (4)}$$

or $$d = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (Y_i - \bar{Y})^2} \quad \text{.... (5)}$$

or $$d = \frac{\sum x_i y_i}{\sum y_i^2} \quad \text{.... (6)}$$

$$= \frac{\frac{1}{n}\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum (Y_i - \bar{Y})^2} = \frac{Cov(X,Y)}{\sigma_Y^2} \quad \text{.... (7)}$$

Also $$d = \frac{n\sum X_i Y_i - (\sum X_i)(\sum Y_i)}{n\sum Y_i^2 - (\sum Y_i)^2} \quad \text{.... (8)}$$

This expression is useful for calculating the value of d. Another short-cut formula for the calculation of d is given by

$$d = \frac{b}{k}\left[\frac{n\sum u_i v_i - (\sum u_i)(\sum v_i)}{n\sum v_i^2 - (\sum v_i)^2}\right] \quad \text{.... (9)}$$

where $$u_i = \frac{X_i - A}{b} \text{ and } v_i = \frac{Y_i - B}{k}$$

Consider equation (7)

$$d = \frac{Cov(X,Y)}{\sigma_Y^2} = \frac{r\sigma_X\sigma_Y}{\sigma_Y^2} = r \cdot \frac{\sigma_X}{\sigma_Y} \quad \text{.... (10)}$$

Substituting the value of c from equation (3) into line of regression of X on Y we have

$$X_C = \bar{X} - d\bar{Y} + dY_i \text{ or } (X_C - \bar{X}) = d(Y_i - \bar{Y}) \quad \text{.... (11)}$$

or $$(X_{Ci} - \bar{X}) = r \cdot \frac{\sigma_X}{\sigma_Y}(Y_i - \bar{Y}) \quad \text{.... (12)}$$

Remarks: It should be noted here that the two lines of regression are different because these have been obtained in entirely two different ways. In case of regression of Y on X, it is assumed that the values of X are given and the values of Y are estimated by minimising $S(Y_i - Y_{Ci})^2$ while in case of regression of X on Y, the values of Y are assumed to be given and the values of X are estimated by minimising $S(X_i - X_{Ci})^2$. Since these two lines have been estimated on the basis of different assumptions, they are not reversible, i.e., it is not possible to obtain one line from the other by mere transfer of terms. There is, however, one situation when these two lines will coincide. From the study of correlation we may recall that when $r = \pm 1$, there is perfect correlation between the variables and all the points lie on a straight line. Therefore, both the lines of regression coincide and hence they are also reversible in this

case. By substituting $r = \pm 1$ in equation (12) or (24) it can be shown that the lines of regression in both the cases become

$$\left(\frac{Y_i - \bar{Y}}{\sigma_Y}\right) = \pm\left(\frac{X_i - \bar{X}}{\sigma_X}\right)$$

Further when $r = 0$, equation $Y_G - Y = r \times \frac{\sigma_Y}{\sigma_X}\left(X_i - \bar{X}\right)$ becomes $Y_G = \bar{Y}$ and equation (12) becomes $X_G = \bar{X}$. These are the equations of lines parallel to X-axis and Y-axis respectively. These lines also intersect at the point $\left(\bar{X}, \bar{Y}\right)$ and are mutually perpendicular at this point.

Correlation Coefficient and the Two Regression Coefficients

Since $b = r \times \frac{\sigma_Y}{\sigma_X}$ and $d = r \times \frac{\sigma_X}{\sigma_Y}$, we have

$b.d = r\frac{\sigma_Y}{\sigma_X} \times r\frac{\sigma_X}{\sigma_Y} = r^2$ or $r = \sqrt{b.d}$. This shows that correlation coefficient is the geometric mean of the two regression coefficients.

Remarks: The following points should be kept in mind about the coefficient of correlation and the regression coefficients:

(i) Since $r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$, $b = \frac{Cov(X,Y)}{\sigma_X^2}$ and $d = \frac{Cov(X,Y)}{\sigma_Y^2}$, therefore the sign of r, b and d will always be same and this will depend upon the sign of Cov (X, Y).

(ii) Since $bd = r^2$ and $0 \le r^2 \le 1$, therefore either both b and d are less than unity or if one of them is greater than unity, the other must be less than unity such that $0 \le b.d \le 1$ is always true.

Example 9.1: Obtain the two regression equations and find correlation coefficient between X and Y from the following data:

X	:	10	9	7	8	11
Y	:	6	3	2	4	5

Solution:

Calculation Table

X	Y	XY	X^2	Y^2
10	6	60	100	36
9	3	27	81	9
7	2	14	49	4
8	4	32	64	16
11	5	55	121	25
45	20	188	415	90

(a) ***Regression of Y on X***

$$b = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{n\sum X^2 - \left(\sum X\right)^2} = \frac{5\times 188 - 45\times 20}{5\times 415 - (45)^2} = 0.8$$

Also, $\bar{X} = \frac{45}{5} = 9$ and $\bar{Y} = \frac{20}{5} = 4$

Now $a = \bar{Y} - b\bar{X} = 4 - 0.8 \times 9 = -3.2$

∴ Regression of Y on X is $Y_C = -3.2 + 0.8X$

(b) ***Regression of X on Y***

$$d = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{n\sum Y^2 - \left(\sum Y\right)^2} = \frac{5\times 188 - 45\times 20}{5\times 90 - (20)^2} = 0.8$$

Also, $c = \bar{X} - d\bar{Y} = 9 - 0.8 \times 4 = 5.8$

∴ The regression of X on Y is $X_C = 5.8 + 0.8Y$

(c) ***Coefficient of correlation*** $r = \sqrt{b.d} = \sqrt{0.8\times 0.8} = 0.8$

9.3 ESTIMATION USING THE REGRESSION LINE

We may recall that Y_C and X_C are the estimated values from the regressions of Y on X and X on Y respectively.

Consider the regression equation $Y_{Ci} - \bar{Y} = b\left(X_i - \bar{X}\right)$.

Taking sum over all the observations, we get

$$\sum\left(Y_{Ci} - \bar{Y}\right) = b\sum\left(X_i - \bar{X}\right) = 0$$

$$\Rightarrow \quad \sum Y_{Ci} - n\bar{Y} = 0 \quad \text{or} \quad \frac{\sum Y_{Ci}}{n} = \bar{Y}_C = \bar{Y}. \qquad \ldots. (1)$$

Similarly, it can be shown that $\bar{X}_C = \bar{X}$.

This implies that the mean of the estimated values is also equal to the mean of the observed values.

9.4 MEAN AND VARIANCE OF 'e' VALUES

(i) ***Mean of e_i values***

We know that $e_i = Y_i - Y_{Ci}$.

Taking sum over all the observations, we have

$$\sum e_i = \sum (Y_i - Y_{Ci}) = \sum Y_i - \sum Y_{Ci} = 0 \text{ [from equation (1)]}$$

∴ Mean of e_i values is equal to zero.

(ii) ***Variance of e_i values***

The variance of e_i values, in case of regression of Y on X, is given by

$$S^2_{Y.X} = \frac{1}{n}\sum (e_i - 0)^2 = \frac{1}{n}\sum (Y_i - Y_{Ci})^2 \quad \ldots. (2)$$

[Note that $\sum (Y_i - Y_{Ci})^2$ is the magnitude of unexplained variation in Y]

$$S^2_{Y.X} = \frac{1}{n}\sum \left[(Y_i - \bar{Y}) - b(X_i - \bar{X}) \right]^2$$

$$= \frac{\sum (Y_i - \bar{Y})^2}{n} + \frac{b^2 \sum (X_i - \bar{X})^2}{n} - \frac{2b \sum (X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

$$= \sigma_Y^2 + b^2\sigma_X^2 - 2b \cdot b\sigma_X^2 = \sigma_Y^2 - b^2\sigma_X^2$$

$$= \sigma_Y^2 - r^2\sigma_Y^2 = \sigma_Y^2 (1 - r^2)$$

Similarly, it can be shown that the mean of e'_i (= $X_i - X_{Ci}$) values, in case of regression of X on Y, is also equal to zero. Further, their variance, i.e.,

$$S^2_{X.Y} = \sigma_X^2 (1 - r^2)$$

Alternatively equation (2) can be written as,

$$S^2_{Y.X} = \frac{1}{n}\Sigma (Y_i - Y_{ci}) Y_i = \frac{1}{n}\left[\Sigma Y_i^2 - a\Sigma Y_i - b\Sigma X_i Y_i \right]$$

Similarly, we can write,

$$S^2_{X.Y} = \frac{1}{n}\left[\Sigma X_i^2 - c\Sigma X_i - d\Sigma X_i Y_i \right]$$

Remarks:

The above expressions for the variance are based on the following:

$$(Y_i - Y_{ci})^2 = \Sigma (Y_i - Y_{ci})(Y_i - Y_{ci})$$
$$= \Sigma (Y_i - Y_{ci}) Y_i - \Sigma (Y_i - Y_{ci}) Y_{ci}$$

It can be shown that the last term is zero.

$$\Sigma (Y_i - Y_{ci}) Y_{ci} = \Sigma [(Y_i - \bar{Y}) - b(X_i - \bar{X})][\bar{Y} + b(X_i - \bar{X})]$$
$$= \bar{Y}\Sigma (Y_i - \bar{Y}) - b\bar{Y}\Sigma (X_i - \bar{X}) + b\Sigma (X_i - \bar{X})(Y_i - \bar{Y}) - b^2\Sigma (X_i - \bar{X})^2$$
$$= 0 - 0 + b^2\Sigma (X_i - \bar{X})^2 - b^2\Sigma (X_i - \bar{X})^2 = 0$$

Method of Least Squares

This is one of the most popular methods of fitting a mathematical trend. The fitted trend is termed as the best in the sense that the sum of squares of deviations of observations, from it, is minimized. We shall use this method in the fitting of following trends:

1. Fitting of Linear Trend
2. Fitting of Parabolic Trend
3. Fitting of Exponential Trend

Standard Error of the Estimate

The standard error of the estimate of regression is given by the positive square root of the variance of e_i values.

The standard error of the estimate of regression of Y on X or simply the standard error of the estimate of Y is given as, $S_{Y.X} = \sigma_Y \sqrt{1-r^2}$.

Similarly, $S_{Y.X} = \sigma_Y \sqrt{1-r^2}$ is the standard error of the estimate X.

Remarks:

According to the theory of estimation, to be discussed in Chapter 21, an unbiased estimate of the variance of e_i values is given by

$$s^2_{Y.X} = \frac{\sum e_i^2}{n-2} = \frac{n}{n-2} \times \frac{\sum e_i^2}{n} = \frac{n}{n-2} \times \sigma_Y^2 \left(1-r^2\right)$$

$\therefore$ The standard errors of the estimate of Y and that of X are written as

$$s_{Y.X} = \sigma_Y \sqrt{\frac{n}{(n-2)}\left(1-r^2\right)} \text{ and } s_{X.Y} = \sigma_X \sqrt{\frac{n}{(n-2)}\left(1-r^2\right)} \text{ respectively.}$$

Note that difference between these standard errors tend to be equal to the standard errors for large values of n. In practice, the value of $n > 30$ may be treated as large.

Example 9.2: From the following data, compute (i) the coefficient of correlation between X and Y, (ii) the standard error of the estimate of Y:

$\sum x^2 = 24 \quad \sum y^2 = 42 \quad \sum xy = 30 \quad N = 10$, where $x = X - \bar{X}$ and $y = Y - \bar{Y}$.

Solution: The coefficient of correlation between X and Y is given by

$$r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}} = \frac{30}{\sqrt{24}\sqrt{42}} = 0.94$$

The standard error of the estimate of Y is given by ($n < 30$)

$$s_{Y.X} = \sqrt{\frac{\left(1-r^2\right)\sum y^2}{n-2}} = \sqrt{\frac{\left(1-0.94^2\right) \times 42}{8}} = 0.79$$

Example 9.3: For 100 items, it is given that the regression equations of Y on X and X on Y are $8X - 10Y + 66 = 0$ and $40X - 18Y = 214$ respectively. Compute the arithmetic means of X and Y and the coefficient of determination. If the standard deviation of X is given to be 3, compute the standard error of the estimate of Y.

Solution:

(a) ***The means of X and Y:*** Since the lines of regression pass through the point $(\bar{X}, \bar{Y})$, the simultaneous solution of the given regression equations would give the mean values of X and Y as $\bar{X} = 13, \bar{Y} = 17$.

(b) ***The coefficient of determination:*** We assume that $8X - 10Y + 66 = 0$ is the regression of Y on X and $40X - 18Y = 214$ is the regression of X on Y. Thus, the respective regression coefficients b and d are given by $\frac{8}{10}$ and $\frac{18}{40}$.

$\therefore$ The coefficient of determination $r^2 = b.d = \frac{8}{10} \times \frac{18}{40} = 0.36$

(c) ***The standard error of the estimate of Y:*** We know that $\sigma_{Y.X} = \sigma_Y \sqrt{1 - r^2}$.

To find s_y we use the relation $b = r \times \frac{\sigma_Y}{\sigma_X}$.

Also $r^2 = \frac{9}{25}$ $\therefore$ $r = \frac{3}{5}$ Thus, $\sigma_Y = \frac{b.\sigma_X}{r} = \frac{8}{10} \times \frac{5}{3} \times 3 = 4$

Hence, $\sigma_{Y.X} = 4\sqrt{1 - 0.36} = 3.2$

Check Your Progress 1

Fill in the blanks:

1. The general form of the line of regression of Y on X is .....
2. The term regression was first introduced by ..... in 1877

9.5 DEFINITION OF CORRELATION

Various experts have defined correlation in their own words and their definitions, broadly speaking, imply that correlation is the degree of association between two or more variables. Some important definitions of correlation are given below:

(i) "If two or more quantities vary in sympathy so that movements in one tend to be accompanied by corresponding movements in other(s) then they are said to be correlated."

–L.R. Connor

(ii) "Correlation is an analysis of covariation between two or more variables."

–A.M. Tuttle

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

–Croxton and Cowden

(iv) "Correlation analysis attempts to determine the 'degree of relationship' between variables ".

–Ya Lun Chou

Correlation Coefficient: It is a numerical measure of the degree of association between two or more variables.

Scatter Diagram

Let the bivariate data be denoted by (X_i, Y_i), where $i = 1, 2 \ldots\ldots n$. In order to have some idea about the extent of association between variables X and Y, each pair (X_i, Y_i), $i = 1, 2\ldots\ldots n$, is plotted on a graph. The diagram, thus obtained, is called a Scatter Diagram.

Each pair of values (X_i, Y_i) is denoted by a point on the graph. The set of such points (also known as dots of the diagram) may cluster around a straight line or a curve or may not show any tendency of association. Various possible situations are shown with the help of Figure 9.2.

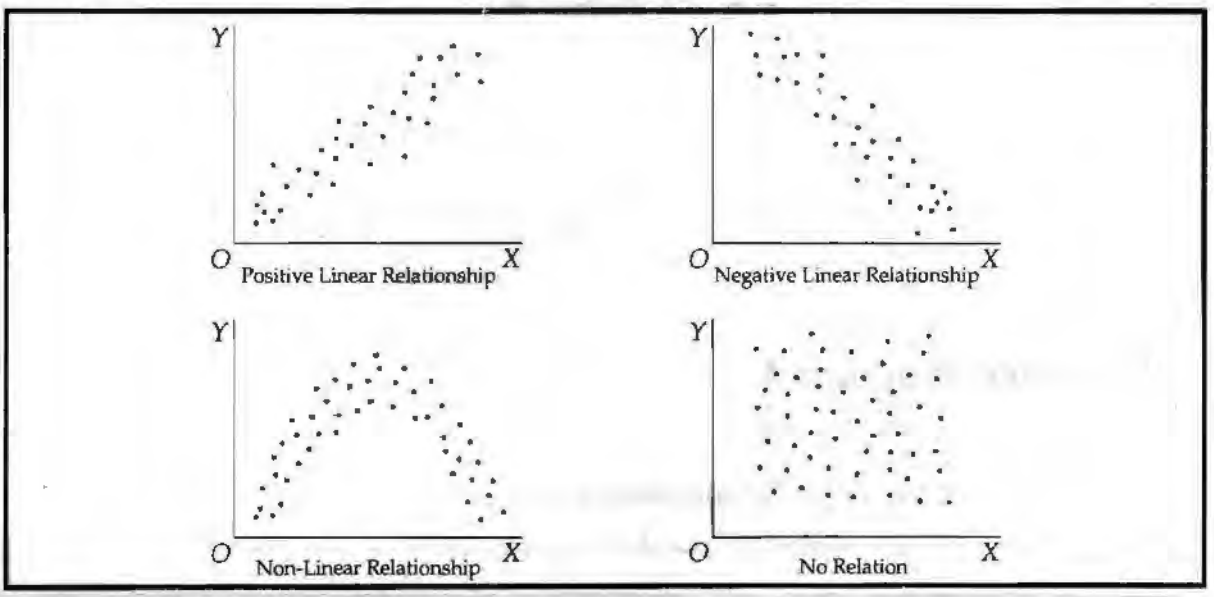


Figure 9.2

Scatter Diagram

If all the points or dots lie exactly on a straight line or a curve, the association between the variables is said to be perfect. This is shown in Figure 9.3.

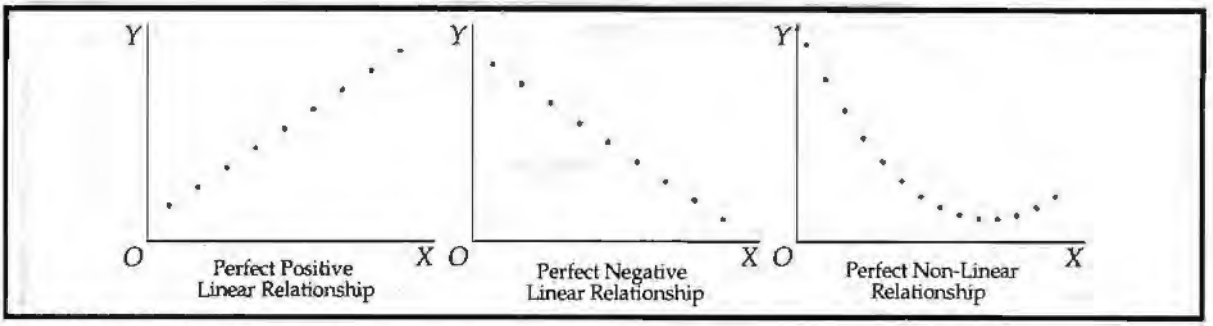


Figure 9.3

Scatter Diagram

A scatter diagram of the data helps in having a visual idea about the nature of association between two variables. If the points cluster along a straight line, the association between variables is linear. Further, if the points cluster along a curve, the corresponding association is non-linear or curvilinear. Finally, if the points neither cluster along a straight line nor along a curve, there is absence of any association between the variables.

It is also obvious from the above figure that when low (high) values of X are associated with low (high) value of Y, the association between them is said to be positive. Contrary to this, when low (high) values of X are associated with high (low) values of Y, the association between them is said to be negative.

This chapter deals only with linear association between the two variables X and Y. We shall measure the degree of linear association by the Karl Pearson's formula for the coefficient of linear correlation.

Correlation Analysis

The simplest way to find out qualitatively the correlation is to plot the data. In the case of our example, a strong *positive* correlation between y and x is evident, i.e., the plot reveals that as the weight increases, the fuel consumption increases as well. How can we quantify the degree of correlation? This is usually done by specifying the correlation coefficient R, defined as

$$= \frac{1}{n-1}\sum_{i=1}^{n}\frac{x_i-\mu_x}{\sigma_x}\frac{y_i-\mu_y}{\sigma_y} \tag{1}$$

where μ_x and σ_x denote the sample mean and the sample standard deviation respectively for the variable x and μ_y and σ_y denote the sample mean and the sample standard deviation respectively for the variable y.

Now, let's assume that a perfect linear relationship exists between the variables x and y. i.e., $y_i = ax_i + b$ for $i = 1, 2, \ldots\ldots n$ with $a \neq 0$. Now verify using the definitions of the mean and the variance that $= a + b$ and $= |a|$. This implies from Eq. 1 that $R = a/|a|$. Or in other words, $R = 1$ if $a > 0$ and $R = -1$ if $a < 0$. The case $R = 1$ corresponds to the maximum possible linear positive association between x and y, meaning that all the data points will lie exactly on a straight line of positive slope. Similarly, $R = -1$ corresponds to the maximum possible negative association between the statistical variables x and y. In general, $-1 \leq R \leq 1$ with the magnitude and the sign of R representing the *strength* and *direction* respectively of the association between the two variables. For the data given in Figure 9.4, $R = 0.977$ implying a strong positive correlation between the fuel consumption and the weight of the automobile.

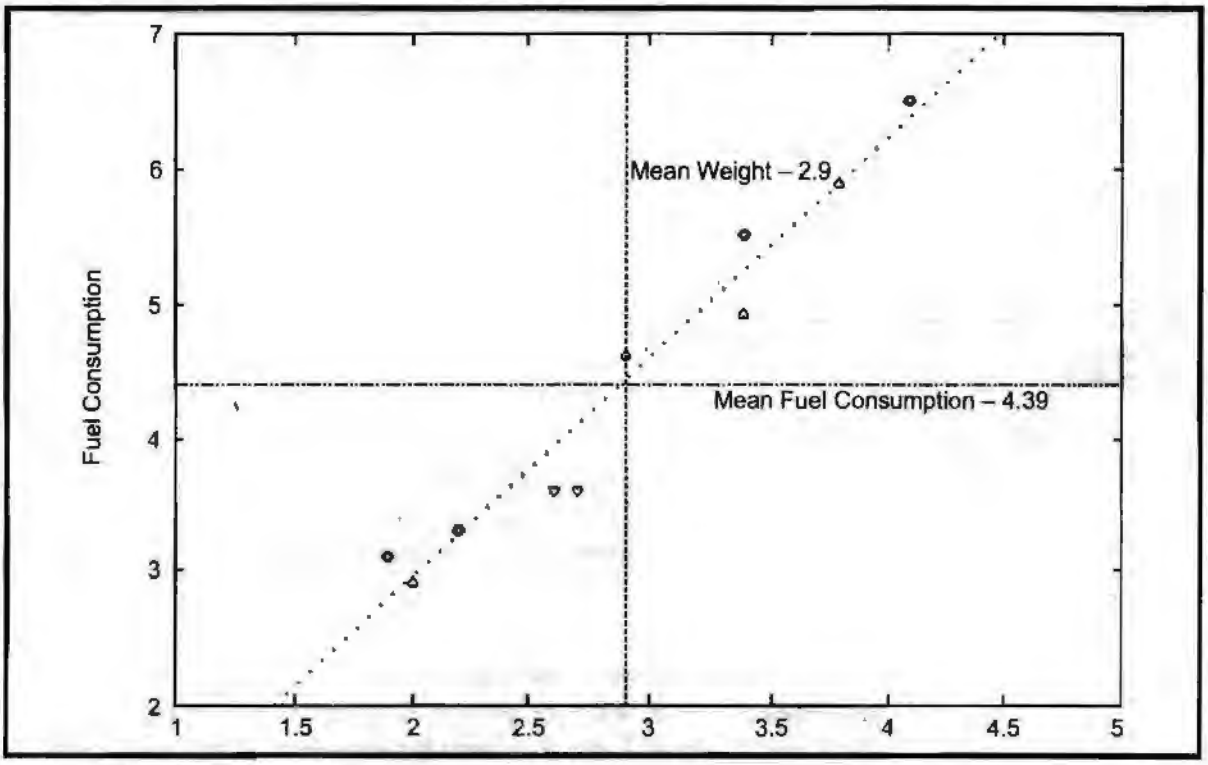


Figure 9.4

Fuel Consumption vs. Weight and the Best Fit Line

Prediction Intervals

The prediction interval estimates what future values will be, based upon present or past background samples taken. It tells tell the next data point sampled. As few as one future value can be estimated, and as few as four background values can be used to determine prediction limits (the minimum recommended in order to determine a standard deviation).

Assume that the data really are randomly sampled from a Gaussian distribution. Collect a sample of data and calculate a prediction interval. Then sample one more value from the population. If you do this many times, you'd expect that next value to lie within that prediction interval in 95% of the samples. The key point is that the prediction interval tells you about the distribution of values, not the uncertainty in determining the population mean. Prediction intervals must account for both the uncertainty in knowing the value of the population mean, plus data scatter. So a prediction interval is always wider than a confidence interval.

The prediction interval attempts to determine what future values will be with a degree of confidence, just as in the confidence and tolerance intervals. For example, we may attempt to predict that the next set of samples will fall within a determined range, with 99% confidence. To calculate prediction limits, we first must know a sample mean and standard deviation, based upon background data of sample size, *n*. Once we decide how many sampling periods and how many samples will be collected per sampling period, we can determine the prediction interval by using the same generic equation:

$$\bar{X} \pm K \cdot s$$

Where this time K is determined by the below equation. Recall, again, that we may be interested in a 1-tailed interval, which would simply have a "+" or a "–" in the above equation, and we would use 1-tailed values instead of 2-tailed values for *K*.

$$K = t_{1-\alpha/k,\,n-1}\sqrt{\frac{1}{n}+\frac{1}{m}}$$

for

k = number of sampling periods interested in

m = number of samples per sampling period (usually m = 1)

n = number of background samples

$1-\alpha/k$ = level of confidence (use $1-\alpha/2k$ for 2-tailed interval)

Let's define that word 'expect' used in defining a prediction interval. It means there is a 50% chance that you'd see the value within the interval in more than 95% of the samples, and a 50% chance that you'd see the value within the interval in less than 95% of the samples.

Making Inference about Population Parameter

Population parameter is the feature or characteristic of a population whose value you want to determine. It is the mean of some variable in a population, or the median, or the standard deviation, are all population parameters. Generally, their values ***define*** that population.

Parametric statistical inference may take the form of:

1. ***Estimation:*** on the basis of sample data we estimate the value of some parameter of the population from which the sample was randomly drawn.
2. ***Hypothesis Testing:*** We test the null hypothesis that a specified parameter (I shall use θ to stand for the parameter being estimated) of the population has a specified value.

One must know the **sampling distribution** of the estimator (the statistic used to estimate θ - I shall use to stand for the statistic used to estimate θ) to make full use of the estimator. The sampling distribution of a statistic is the distribution that would be obtained if you repeatedly drew samples of a specified size from a specified population and computed on each sample. In other words, it is the probability distribution of a statistic.

Coefficients of Determination and Correlation

We recall that in the line of regression $Y_C = a + bX$, X is used to estimate the value of Y. Further, the estimate of Y, independently of X, is given by a constant. Let this constant be A. Thus, we can write $Y_C = A$.

Given the observations $Y_1, Y_2, \ldots\ldots Y_n$, A will be the best estimate of Y if $S = \sum_{i=1}^{n}(Y_i - A)^2$ is minimum.

The necessary condition for minimum of S is $\frac{\partial S}{\partial A} = 0$.

i.e., $2\Sigma(Y_i - A) = 0$ or $\Sigma Y_i - nA = 0$ or $A = \bar{Y}$.

∴ The best estimate (an estimate having minimum sum of squares of errors) of Y, independently of X, is given by $Y_C = \bar{Y}$.

Remarks: If X and Y are independent variables, the two lines of regression are $Y_C = \bar{Y}$ and $X_C = \bar{X}$.

Very often, when we use X for the estimation of Y, we are interested in knowing how far the use of X enables us to explain the variations in Y values from $\bar{Y}$ or, in other words, how much of the variations in Y, from $\bar{Y}$, are being explained by the regression equation $Y_{Ci} = a + bX_i$? To answer this question, we write

$$Y_i - \bar{Y} = Y_i - Y_{Ci} + Y_{Ci} - \bar{Y} \quad \text{(Subtracting and adding } Y_{Ci})$$

$$\text{or} \quad Y_i - \bar{Y} = (Y_i - Y_{Ci}) + (Y_{Ci} - \bar{Y})$$

Squaring both sides and taking sum over all the observations, we have

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - Y_{Ci})^2 + \sum(Y_{Ci} - \bar{Y})^2 + 2\sum(Y_i - Y_{Ci})(Y_{Ci} - \bar{Y}) \qquad \ldots(1)$$

Consider the product term

$$2\sum(Y_i - Y_{Ci})(Y_{Ci} - \bar{Y}) = 2\sum\left[\{Y_i - \bar{Y} - b(X_i - \bar{X})\}\{b(X_i - \bar{X})\}\right]$$

$$= 2b\sum(Y_i - \bar{Y})(X_i - \bar{X}) - 2b^2\sum(X_i - \bar{X})^2$$

$$= 2b^2\sum(X_i - \bar{X})^2 - 2b^2\sum(X_i - \bar{X})^2 = 0$$

Thus, equation (1) becomes

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - Y_{Ci})^2 + \sum(Y_{Ci} - \bar{Y})^2 \qquad \ldots. (2)$$

From the above figure, we note that $Y_{Ci} - \bar{Y}$ is the deviation of the estimated value from $\bar{Y}$. This deviation has occurred because X and Y are related by the regression equation $Y_{Ci} = a + bX_i$, so that the estimate of Y is Y_{Ci} when $X = X_i$. Similar type of deviations would occur for other values of X. Thus, the magnitude of the term $\sum(Y_{Ci} - \bar{Y})^2$ gives the strength of the relationship, $Y_{Ci} = a + bX_i$, between X and Y or, equivalently, the variations in Y that are explained by the regression equation.

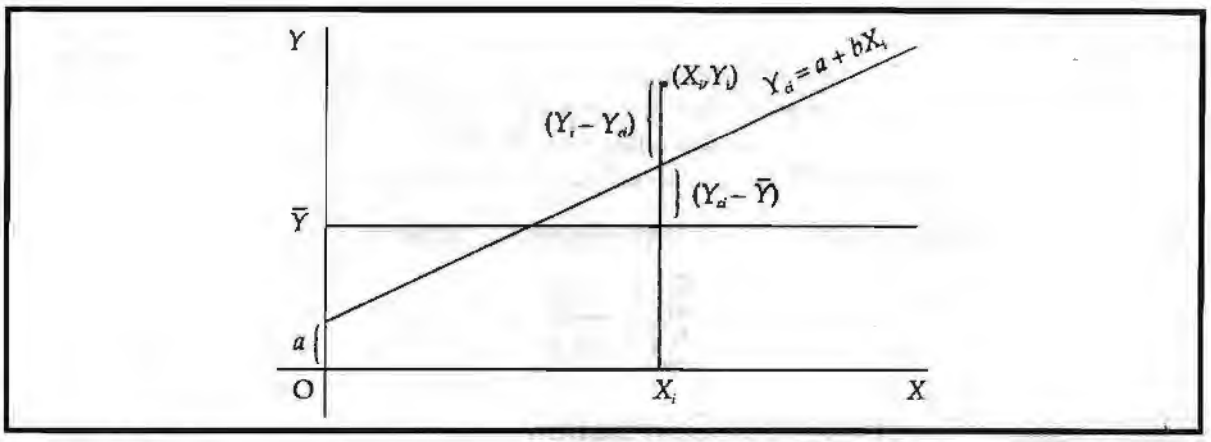


Figure 9.5

Lines of regression

The other term $Y_i - Y_{Ci}$ gives the deviation of i th observed value from the regression line and thus the magnitude of the term $\sum(Y_i - Y_{Ci})^2$ gives the variations in Y about the line of regression. These variations are also known as unexplained variations in Y.

Adding the two types of variations, we get the magnitude of total variations in Y. Thus, equation (2) can also be written as

Total variations in Y = Unexplained variations in Y + Explained variations in Y.

Dividing both sides of equation (2) by $\sum(Y_i - \bar{Y})^2$, we have

$$1 = \frac{\sum(Y_i - Y_{Ci})^2}{\sum(Y_i - \bar{Y})^2} + \frac{\sum(Y_{Ci} - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \quad \text{.... (3)}$$

or 1 = Proportion of unexplained variations + Proportion of variations explained by the regression equation.

The proportion of variation explained by regression equation is called the coefficient of determination.

Thus, the coefficient of determination $= \frac{\sum(Y_{Ci} - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$

$$= \frac{b^2\sum(X_i - \bar{X})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\left[\sum(X_i - \bar{X})(Y_i - \bar{Y})\right]^2}{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2} = r^2$$

This result shows that the coefficient of determination is equal to the square of the coefficient of correlation, i.e., r^2 gives the proportion of variations explained by each regression equation.

Remarks:

(i) It should be obvious from the above that it is desirable to calculate the coefficient of correlation prior to the fitting of a regression line. If r^2 is high enough, the fitted line will explain a greater proportion of the variations in the dependent variable. A low value of r^2 would, however, indicate that the proposed fitting of regression would not be of much use.

(ii) The expression for the coefficient of determination for regression of X on Y can be written in a similar way. Here we can write $r^2 = \frac{\sum(X_C - \bar{X})^2}{\sum(X_i - \bar{X})^2}$.

Limitations of Coefficient of Correlation

This measure, however, suffers from certain limitations, given below:

1. Coefficient of correlation r does not give any idea about the existence of cause and effect relationship between the variables. It is possible that a high value of r is obtained although none of them seem to be directly affecting the other. Hence, any interpretation of r should be done very carefully.
2. It is only a measure of the degree of linear relationship between two variables. If the relationship is not linear, the calculation of r does not have any meaning.
3. Its value is unduly affected by extreme items.
4. If the data are not uniformly spread in the relevant quadrants, the value of r may give a misleading interpretation of the degree of relationship between the two variables. For example, if there are some values having concentration around a point in first quadrant and there is similar type of concentration in third quadrant, the value of r will be very high although there may be no linear relation between the variables.
5. As compared with other methods, to be discussed later in this lesson, the computations of r are cumbersome and time consuming.

Probable Error of r

It is an old measure to test the significance of a particular value of r without the knowledge of test of hypothesis. Probable error of r, denoted by P.E.(r) is 0.6745 times its standard error. The value 0.6745 is obtained from the fact that in a normal distribution $\bar{r} \pm 0.6745 \times S.E.$ covers 50% of the total distribution.

According to Horace Secrist *"The probable error of correlation coefficient is an amount which if added to and subtracted from the mean correlation coefficient, gives limits within which the chances are even that a coefficient of correlation from a series selected at random will fall."*

Since standard error of r, i.e., $S.E._r = \frac{1-r^2}{\sqrt{n}}$, $\therefore P.E.(r) = 0.6745 \times \frac{1-r^2}{\sqrt{n}}$

Uses of P.E.(r)

(i) It can be used to specify the limits of population correlation coefficient r (rho) which are defined as $r - \text{P.E.}(r) \le r \le r + \text{P.E.}(r)$, where r denotes correlation coefficient in population and r denotes correlation coefficient in sample.

(ii) It can be used to test the significance of an observed value of r without the knowledge of test of hypothesis. By convention, the rules are:

(a) If $|r| < 6$ P.E.(r), then correlation is not significant and this may be treated as a situation of no correlation between the two variables.

(b) If $|r| > 6$ P.E.(r), then correlation is significant and this implies presence of a strong correlation between the two variables.

(c) If correlation coefficient is greater than 0.3 and probable error is relatively small, the correlation coefficient should be considered as significant.

Example 9.4: Find out correlation between age and playing habit from the following information and also its probable error.

Age	:	15	16	17	18	19	20
No. of Students	:	250	200	150	120	100	80
Regular Players	:	200	150	90	48	30	12

Solution: Let X denote age, p the number of regular players and q the number of students. Playing habit, denoted by Y, is measured as a percentage of regular players in an age group, i.e., $Y = (p/q)\times 100$.

Table for calculation of *r*

X	q	p	Y	$u = X - 17$	$v = Y - 40$	uv	u^2	v^2
15	250	200	80	−2	40	−80	4	1600
16	200	150	75	−1	35	−35	1	1225
17	150	90	60	0	20	0	0	400
18	120	48	40	1	0	0	1	0
19	100	30	30	2	−10	−20	4	100
20	80	12	15	3	−25	−75	9	625
Total				3	60	−210	19	3950

$$r_{XY} = \frac{-6\times 210 - 3\times 60}{\sqrt{6\times 19 - 9}\sqrt{6\times 3950 - 3600}} = -0.99$$

Probable error of r, i.e., $P.E.(r) = 0.6745 \times \dfrac{\left[1-(0.99)^2\right]}{\sqrt{6}} = 0.0055$

Example 9.5: Test the significance of correlation for the values based on the number of observations (i) 10, and (ii) 100 and r = 0.4 and 0.9.

Solution:

(i) (a) Consider $n = 10$ and $r = 0.4$. Thus, $P.E.(r) = 0.6745 \times \frac{1-0.4^2}{\sqrt{10}} = 0.179$ and

6 *P.E.* = 6 × 0.179 = 1.074. Since $|r| < 6$ *P.E.*, r is not significant.

(b) Take $n = 10$ and $r = 0.9$. Thus, $P.E. = 0.6745 \times \frac{1-0.9^2}{\sqrt{10}} = 0.041$ and 6 *P.E.* = 6 × 0.041 = 0.246. Since $|r| > 6$ *P.E.*, r is highly significant.

(ii) (a) Take $n = 100$ and $r = 0.4$. Thus, $6P.E. = 6 \times 0.6745 \frac{(1-0.4^2)}{\sqrt{100}} = 0.34$

Since $|r| > 6$ *P.E.*, r is significant.

(b) Take $n = 100$ and $r = 0.9$. Thus, $6P.E. = 6 \times 0.6745 \frac{(1-0.9^2)}{\sqrt{100}} = 0.077$

Since $|r| > 6$ *P.E.*, r is significant.

9.6 REGRESSION AND CORRELATION ANALYSIS

As in case of calculation of correlation coefficient, we can directly write the formula for the two regression coefficients for a bivariate frequency distribution as given below:

$$b = \frac{N\sum\sum f_{ij}X_iY_j - \left(\sum f_iX_i\right)\left(\sum f_j'Y_j\right)}{N\sum f_iX_i^2 - \left(\sum f_iX_i\right)^2}$$

or, if we define $u_i = \frac{X_i - A}{h}$ and $v_j = \frac{Y_j - B}{k}$,

$$b = \frac{k}{h}\left[\frac{N\sum\sum f_{ij}u_iv_j - \left(\sum f_iu_i\right)\left(\sum f_j'v_j\right)}{N\sum f_iu_i^2 - \left(\sum f_iu_i\right)^2}\right]$$

Similarly, $$d = \frac{N\sum\sum f_{ij}X_iY_j - \left(\sum f_iX_i\right)\left(\sum f_j'Y_j\right)}{N\sum f_j'Y_j^2 - \left(\sum f_j'Y_j\right)^2}$$

or $$d = \frac{h}{k}\left[\frac{N\sum f_{ij}u_iv_j - \left(\sum f_iu_i\right)\left(\sum f_j'v_j\right)}{N\sum f_j'v_j^2 - \left(\sum f_j'v_j\right)^2}\right]$$

Example 9.6: By calculating the two regression coefficients obtain the two regression lines from the following data:

$Y \rightarrow$ / $X \downarrow$	0 – 5	5 – 10	10 – 15
0 – 10	2	5	7
10 – 20	1	3	2
20 – 30	8	4	0

Solution: The mid points of X-values are 5, 15, 25.

Let $u = \dfrac{X-15}{10}$, $\therefore$ Corresponding u-values become – 1, 0, 1

Similarly, the mid-points of Y-values are 2.5, 7.5, 12.5

Let $v = \dfrac{Y-7.5}{5}$, $\therefore$ Corresponding v-values become – 1, 0, 1

Calculation Table

u \ v	–1	0	1	f_i	$f_i u_i$	$f_i u_i^2$	$f_{ij} u_i v_j$
–1	2 [2]	5 [0]	7 [–7]	14	–14	14	–5
0	1 [0]	3 [0]	2 [0]	6	0	0	0
1	8 [–8]	4 [0]	0 [0]	12	12	12	–8
f_j'	11	12	9	32	–2	26	–13
$f_j v_j'$	–11	0	9	–2			
$f_j v_j^2$	11	0	9	20			

From the table $N = 32$ (total frequency)

(a) ***Regression of Y on X***

Regression Coefficient (here $h = 10$ and $k = 5$)

$$b = \left[\frac{-32 \times 13 - 2 \times 2}{32 \times 26 - 4}\right] \times \frac{5}{10} = \frac{-416-4}{832-4} \times \frac{1}{2} = -0.25$$

Also, $\bar{X} = 15 + \dfrac{10(-2)}{32} = 14.73$ and $\bar{Y} = 7.5 + \dfrac{5(-2)}{32} = 7.19$

$\therefore$ $a = \bar{Y} - b\bar{X} = 7.19 + 0.25 \times 14.73 = 10.87$

Hence, the regression of Y on X becomes $Y_c = 10.87 - 0.25X$

(b) ***Regression of X on Y***

Regression coefficient $d = \left[\dfrac{-420}{32 \times 20 - 4}\right] \times \dfrac{10}{5} = -1.32$

Also, $c = \bar{X} - d\bar{Y}$ =14.73 + 1.32 × 7.19 = 24.22

Hence, the regression of X on Y becomes X_c = 24.22 – 1.32Y

Check Your Progress 2

Fill in the blanks:

1. .....is an analysis of co-variation between two or more variables.
2. .....r does not give any idea about the existence of cause and effect relationship between the variables.

9.7 SUMMARY

The regression equations are useful for predicting the value of dependent variable for given value of the independent variable. As pointed out earlier, the nature of a regression equation is different from the nature of a mathematical equation, e.g., if $Y = 10 + 2X$ is a mathematical equation then it implies that Y is exactly equal to 20 when $X = 5$.

However, if $Y = 10 + 2X$ is a regression equation, then $Y = 20$ is an average value of Y when $X = 5$.

The term 'Regression', originated in this particular context, is now used in various fields of study, even though there may be no existence of any regressive tendency.

For a bivariate data (X_i, Y_i), $i = 1, 2, \ldots\ldots n$, we can have either X or Y as independent variable. If X is independent variable then we can estimate the average values of Y for a given value of X. When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation.

So far we have considered distributions relating to single characteristics. Such distributions are known as Univariate Distribution. When various units under consideration are observed simultaneously, with regard to two characteristics, we get a Bivariate Distribution. For example, the simultaneous study of the heights and weights of students of a college. For such data also, we can compute mean, variance, skewness etc. for each individual characteristics. In addition to this, in the study of a bivariate distribution, we are also interested in knowing whether there exists some relationship between two characteristics or in other words, how far the two variables, corresponding to two characteristics, tend to move together in same or opposite directions i.e. how far they are associated.

9.8 KEYWORDS

- Correlation
- Standard Error
- Coefficient of determination
- Spearman's Rank correlation
- Covariance
- Regression analysis

9.10 REVIEW QUESTIONS

1. Distinguish between correlation and regression. Discuss least square method of fitting regression.

2. What do you understand by linear regression? Why there are two lines of regression? Under what condition(s) can there be only one line?

3. Write a note on the standard error of the estimate.

4. The regression line gives only a 'best estimate' of the quantity in question. We may assess the degree of uncertainty in this estimate by calculating its standard error. Explain.

5. Given a scatter diagram of a bivariate data involving two variables X and Y. Find the conditions of minimisation of and hence derive the normal equations for the linear regression of Y on X. What sum is to be minimised when X is regressed on Y? Write down the normal equation in this case.

6. What is the method of least squares? Show that the two lines of regression obtained by this method are irreversible except when $r = \pm 1$. Explain.

7. (a) Define correlation between two variables.

   (b) Define the concept of covariance. How do you interpret it?

8. Define correlation and discuss its significance in statistical analysis. Does it signify 'cause and effect' relationship between the two variables?

9. (a) What do you understand by the coefficient of linear correlation? Explain the significance and limitations of this measure in any statistical analysis.

   (b) Write down an expression for the Karl Pearson's coefficient of linear correlation. Why is it termed as the coefficient of linear correlation? Explain.

10. (a) Describe the method of obtaining the Karl Pearson's formula of coefficient of linear correlation. What do positive and negative values of this coefficient indicate?

    (b) Does a zero value of Karl Pearson's coefficient of correlation between two variables X and Y imply that X and Y are not related? Explain.

11. Define product moment coefficient of correlation. What are the advantages of the study of correlation?

12. Show that the coefficient of correlation, r, is independent of change of origin and scale.

13. Prove that the coefficient of correlation lies between -1 and $+1$.

14. "If two variables are independent the correlation between them is zero, but the converse is not always true". Explain the meaning of this statement.

15. Calculate the coefficient of correlation by Karl Pearson's method for the following data relating to the money supply (in crores of Rs) and deposit money with the public (in crores of Rs):

Year	Money Supply	Deposit Money	Year	Money Supply	Deposit Money
1961	29	8	1966	46	15
1962	30	9	1967	50	18
1963	33	9	1968	54	20
1964	38	12	1969	58	21
1965	41	14	1970	70	25

What conclusions do you draw?

Answers to Check Your Progress

Check Your Progress 1

1. $YC_i = a + bX_i$
2. Sir Francis Galton

Check Your Progress 2

1. Correlation
2. Coefficient of Correlation

9.10 REFERENCES AND FURTHER READING

- Pallant, J. (2020). SPSS survival manual (7th ed.). Open University Press.
- Vogt, W. P., & Johnson, R. (2021). Dictionary of statistics & methodology: A nontechnical guide for the social sciences (4th ed.). Sage Publications.
- Gravetter, F. J., & Wallnau, L. B. (2022). Statistics for The Behavioral Sciences (10th ed.). Cengage Learning.
- Laerd Statistics. (2023). Statistical analyses using SPSS. Retrieved from https://statistics.laerd.com
- UCLA Statistical Consulting Group. (2024). Statistical methods for the social sciences. Retrieved from https://stats.oarc.ucla.edu

BLOCK – IV

UNIT 10

Time Series and Forecasting

CHAPTER OUTLINE

10.1 INTRODUCTION

A series of observations, on a variable, recorded after successive intervals of time is called a time series. The successive intervals are usually equal time intervals, e.g., it can be 10 years, a year, a quarter, a month, a week, a day, an hour, etc. The data on the population of India is a time series data where time interval between two successive figures is 10 years. Similarly figures of national income, agricultural and industrial production, etc., are available on yearly basis.

It should be noted here that the time series data are bivariate data in which one of the variables is time. This variable will be denoted by t. The symbol Y_t will be used to denote the observed value, at point of time t, of the other variable. If the data pertains to n periods, it can be written as (t, Y_t), $t = 1, 2, \ldots. n$.

10.2 VARIATIONS IN TIME SERIES

An observed value of a time series, Y_t, is the net effect of many types of influences such as changes in population, techniques of production, seasons, level of business activity, tastes and habits, incidence of fire floods, etc. It may be noted here that different types of variables may be affected by different types of factors, e.g., factors affecting the agricultural output may be entirely different from the factors affecting industrial output. However, for the purpose of time series analysis, various factors are classified into the following three general categories applicable to any type of variable.

1. Secular Trend or simply Trend
2. Periodic or Oscillatory Variations
   (i) Seasonal Variations
   (ii) Cyclical Variations
3. Random or Irregular Variations

Secular Trend

Secular trend or simply trend is the general tendency of the data to increase or decrease or stagnate over a long period of time. Most of the business and economic time series would reveal a tendency to increase or to decrease over a number of years. For example, data regarding industrial production, agricultural production, population, bank deposits, deficit financing, etc., show that, in general, these magnitudes have been rising over a fairly long period. As opposed to this, a time series may also reveal a declining trend, e.g., in the case of substitution of one commodity by another, the demand of the substituted commodity would reveal a declining trend such as the demand for cotton clothes, demand for coarse grains like bajra, jowar, etc. With the improved medical facilities, the death rate is likely to show a declining trend, etc. The change in trend, in either case, is attributable to the fundamental forces such as changes in population, technology, composition of production, etc.

According to A.E. Waugh, secular trend is, "that irreversible movement which continues, in general, in the same direction for a considerable period of time". There are two parts of this definition; (i) movement in same direction, which implies that if the values are increasing (or decreasing) in successive periods, the tendency continues; and (ii) a considerable period of time. There is no specific period which can be called as a long period. Long periods are different for different situations. For example, in

cases of population or output trends, the long period could be 10 years while it could be a month for the daily demand trend of vegetables. It should, however, be noted that longer is the period the more significant would be the trend. Further, it is not necessary that the increase or decrease of values must continue in the same direction for the entire period. The data may first show a rising (or falling) trend and subsequently a falling (or rising) trend.

Periodic Variations

These variations, also known as oscillatory movements, repeat themselves after a regular interval of time. This time interval is known as the period of oscillation. These oscillations are shown in Figure 10.1:

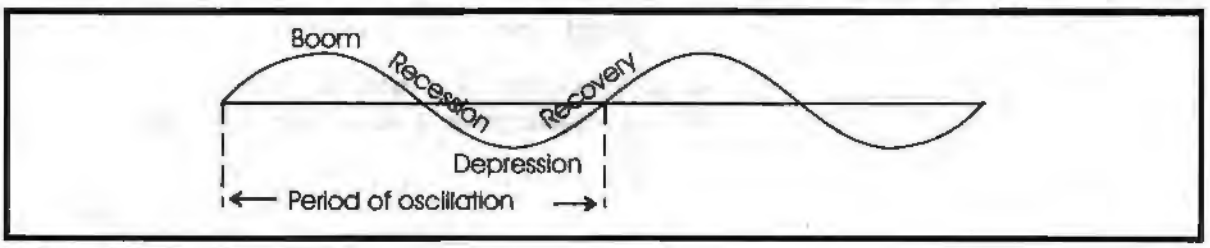


Figure 10.1

Periodic Variations

The oscillatory movements are termed as Seasonal Variations if their period of oscillation is equal to one year, and as Cyclical Variations if the period is greater than one year.

A time series, where the time interval between successive observations is less than or equal to one year, may have the effects of both the seasonal and cyclical variations. However, the seasonal variations are absent if the time interval between successive observations is greater than one year.

Although the periodic variations are more or less regular, they may not necessarily be uniformly periodic, i.e., the pattern of their variations in different periods may or may not be identical in respect of time period and size of periodic variations. For example, if a cycle is completed in five years then its following cycle may take greater or less than five years for its completion.

Causes of Seasonal Variations

The main causes of seasonal variations are:

1. Climatic Conditions
2. Customs and Traditions

1. ***Climatic Conditions:*** The changes in climatic conditions affect the value of time series variable and the resulting changes are known as seasonal variations. For example, the sale of woolen garments is generally at its peak in the month of November because of the beginning of winter season. Similarly, timely rainfall may increase agricultural output, prices of agricultural commodities are lowest during their harvesting season, etc., reflect the effect of climatic conditions on the value of time series variable.
2. ***Customs and Traditions:*** The customs and traditions of the people also give rise to the seasonal variations in time series. For example, the sale of garments and ornaments may be highest during the marriage season, sale of sweets during Diwali, etc., are variations that are the results of customs and traditions of the people.

It should be noted here that both of the causes, mentioned above, occur regularly and are often repeated after a gap of less than or equal to one year.

Objectives of Measuring Seasonal Variations

The main objectives of measuring seasonal variations are:

1. To analyse the past seasonal variations.
2. To predict the value of a seasonal variation which could be helpful in short-term planning?
3. To eliminate the effect of seasonal variations from the data.

Methods of Measuring Seasonal Variations

The measurement of seasonal variation is done by isolating them from other components of a time series. There are four methods commonly used for the measurement of seasonal variations. These methods are:

1. Method of Simple Averages
2. Ratio to Trend Method
3. Ratio to Moving Average Method
4. Method of Link Relatives

Method of Simple Averages

This method is used when the time series variable consists of only the seasonal and random components. The effect of taking average of data corresponding to the same period (say 1st quarter of each year) is to eliminate the effect of random component and thus, the resulting averages consist of only seasonal component. These averages are then converted into seasonal indices.

Example 10.1: Assuming that trend and cyclical variations are absent, compute the seasonal index for each month of the following data of sales (in Rs '000) of a company.

Year	*Jan*	*Feb*	*Mar*	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	*Dec*
1987	46	45	44	46	45	47	46	43	40	40	41	45
1988	45	44	43	46	46	45	47	42	43	42	43	44
1989	42	41	40	44	45	45	46	43	41	40	42	45

Solution:

Calculation Table

Year	*Jan*	*Feb*	*Mar*	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	*Dec*
1987	46	45	44	46	45	47	46	43	40	40	41	45
1988	45	44	43	46	46	45	47	42	43	42	43	44
1989	42	41	40	44	45	45	46	43	41	40	42	45
Total	133	130	127	136	136	137	139	128	124	122	126	134
A_i	44.3	43.3	42.3	45.3	45.3	45.7	46.3	42.7	41.3	40.7	42.0	44.7
S.I.	101.4	99.1	96.8	103.7	103.7	104.6	105.9	97.7	94.5	93.1	96.1	102.3

In the above table, A_i denotes the average and S.I. the seasonal index for a particular month of various years. To calculate the seasonal index, we compute grand average G, given by $G = \frac{\Sigma A_i}{12} = \frac{524}{12} = 43.7$. Then the seasonal index for a particular month is given by $S.I. = \frac{A_i}{G} \times 100$.

Further, ΣS.I. = 1198.9 ≠ 1200. Thus, we have to adjust these values such that their total is 1200. This can be done by multiplying each figure by $\frac{1200}{1198.9}$. The resulting figures are the adjusted seasonal indices, as given below:

Jan	*Feb*	*Mar*	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	*Dec*
101.5	99.2	96.9	103.8	103.8	104.7	106.0	97.8	94.6	93.2	96.2	102.3

Remarks: The totals equal to 1200, in case of monthly indices and 400, in case of quarterly indices, indicate that the ups and downs in the time series, due to seasons, neutralise themselves within that year. It is because of this that the annual data are free from seasonal component.

Example 10.2: Compute the seasonal index from the following data by the method of simple averages.

Year	*Quarter*	*Y*	*Year*	*Quarter*	*Y*	*Year*	*Quarter*	*Y*
1980	*I*	106	1982	*I*	90	1984	*I*	80
	II	124		*II*	112		*II*	104
	III	104		*III*	101		*III*	95
	IV	90		*IV*	85		*IV*	83
1981	*I*	84	1983	*I*	76	1985	*I*	104
	II	114		*II*	94		*II*	112
	III	107		*III*	91		*III*	102
	IV	88		*IV*	76		*IV*	84

Solution:

Calculation of Seasonal Indices

Years	*1st Qr*	*2nd Qr*	*3rd Qr*	*4th Qr*
1980	106	124	104	90
1981	84	114	107	88
1982	90	112	101	85
1983	76	94	91	76
1984	80	104	95	83
1985	104	112	102	84
Total	540	660	600	506
A_i	90	110	100	84.33
$\frac{A_i}{G} \times 100$	93.67	114.49	104.07	87.77

We have $G = \frac{\sum A_i}{4} = \frac{384.33}{4} = 96.08$. Further, since the sum of terms in the last row of the table is 400, no adjustment is needed. These terms are the seasonal indices of respective quarters.

Merits and Demerits

This is a simple method of measuring seasonal variations which is based on the unrealistic assumption that the trend and cyclical variations are absent from the data. However, we shall see later that this method, being a part of the other methods of measuring seasonal variations, is very useful.

Ratio to Trend Method

This method is used when cyclical variations are absent from the data, i.e., the time series variable Y consists of trend, seasonal and random components.

Using symbols, we can write $Y = T.S.R$

Various steps in the computation of seasonal indices are:

1. Obtain the trend values for each month or quarter, etc., by the method of least squares.
2. Divide the original values by the corresponding trend values. This would eliminate trend values from the data. To get figures in percentages, the quotients are multiplied by 100.

   Thus, we have

$$\frac{Y}{T} \times 100 = \frac{T.S.R}{T} \times 100 = S.R.100$$

3. Finally, the random component is eliminated by the method of simple averages.

Example 10.3: Assuming that the trend is linear, calculate seasonal indices by the ratio to moving average method from the following data:

Quarterly Output of Coal in 4 Years (in thousand tonnes)

Year	*I*	*II*	*III*	*IV*
1982	65	58	56	61
1983	68	63	63	67
1984	70	59	56	52
1985	60	55	51	58

Solution: By adding the values of all the quarters of a year, we can obtain annual output for each of the four years. Fit a linear trend to the data and obtain trend values for each quarter.

Year	*Output*	$X = 2(t - 1983.5)$	XY	X^2
1982	240	−3	−720	9
1983	261	−1	−261	1
1984	237	1	237	1
1985	224	3	672	9
Total	962	0	−72	20

From the above table, we get $a = \frac{962}{4} = 240.5$ and $b = \frac{-72}{20} = -3.6$

Thus, the trend line is $Y = 240.5 - 3.6X$, Origin : 1st January 1984, unit of X : 6 months.

The quarterly trend equation is given by

$Y = \frac{240.5}{4} - \frac{3.6}{8}X$ or $Y = 60.13 - 0.45X$, Origin : 1st January 1984, unit of X : 1 quarter (i.e., 3 months).

Shifting origin to 15th Feb. 1984, we get

$Y = 60.13 - 0.45(X + \frac{1}{2}) = 510.9 - 0.45X$, origin I-quarter, unit of X = 1 quarter.

The table of quarterly values is given by

Year	*I*	*II*	*III*	*IV*
1982	63.50	63.05	62.60	62.15
1983	61.70	61.25	60.80	60.35
1984	59.90	59.45	59.00	58.55
1985	58.10	57.65	57.20	56.75

The table of Ratio to Trend Values, i.e., $\frac{Y}{T} \times 100$

Years	*I*	*II*	*III*	*IV*
1982	102.36	91.99	89.46	98.15
1983	110.21	102.86	103.62	111.02
1984	116.86	99.24	94.92	88.81
1985	103.27	95.40	89.16	102.20
Total	432.70	389.49	377.16	400.18
Average	108.18	97.37	94.29	100.05
S.I.	108.20	97.40	94.32	100.08

Note: Grand Average, $G = \frac{399.89}{4} = 99.97$

Example 10.4: Find seasonal variations by the ratio to trend method, from the following data:

Year	*I-Qr*	*II-Qr*	*III-Qr*	*IV-Qr*
1975	30	40	36	34
1976	34	52	50	44
1977	40	58	54	48
1978	54	76	68	62
1979	80	92	86	82

Solution: First we fit a linear trend to the annual totals.

Years	*Annual Totals* (Y)	X	XY	X^2
1975	140	−2	−280	4
1976	180	−1	−180	1
1977	200	0	0	0
1978	260	1	260	1
1979	340	2	680	4
Total	1120	0	480	10

Now $a = \frac{1120}{5} = 224$ and $b = \frac{480}{10} = 48$.

∴ The trend equation is $Y = 224 + 48X$, origin : 1st July 1977, unit of X = 1 year.

The quarterly trend equation is $Y = \frac{224}{4} + \frac{48}{16}X = 56 + 3X$, origin : 1st July 1977, unit of X = 1 quarter.

Shifting the origin to III quarter of 1977, we get

$$Y = 56 + 3(X + \frac{1}{2}) = 57.5 + 3X$$

Table of Quarterly Trend Values

Year	*I*	*II*	*III*	*IV*
1975	27.5	30.5	33.5	36.5
1976	39.5	42.5	45.5	48.5
1977	51.5	54.5	57.5	60.5
1978	63.5	66.5	69.5	72.5
1979	75.5	78.5	81.5	84.5

Ratio to Trend Values

Year	*I*	*II*	*III*	*IV*
1975	109.1	131.1	107.5	93.2
1976	86.1	122.4	109.9	90.7
1977	77.7	106.4	93.9	79.3
1978	85.0	114.3	97.8	85.5
1979	106.0	117.2	105.5	97.0
Total	463.9	591.4	514.6	445.7
A_i	92.78	118.28	102.92	89.14
S.I.	92.10	117.35	102.11	88.44

Note that the Grand Average $G = \frac{403.12}{4} = 100.78$. Also check that the sum of indices is 400.

Remarks: If instead of multiplicative model we have an additive model, then $Y = T + S + R$ or $S + R = Y - T$. Thus, the trend values are to be subtracted from the Y values. Random component is then eliminated by the method of simple averages.

Merits and Demerits

It is an objective method of measuring seasonal variations. However, it is very complicated and doesn't work if cyclical variations are present.

Ratio to Moving Average Method

The ratio to moving average is the most commonly used method of measuring seasonal variations. This method assumes the presence of all the four components of a time series. Various steps in the computation of seasonal indices are as follows:

(i) Compute the moving averages with period equal to the period of seasonal variations. This would eliminate the seasonal component and minimize the effect of random component. The resulting moving averages would consist of trend, cyclical and random components.

(ii) The original values, for each quarter (or month) are divided by the respective moving average figures and the ratio is expressed as a percentage, i.e.,

$$\frac{Y}{M.A.} = \frac{TCSR}{TCR'} = SR''$$

where R' and R'' denote the changed random components.

(iii) Finally, the random component R'' is eliminated by the method of simple averages.

Example 10.5: Given the following quarterly sale figures, in thousand of rupees, for the year 1986-1989, find the specific seasonal indices by the method of moving averages.

Year	*I*	*II*	*III*	*IV*
1986	34	33	34	37
1987	37	35	37	39
1988	39	37	38	40
1989	42	41	42	44

Solution: Calculation of Ratio to Moving Averages

Year / Quarter	*Sales*	*4 – Period Moving Total*	*Centred Total*	*4 Period M*	$\frac{Y}{M} \times 100$
1986 *I*	34		...	...	...
II	33 →		...	...	...
III	34 →	138 →	279	34.9	97.4
IV	37 →	141 →	284	35.5	104.2
1987 *I*	37 →	143 →	289	36.1	102.5
II	35 →	146 →	294	36.8	95.1
III	37 →	148 →	298	37.3	99.2
IV	39 →	150 →	302	37.8	103.2
1988 *I*	39 →	152 →	305	38.1	102.4
II	37 →	153 →	307	38.4	96.4
III	38 →	154 →	311	38.9	97.7
IV	40 →	157 →	318	39.8	100.5
1989 *I*	42 →	161 →	326	40.8	102.9
II	41 →	165 →	334	41.8	98.1
III	42	169	...	...	...
IV	44		...	...	...

Calculation of Seasonal Indices

Year	*I*	*II*	*III*	*IV*
1986	–	–	97.4	104.2
1987	102.5	95.1	99.2	103.2
1988	102.4	96.4	97.7	100.5
1989	102.9	98.1	–	–
Total	307.8	289.6	294.3	307.9
A_i	102.6	96.5	98.1	102.6
S.I.	102.7	96.5	98.1	102.7

Note that the Grand Average $G = \frac{399.8}{4} = 99.95$. Also check that the sum of indices is 400.

Example 10.6: The following table gives the figures of imports of a certain commodity during four years. Determine the seasonal variations by ratio to moving average method on the assumption that various components are additive.

Year	I	II	III	IV
1988	22	26	25	27
1989	24	20	29	35
1990	31	27	38	40
1991	41	40	46	48

Solution:

Calculation of Moving Averages and Short-term fluctuations

Year / Quarter	Imports (Y)	4 – Quarter Moving Total	Centred Total	4 Quarter M.A.	Short Term Fluctuations (Y– M.A.)
1988 I	22		...	...	...
II	26 →		...	...	...
III	25 →	100 →	202	25.2	– 0.2
IV	27 →	102 →	198	24.8	2.2
1989 I	24 →	96 →	196	24.5	– 0.5
II	20 →	100 →	208	26.0	– 6.0
III	29 →	108 →	223	27.9	1.1
IV	35 →	115 →	237	29.6	5.4
1990 I	31 →	122 →	253	31.6	– 0.6
II	27 →	131 →	267	33.4	– 6.4
III	38 →	136 →	282	35.3	2.7
IV	40 →	146 →	305	38.1	1.9
1991 I	41 →	159 →	326	40.8	0.2
II	40 →	167 →	342	42.8	– 2.8
III	46	175	...	...	...
IV	48		...	...	...

Calculation of Seasonal Fluctuations

Year	I	II	III	IV
1988	–	–	– 0.2	2.2
1989	– 0.5	– 0.6	1.1	5.4
1990	– 0.6	– 6.4	2.7	1.9
1991	0.2	– 2.8	–	–
Total	– 0.9	– 15.2	3.6	9.5
A_i	– 0.30	– 5.07	1.20	3.17

Since the total of these averages is not equal to zero, these are required to be adjusted. Sum of averages $\Sigma A_i = -1$.

Therefore, the correction factor $= \frac{\sum A_i}{4} = -0.25$. This value is subtracted from each average to get seasonal fluctuations. Thus, the seasonal fluctuations of I, II, III and IV quarters are –0.05, –4.82, 1.45 and 3.42 respectively.

Merits and Demerits

This method assumes that all the four components of a time series are present and, therefore, widely used for measuring seasonal variations. However, the seasonal variations are not completely eliminated if the cycles of these variations are not of regular nature. Further, some information is always lost at the ends of the time series.

Link Relatives Method

This method is based on the assumption that the trend is linear and cyclical variations are of uniform pattern. As discussed in earlier chapter, the link relatives are percentages of the current period (quarter or month) as compared with previous period. With the computation of link relatives and their average, the effect of cyclical and random component is minimised. Further, the trend gets eliminated in the process of adjustment of chained relatives. The following steps are involved in the computation of seasonal indices by this method:

(i) Compute the link relative (L.R.) of each period by dividing the figure of that period with the figure of previous period. For example, link relative of 3rd quarter

$$= \frac{\text{figure of 3rd quarter}}{\text{figure of 2nd quarter}} \times 100$$

(ii) Obtain the average of link relatives of a given quarter (or month) of various years. A.M. or M_4 can be used for this purpose. Theoretically, the later is preferable because the former gives undue importance to extreme items.

(iii) These averages are converted into chained relatives by assuming the chained relative of the first quarter (or month) equal to 100. The chained relative (C.R.) for the current period (quarter or month)

$$= \frac{C.R. \text{ of the previous period} \times L.R. \text{ of the current period}}{100}$$

(iv) Compute the C.R. of first quarter (or month) on the basis of the last quarter (or month). This is given by

$$\frac{C.R. \text{ of last quarter (or month)} \times \text{average } L.R. \text{ of 1st quarter (or month)}}{100}$$

This value, in general, be different from 100 due to long term trend in the data. The chained relatives, obtained above, are to be adjusted for the effect of this trend. The adjustment factor is

$$d = \frac{1}{4}[\text{New } C.R. \text{ for 1st quarter} - 100] \text{ for quarterly data}$$

and $d = \frac{1}{12}[\text{New } C.R. \text{ for 1st month} - 100]$ for monthly data.

On the assumption that the trend is linear, d, $2d$, $3d$, etc., is respectively subtracted from the 2nd, 3rd, 4th, etc., quarter (or month).

(v) Express the adjusted chained relatives as a percentage of their average to obtain seasonal indices.

(vi) Make sure that the sum of these indices is 400 for quarterly data and 1200 for monthly data.

Example 10.7: Determine the seasonal indices from the following data by the method of link relatives:

Year	*1st Qr*	*2nd Qr*	*3rd Qr*	*4th Qr*
1985	26	19	15	10
1986	36	29	23	22
1987	40	25	20	15
1988	46	26	20	18
1989	42	28	24	21

Solution:

Calculation Table

Year	*I*	*II*	*III*	*IV*
1985	–	73.1	78.9	66.7
1986	360.0	80.5	79.3	95.7
1987	181.8	62.5	80.0	75.0
1988	306.7	56.5	76.9	90.0
1989	233.3	66.7	85.7	87.5
Total	1081.8	339.3	400.8	414.9
Mean	270.5	67.9	80.2	83.0
C.R.	100.0	67.9	54.5	45.2
C.R.(adjusted)	100.0	62.3	43.3	28.4
S.I.	170.9	106.5	74.0	48.6

The chained relative (C.R.) of the 1st quarter on the basis of C.R. of the 4th quarter $= \frac{270.5 \times 45.2}{100} = 122.3$.

The trend adjustment factor $d = \frac{1}{4}(122.3 - 100) = 5.6$

Thus, the adjusted C.R. of 1st quarter = 100

and for 2nd = 67.9 – 5.6 = 62.3

for 3rd = 54.5 – 2 × 5.6 = 43.3

for 4th = 45.2 – 3 × 5.6 = 28.4

The grand average of adjusted C.R., $G = \frac{100 + 62.3 + 43.3 + 28.4}{4} = 58.5$

The seasonal index of a quarter $= \frac{\text{Adjusted } C.R. \times 100}{G}$.

Merits and Demerits

This method is less complicated than the ratio to moving average and the ratio to trend methods. However, this method is based upon the assumption of a linear trend which may not always hold true.

Remarks: Looking at the merits and demerits of various methods of measuring seasonal variations, we find that the ratio to moving average method is most general and, therefore, most popular method of measuring seasonal variations.

Causes of Cyclical Variations

Cyclical variations are revealed by most of the economic and business time series and, therefore, are also termed as trade (or business) cycles. Any trade cycle has four phases which are respectively known as boom, recession, depression and recovery phases. These phases are shown in Fig. 10.1. Various phases repeat themselves regularly one after another in the given sequence. The time interval between two identical phases is known as the period of cyclical variations. The period is always greater than one year. Normally, the period of cyclical variations lies between 3 to 10 years.

Objectives of Measuring Cyclical Variations

The main objectives of measuring cyclical variations are:

(i) To analyse the behaviour of cyclical variations in the past.

(ii) To predict the effect of cyclical variations so as to provide guidelines for future business policies.

Random or Irregular Variations

As the name suggests, these variations do not reveal any regular pattern of movements. These variations are caused by random factors such as strikes, floods, fire, war, famines, etc. Random variations are that component of a time series which cannot be explained in terms of any of the components discussed so far. This component is obtained as a residue after the elimination of trend, seasonal and cyclical components and hence is often termed as residual component.

Random variations are usually short-term variations but sometimes their effect may be so intense that the value of trend may get permanently affected.

10.3 TREND ANALYSIS

The following are the principal methods of measuring trend from a given time series:

I. Mathematical Trends
   - (i) Method of Least Squares
     - (a) Fitting of Linear Trend
     - (b) Fitting of Parabolic Trend
     - (c) Fitting of Exponential Trend

Mathematical Trends

The method of fitting a mathematical trend to given time series data is perhaps the most popular and satisfactory. The form of mathematical equation used for the determination of trend depends upon the

nature of the broad idea of trend, obtained by graphic representation of data or otherwise. Some popularly known forms of trend are linear, parabolic, and exponential and growth curves.

Method of Least Squares

This is one of the most popular methods of fitting a mathematical trend. The fitted trend is termed as the best in the sense that the sum of squares of deviations of observations, from it, is minimized. We shall use this method in the fitting of following trends:

1. ***Linear Trend:*** The general form of a linear trend is given by the equation $Y_t = a + bt$, where t denotes time, Y_t is the trend value (note that trend values, in mathematical models, will be denoted by Y_t rather than by T_t for the sake of convenience) of variable at time t and a (> 0) and b (a real number) are constants. The constant a can be interpreted as the value of trend (Y_t) when $t = 0$ and b gives the change in Y_t per unit change in time. It should be noted that the rate of change of Y_t is always constant in case of a linear trend. This implies that for equal absolute changes in t, there are correspondingly equal absolute changes in Y_t. Further, a linear trend can be rising or falling according as $b > 0$ or < 0, as shown in the following figure.

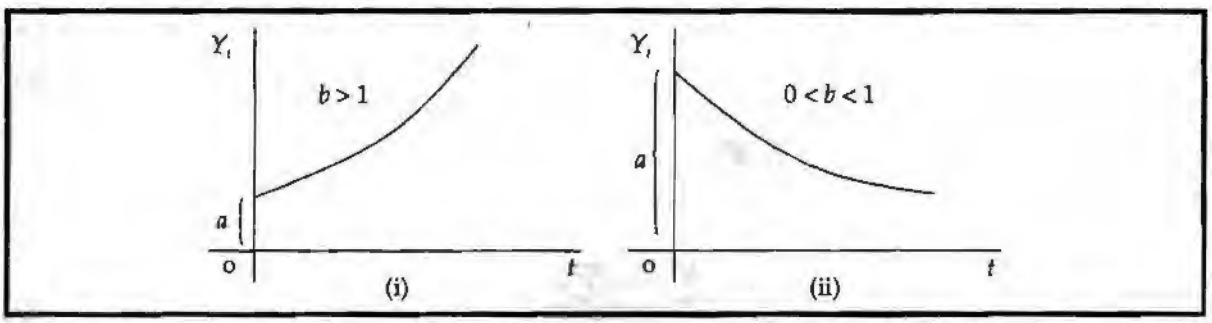


Figure 10.2

Linear Trend

2. ***Parabolic Trend:*** The general form of a parabolic trend is $Y_t = a + bt + ct^2$, where a, b and c are constants. Here the rate of change of Y_t is different at different time periods. The possible shapes of parabolic trends are shown below:

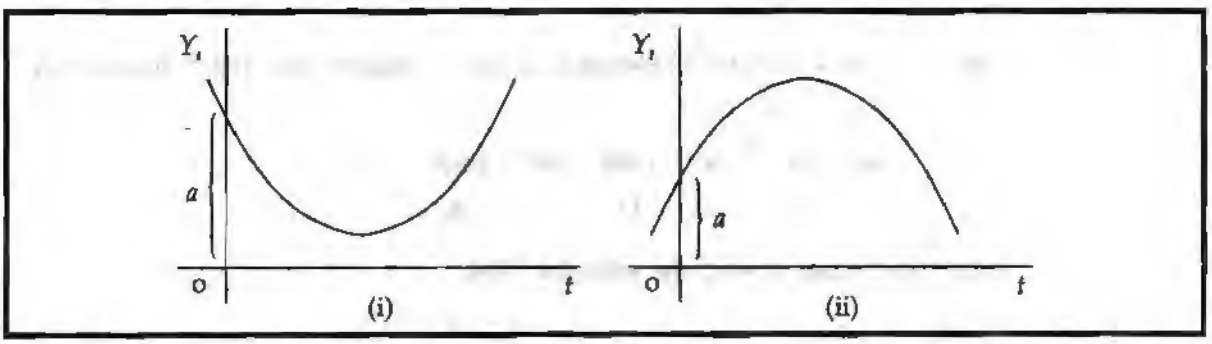


Figure 10.3

Parabolic Trend

We note that the rate of change of Y_t is increasing in the first case while it is decreasing in the second.

Fitting of Parabolic Trend

The mathematical form of a parabolic trend is given by $Y_t = a + bt + ct^2$ or $Y = a + bt + ct^2$ (dropping the subscript for convenience). Here a, b and c are constants to be determined from the given data.

Using the method of least squares, the normal equations for the simultaneous solution of a, b, and c are :

$$\Sigma Y = na + b\Sigma t + c\Sigma t^2$$

$$\Sigma tY = a\Sigma t + b\Sigma t^2 + c\Sigma t^3$$

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

By selecting a suitable year of origin, i.e., define X = t - origin such that $\Sigma X = 0$, the computation work can be considerably simplified. Also note that if $\Sigma X = 0$, then ΣX^3 will also be equal to zero. Thus, the above equations can be rewritten as:

$$\Sigma Y = na + c\Sigma X^2 \quad \text{.... (i)}$$

$$\Sigma XY = b\Sigma X^2 \quad \text{.... (ii)}$$

$$\Sigma X^2Y = a\Sigma X^2 + c\Sigma X^4 \quad \text{.... (iii)}$$

From equation (ii), we get $b = \dfrac{\sum XY}{\sum X^2}$ .... (iv)

Further, from equation (i), we get $a = \dfrac{\sum Y - c\sum X^2}{n}$ .... (v)

And from equation (iii), we get $c = \dfrac{n\sum X^2Y - \left(\sum X^2\right)\left(\sum Y\right)}{n\sum X^4 - \left(\sum X^2\right)^2}$ .... (vi)

Thus, equations (iv), (v) and (vi) can be used to determine the values of the constants a, b and c.

Example 10.8

Fit a parabolic trend $Y = a + bt + ct^2$ to the following data, where t denotes years and Y denotes output (in thousand units).

t	:	1981	1982	1983	1984	1985	1986	1987	1988	1989
Y	:	2	6	7	8	10	11	11	10	9

Also compute the trend values. Predict the value for 1990.

Solution

Calculation Table

t	Y	X = t − 1985	XY	X^2Y	X^2	X^3	X^4	Trend Values
1981	2	−4	−8	32	16	−64	256	2.28
1982	6	−3	−18	54	9	−27	81	5.02
1983	7	−2	−14	28	4	−8	16	7.22
1984	8	−1	−8	8	1	−1	1	8.88
1985	10	0	0	0	0	0	0	10.00
1986	11	1	11	11	1	1	1	10.58
1987	11	2	22	44	4	8	16	10.62
1988	10	3	30	90	9	27	81	10.12
1989	9	4	36	144	16	64	256	9.08
Total	74	0	51	411	60	0	708	

From the above table, we can write

$$b = \frac{51}{60} = 0.85$$

$$c = \frac{9 \times 411 - 60 \times 74}{9 \times 708 - (60)^2} = -0.27$$

$$a = \frac{74 - (-0.27) \times 60}{9} = 10.0$$

∴ The fitted trend equation is $Y = 10.0 + 0.85X - 0.27X^2$,

with origin = 1985 and unit of X = 1 year.

Various trend values are calculated by substituting appropriate values of X in the above equation. These values are shown in the last column of the above table.

The predicted value for 1990 is given by

$$Y = 10.0 + 0.85 \times 5 - 0.27 \times 25 = 7.5$$

Example 10.9

The prices of a commodity during 1981-86 are given below. Fit a second degree parabola to the following data. Calculate the trend values and estimate the price of the commodity in 1986.

Year	:	1981	1982	1983	1984	1985	1986
Price	:	110	114	120	138	152	218

Solution

Calculation Table

Year (t)	Price (Y)	$X = 2(t-1983.5)$	XY	X^2Y	X^2	X^4	Trend Values
1981	110	−5	−550	2750	25	625	114.40
1982	114	−3	−342	1026	9	81	109.12
1983	120	−1	−120	120	1	1	116.08
1984	138	1	138	138	1	1	135.28
1985	152	3	456	1368	9	81	166.72
1986	218	5	1090	5450	25	625	210.40
	852	0	672	10852	70	1414	

From the above table, we get

$$b = \frac{672}{70} = 9.6,\ c = \frac{6 \times 10852 - 70 \times 852}{6 \times 1414 - (70)^2} = 1.53 \text{ and } a = \frac{852 - 1.53 \times 70}{6} = 124.15$$

∴ The equation of parabolic trend is $Y = 124.15 + 10.6X + 1.53X^2$, with year of origin = 1983.5 or 1st January, 1984 and the unit of $X = \frac{1}{2}$ year.

The calculated trend values are shown in the last column of the above table.

The price of the commodity in 1986 is obtained by substituting X = 5, in the above equation.

Thus, $Y = 124.15 + 9.5 \times 5 + 1.53 \times 25 = 210.4$

Exponential Trend

The general form of an exponential trend is given by the equation $Y_t = a.b^t$, where a and b are positive constants. This implies that values

Y_t changes by a constant percentage per unit of time. For example, if $a = 50$ and $b = 1.05$, then

$Y_1 = 50 \times 1.05 \Rightarrow$ 5% increase in the value of a.

Similarly, $Y_2 = 50 \times (1.05)^2 = Y_1 \times 1.05 \Rightarrow$ 5% increase in the value of Y_1 and so on.

We note that when $b > 1$, the exponential trend is increasing. In a similar way, it would be decreasing when $0 < b < 1$, as shown in the Figure 9.4.

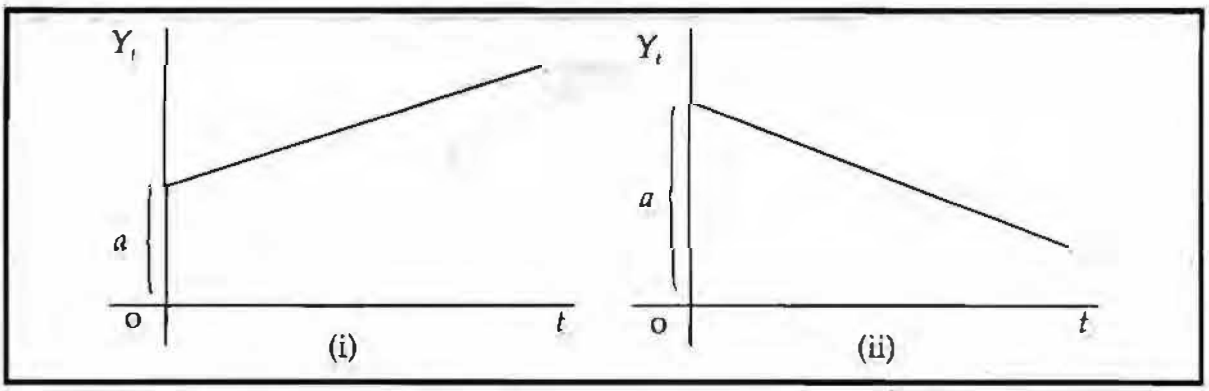


Figure 10.4

Exponential Trend

Check Your Progress 1

Fill in the blanks

1. ..... is the general tendency of the data to increase or decrease or stagnate over a long period of time.
2. The general form of a parabolic trend is .....
3. .....method is used when cyclical variations are absent from the data.

10.4 TIME SERIES ANALYSIS IN FORECASTING

Time has strange, fascinating and little understood properties. Virtually every process on earth is determined by a time variable. One of the most frequently encountered managerial decision situations involving forecasting is to measure the effect that time has on the sales of a product, the market price of a security, the output of individuals, work shifts, companies, industries, societies and so on. A fundamental conceptual model in all of these situations is the product life cycle concept which goes through four stages – introduction, growth, maturity and decline. Let us look at this concept in greater detail before we apply it.

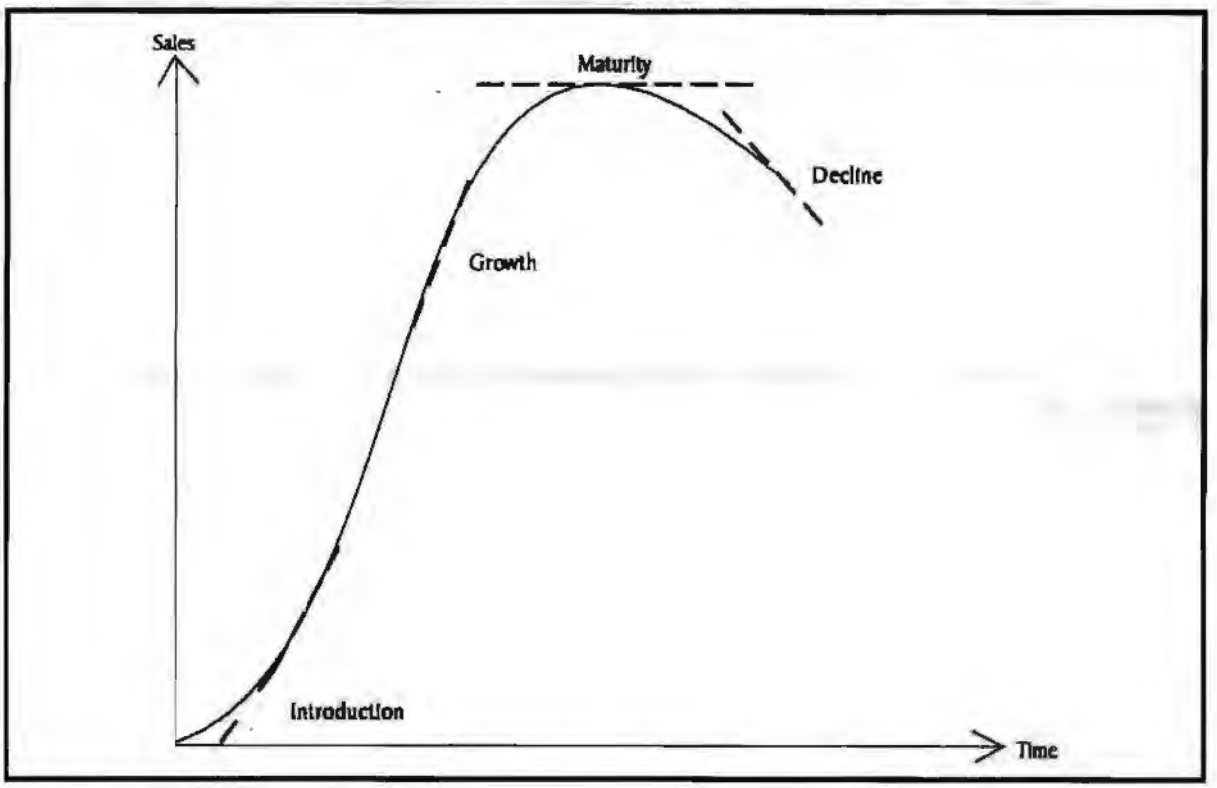


Figure 10.5

Product Life-Cycle

Figure 10.5 depicts various stages in the life of a product. The sales performance of this product goes through the four stages—introduction, growth, maturity and decline. Data have been plotted and regression lines fitted to each of the four environments. Thus, when a sales forecast is made and the target horizon falls within the same stage, the linear fit yields valid results. If, however, the target horizon falls into a future stage, a linear forecast may be erroneous. In this case a curve should be fitted as shown. It is usually lightly speculative to select a forecasting horizon that spans more than two stages.

Another point of interest is the behaviour of the sales variable over the short run. It fluctuates between a succession of peaks and troughs. How do these come about? In order to answer this question, the time series, must be decomposed. Then four independent motors for this behaviour become visible. First there is a long-term or secular trend (T) which is primarily noticeable within each stage of the cycle and over the entire cycle. Secondly cyclical variations (C) which are caused by an economy's business cycles affect product sales. Such cycles, whose origins are little understood, exist for all economies. Thirdly the product's sales may be influenced by the seasonality (S) of the item, and finally there may be the irregular (I) affects of inclement such as weather, strikes and so forth. In equation form the decomposed time series appears as $TS = T + C + S + I$.

This creates a complex situation in time series analysis. Each factor must be quantified and its effect ascertained upon product sales. Let us see how this is done. The long-term trend effect T is reflected in the slope b of the regression equation. We already know how b is calculated even though minor modifications of the decision formulas will be encountered soon. The quantification of the cyclical component C is beyond the scope of this book. However, since business cycles always proceed from peak to trough to new peak and so on, their positive and negative effects upon a product's sales cancel

out in the long-run. Hence in managerial, as opposed to economic, decision making, the sum effect of the business cycles may be set equal to zero. This eliminates the *C* factor from the equation. Seasonality, if present, is something that must be taken into consideration because it is a product-inherent variable and therefore it is under the immediate control of the decision maker. We will quantify the *S* component and keep it in the equation.

Finally, there are the irregular variations. Do we know in July whether the weather will be sunny and mild during the four weeks before Diwali? We don't, but we know that if this happens, Diwali sales will be severely impacted. Can we forecast such horrible weather conditions? Not really. We cannot forecast them because they cannot be quantified—a rather unpleasant characteristic they share with all other type of irregular variations like strikes, earthquakes, power failure, etc. Yet, something strange usually happens after such an irregular variation from "normal" has occurred. Whatever people did not do because of it, like not buying a product, they attempt to catch up with quickly. Therefore the I factor effect may also be assumed to cancel out over time and it may be dropped from the equation which then appears to the manager as $TS = T + S$.

Linear Analysis

We will construct again the best fitting regression line by the method of least squares. In order to illustrate the procedure, let us use a data set from box given below.

Box 10.1: Smart Discount Stores

There are 2117 Smart stores in the India (the chain is building up). It is one of India's most interesting discounters tracing its origins back to 1980's and the opening of the first Smart store. At present Smart has reached an "upgrading" phase like so many discounters before.

Given the data below, perform the indicated analyses.

YEAR	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990
EARNINGS										
PER SHARE	19.0	17.5	20.7	28.4	27.4	23.9	21.1	16.1	8.5	11.1
DIVIDENDS	9.9	9.5	9.0	8.1	6.8	5.0	3.0	2.4	2.2	1.9
PER SHARE										
PRE-TAX	2.1	2.0	3.1	4.9	5.4	5.7	5.8	5.8	3.3	5.3
MARGIN										

It involves the dividend payments per share of the Smart, a well-known discount store chain, for the years 1990 through 19910. Suppose that a potential investor would like to know the dividend payment for 2001. The data are recorded in the work sheet (Table 10.1). First, however, turn your attention to Figure 10.6 which shows the plot for this problem.

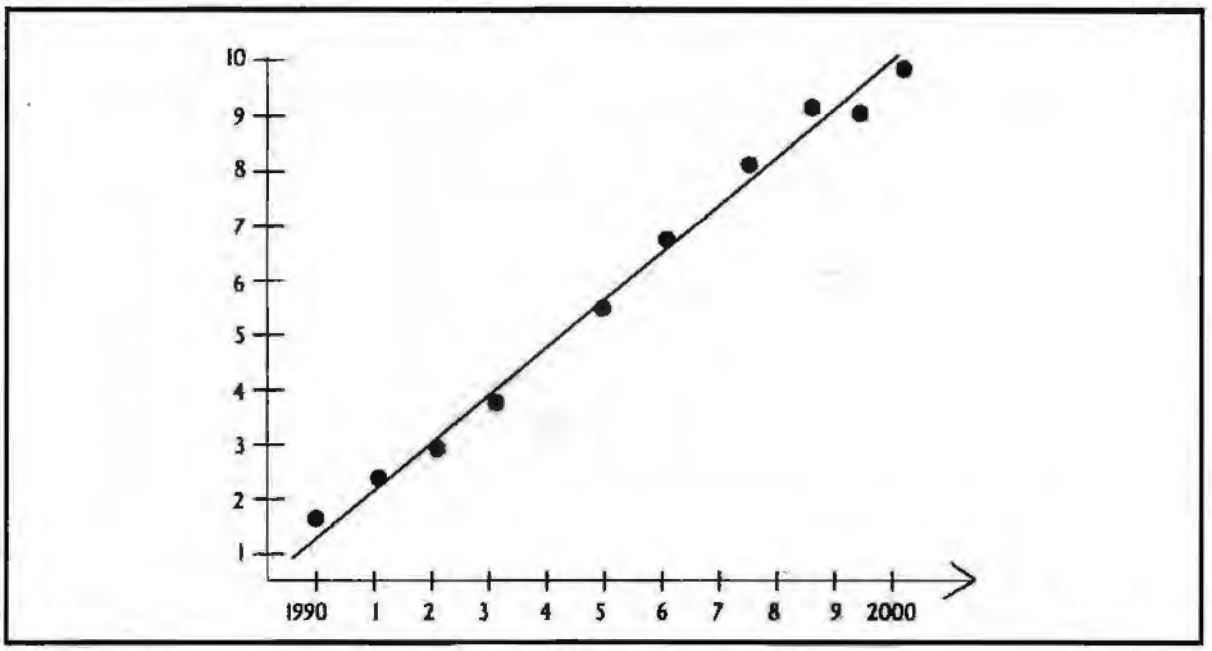


Figure 10.6

Plot of Dividend Values

Think for a moment about the qualitative nature of the time variable. It is expressed in years in this case but could be quarters, months, days, hours, minutes or any other time measurement unit. How does it differ from advertising expenditures, the independent variable that we examine in the preceding section? Is there a difference in the effect that a unit of each has on the dependent variable, or, Rs 1 million in one case and 1 year in the other? Time, as you can readily see is constant. One year has the same effects as any other. This is not true for advertising expenditures, especially when you leave the linear environment and enter the nonlinear environments. Then there may be qualitative difference in the sales impact as advertising expenditures are increased or decreased by unit. Since time is constant in its effect, we may code the variable rather than to use the actual years or other time units x values. This code assigns a 1 to the first time period in the series and continues in unit distances to the nth period. Do not start with a zero as this may cause some computer programs to reject the input. The code is based on the fact that the unit periods are constant, and therefore their sum may be set equal to zero. See what effect this has on the normal equations for the straight line.

$$\Sigma y = na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

If $\Sigma x = 0$, the equations reduce to

$$\Sigma y = na$$

$$\Sigma xy = b\Sigma x^2$$

Which allow the direct solution for a and b as follows

$$a = \frac{\Sigma y}{n}$$

$$b = \frac{\Sigma xy}{\Sigma x^2}$$

This form simplifies the calculations substantially compared to the previous formulas. The code, however, that allows to set $\Sigma x = 0$ must incorporate the integrity of a unit distance series. Thus if the series is odd-numbered, the midpoint is set equal to zero and the code completed by negative and positive unit distances of $x = 1$ where each x unit stands for one year or other time period. If the series is even-numbered, let us say it ran from 1990 to 1999, the two midpoints (1994/1995) are set equal to −1 and +1, respectively. Since there is now a distance of $x = 2$ between + 1 (−1, 0, +1), the code continues by negative and positive units distance of $x = 2$ where each x unit stands for one-half year or other time period.

The worksheet is in Table 10.1 and calculations are as follows:

Table 10.1: Worksheet

YEAR	Code for an Even Series X	YEAR	Code for an Odd Series X	Dividend payments in Rs Y	XY	X^2
1990	-9					
1991	-7	1991	-4	2.2	-8.8	16
1992	-5	1992	-3	2.4	-7.2	9
1993	-3	1993	-2	3.0	-6.0	4
1994	-1	1994	-1	5.0	-5.0	1
1995	1	1995	0	6.8	0	0
1996	3	1996	1	8.1	8.1	1
1997	5	1997	2	9.0	18.0	4
1998	7	1998	3	9.5	28.5	9
1999	9	1999	4	9.9	39.6	16
Total	0		0	55.9	67.2	60

Then $a = \frac{5.59}{9} = 6.211$

$$b = \frac{67.2}{60} = 1.12$$

and $Y_c = 6.211 + 1.12x$

origin 1995

x in 1 year units

The regression equation is plotted in Figure 10.6. Note that in the case of time series analysis, the origin of the code and the x units must be defined as part of the regression equation. In our problem the investor would like to obtain a dividend forecast for 2001. Since the origin is 1995 ($x = 0$) and $x = 1$ year units, the code for 2001 is $x = 6$. Therefore the forecast is $y_c = 6.211+1.12\ (6) = \text{Rs } 12.9$. If the time series had been even numbered, let us say that dividend payments for 1990 had been included in the

forecasting study, the definition under the regression equation would have read

origin 1994/95

x in 6 month units.

Thus, we know that for 1995, $x = 1$; and since we must use $x = 2$ units for each year, the code value for 2001 would be $x = 13$. Once the y_c value has been obtained, b is tested for significance and the 95% confidence interval constructed as previously shown.

Time series analysis is a long-term forecasting tool. Hence, it addresses itself to the trend component T in our time series equation $TS = T + S$. In the dividend forecast, $b = 1.120$ was calculated which means that in the environment that is reflected in the set, smart increased the dividend payments on a average by Rs 1.12 per year. Let us now turn out attention to the seasonal variation component that may be present in a time series. A product's seasonality is shown by the regularly recurring increases or decreases in sales or production that are caused by seasonal influences. In the case of some products, their seasonality is quite apparent.

An obvious example virtually all non-animal agricultural commodities may be cited. Seasonality of other products may be more difficult to detect. Take hogs in order to stay on the farm. Are they seasonal? They are lusty breeders and could not care less about seasonal influences. Yet, there is an induced season by the corn harvest. If corn is plentiful and cheap, farmers raise more hogs. This is known as the corn-hog cycle. Or take automobiles, Indian manufacturers are used to introduce major design or technological changes once every generation. This "season" has now been shortened somewhat. How about computers? There the season even has a special name. It is called a generation and prior to increased competitive pressures within the industry it used to be about seven years long. Our stock market investor knows that stock trades on the Stock Exchanges are seasonal. The daily season is V-shaped starting the trading with a relatively high volume which tapers off toward the lunch hour to pick up again in the afternoon. And so it goes with many other products, not ordinarily thought of as being seasonal.

Let us quantify this seasonality and illustrate how it may be used in a decision situation. There are, as is often the case, a number of decision tools that may be applied. The reader may be familiar with the term ratio-to-moving-average. It is a widely used method for constructing a seasonal index and programs are available in most larger computer libraries. Usually the method assumes a 12 - period season like the twelve months of the year. There is a more efficient method which yields good statistical results. It is especially helpful in manual calculations of the seasonal index and when the number of seasonal periods is small like the four quarters of a year, the six hours of a stock exchange trading day or the five days of a work week. This method is known as simple average and will be used for illustration purposes.

To stay with the investment environment of this chapter section, let us calculate a seasonal index for shares traded on the Stock Exchange from July 2 through July 7, 1999. This period includes the July 4 week-end. Volume of shares (DATA) for each trending day (SEASON) is given in thousands of shares per hour. The Individual steps of the analysis (OPERATIONS) are discussed in detail for each column of the worksheet below.

Table 10.2: Worksheet for Analysis

	Column (2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
					Trading volume ('000)				
Hour	Total Variation (TS)	Trend Variation (T)	Seasonal variation TS-T	Seasonal Index	7/2	7/3	7/6	7/7	Avg. for four days
10-11	0.965	0	0.965	110.6	12.00	12.25	15.44	16.72	14.10
11-12	0.245	0.159	0.086	103.7	10.40	11.75	15.04	16.32	13.38
12-13	–0.885	0.318	–1.117	94.2	10.55	10.06	12.95	15.44	12.25
13-14	–1.555	0.477	–2.032	87.1	9.55	9.46	12.05	15.24	11.58
14-15	0.395	0.636	–0.241	101.1	11.02	11.55	14.82	16.73	13.53
15-16	0.835	0.795	0.040	103.3	11.58	12.25	15.38	16.69	13.97
Average			–0.383	600	10.85	11.22	14.28	16.19	13.135

As you inspect the data columns, you notice the V-shaped season for each trading day. You also notice in the total daily volume that there is a increase in shares traded. Hence, you can expect a positive slope of the regression line. The hourly mean number of shares is indicated also. This is the more important value because we are interested in quantifying a season by the hour for each trading day. Now turn to the operations. In last column the hourly trading activity for the four days has been summed. In this total all time series factors are assumed to be incorporated. You will recall that the positive or negative cyclical and irregular component effect is assumed to cancel out over time. Hence averaging the trading volume over a long term data set eliminates both components, yielding $TS = T + S$. You may ask, are four days a sufficiently long time span? The answer is NO. In a real study you would probably use 15 to 25 yearly averages for each trading hour. In an on-the-job application of this tool, you will have to know the specific time horizon in order to effectively eliminate cyclical and irregular variations. But by and large, what is a long or short time span depend upon situation.

In order to isolate the trend component (T) so that it may be subtracted from column (2) in the Table 10.2, yielding seasonal variation, the slope (b) of the regression line must be calculated. (Rember: b is T.) The necessary calculations are performed below using the mean hourly trading volume for each day. But since we are interested in an index by the hour, the calculated daily b value must be apportioned to each hour. This is accomplished by a further division by six—the number of trading hours. The result is entered in column (3). Note that the origin of a time series is always zero. The origin of the time series is always the first period of the season. In our case this is the 10-11 trading hour. Therefore the first entry in column (3) is always zero to be followed by the equal (since this is a linear analysis) summed increment of the apportioned b-value.

Table 10.3: Worksheet for Trend Calcutation

Day	Code	Average Hourly Trading Volume Per Day		
x	x^1	Y	x y	x^2
7/1	–3	10.85	–32.55	9
7/5	–1	11.22	–11.22	1
7/6	1	14.28	14.28	1
7/7	3	16.19	48.57	9
Total		52.54	19.08	20

$$b = \frac{\sum xy}{\sum x^2}$$

$$= \frac{19.08}{20}$$

$$= 0.954$$

and the apportioned b-value is

$$\frac{0.954}{6} = 0.159$$

It is not necessary to calculate the y-intercept (a) in this analysis unless of course, you wish to combine it with a long-term forecast of daily trading volume. Then, just to review the calculations, you would find:

$$a = \frac{\sum y}{n}$$

$$= \frac{52.54}{4}$$

$$= 13.135$$

and

$$y_c = 13.135 + 0.954\,x$$

origin 7/5 and 7/6

x in half trading day units.

In column (4) $TS - T = S$ is performed. Column (4) is already a measure of seasonal variation. But in order to standardize the answer so that it may be compared with other stock exchange, for example, it is customary to convert the values in column (4) to a seasonal index. Every index has a base of 100 and the values above or below the base indicate percentages of above or below "normal" activity, hence the season. Since the base of column (5) is 100, the mean of the column should be 100 and the total 600 since there are 6 trading hours. In order to convert the obtained values of column (4) to index numbers, each of its entries is added to the total mean and then is divided by the column mean added to total mean and multiplied by 100 yielding the corresponding entry in column (5). It is customary to show index numbers with one significant digit.

Column (6) shows the seasonal effect of this decision variable—share trading on the Stock Exchange. Regardless of heavy or light daily volume, the first hour volume is the heaviest by far. It is 7.4% above what may be considered average trading volume for any given day. Keep in mind that a very limited data set was used in this analysis and while the season, reaching its low point between 1 and 2 p.m., is generally correctly depicted, individual index members may be exaggerated. What managerial action programs would result from analyses such as this? Would traders go out for tea and samosas between 10-11? How about lunch between 1-2? When would brokers call clients with hot or luke-warm tips? Assuming that a decrease in volume means a decrease in prices in general during the trading day, when would a savvy trader buy? When would he sell? Think of some other intervening variables and you have yourself a nice little bull session in one of Dalal Street's watering holes. If, in addition, you make money for yourself or firm, then, you have got it.

Non-linear Analysis

Any number of different curves may be fitted to a data set. The most widely used program in computer libraries, known as CURFIT, offers a minimum of 5 curves plus the straight line. The curves may differ from program to program. So, which ones are the "best" ones? There is no answer. Every forecaster has to decide individually about his pet forecasting tools. They appear to be promising decision tools especially in problem situations that in some way incorporate the life cycle concept and the range of such problems is vast, indeed.

As we know from many empirical studies, achievement is usually normally distributed. Growth, on the other hand, seems to be exponentially distributed. The same holds true for decline. As the life cycle moves from growth to maturity, a parabolic trend may often be used as the forecasting tool. Now, we know—again from all sorts of empirical evidence—that trees don't grow into the high heavens. Even the most spectacular growth must come to an end. Therefore, when using the exponential forecast, care must be taken that the eventual ceiling or floor (in the case of a decline) is not overlooked. The modified exponential trend has the ceiling or floor build in.

One final piece of advice before we start fitting curves. If you can do it by straight line, do it. By extending the planning and forecasting horizon over a reasonable shorter period rather than spectacular but dangerous longer period, the straight line can serve as useful prediction tool.

Check Your Progress 2

Fill in the blanks:

1. As the life cycle moves from growth to maturity, a .....may often be used as the forecasting tool.
2. We will construct again the best fitting regression line by the method of.....
3. The most widely used program in computer libraries, known as ....., offers a minimum of curves plus the straight line.

10.5 SUMMARY

The analysis of time series implies its decomposition into various factors that affect the value of its variable in a given period. It is a quantitative and objective evaluation of the effects of various factors on the activity under consideration. Secular trend or simply trend is the general tendency of the data to increase or decrease or stagnate over a long period of time. Most of the business and economic time series would reveal a tendency to increase or to decrease over a number of years. This time interval is known as the period of oscillation. The oscillatocy movements are termed as Seasonal Variations if their period of oscillation is equal to one year, and as Cyclical Variations if the period is greater than one yeat. The measurement of seasonal variation is done by isolating them from other components of a time series. There are four methods commonly used for the measurement of seasonal variations.

Graphic or Free Hand Curve Method is the simplest method of studying the trend. The given time series data are plotted on a graph paper by taking time on X-axis and the other variable on Y-axis. The method of fitting a mathematical trend to given time series data is perhaps the most popular and satisfactory. The form of mathematical equation used for the determination of trend depends

upon the nature of the broad idea of trend, obtained by graphic representation of data or otherwise. Method of Least Square is one of the most popular methods of fitting a mathematical trend. The fitted trend is termed as the best in the sense that the sum of squares of deviations of observations, from it, is minimized. Any number of different curves may be fitted to a data set. The most widely used program in computer libraries, known as CURFIT, offers a minimum of 5 curves plus the straight line. The curves may differ from program to program.

10.6 KEYWORDS

- Time Series
- Periodic Variations
- Seasonal Variations
- Parabolic Trend
- Trend Analyses
- Least Squares
- Cyclical Variations

10.7 REVIEW QUESTIONS

1. Write short note on:
   (a) Time Series Analysis
   (b) Secular Trend
   (c) Periodic Variations
   (d) Irregular Variations
   (e) Mathematical Trends
   (f) Parabolic Trend
2. "All periodic variations are not necessarily seasonal". Discuss the above statement with a suitable example.
3. Explain the meaning and objectives of time series analysis. Describe briefly the methods of measurement of trend.
4. What is a time series? What are its main components? How would you study the seasonal variations in any time series?
5. Distinguish between secular trend and periodic variations. How would you measure trend in a time series data by the method of least squares? Explain your answer with an example.
6. Fit a straight line trend to the following data on steel production (in M. tonnes). Predict the value for 1992.

Year	:	1985	1986	1987	1988	1989	1990	1991
Production	:	80	84	90	93	98	100	104

7. Fit a straight line trend by method of least squares to the following data on earnings (Rs lakh) of a firm. (a) Assuming that the same trend continues, what would be the predicted earnings for the

year 1987? (b) Convert this equation into a monthly equation with January, 1985 as origin and estimate the values for November, 1984 and April, 1985.

Year	:	1978	1979	1980	1981	1982	1983	1984	1985
Earnings	:	38	40	65	72	79	60	87	95

8. Given below are the figures of production of a sugar factory in '000 tonnes

Year	:	1981	1982	1983	1984	1985	1986	1987
Production	:	77	88	94	85	91	98	90

(i) Fit a straight-line trend by method of least squares.

(ii) Calculate trend values and plot observed values and trend values on a graph.

(iii) Predict the production of factory for 1989 and 1993 on the assumption that the same trend continues.

(iv) Comment on the validity of prediction for 1993.

9. Compute the trend line by the method of least squares from the data on profits (in Rs '000) of a firm, given below:

Year	:	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991
Profits	:	110	125	115	135	150	165	155	175	180	200

10. Fit a linear trend to the following data, on average monthly output, with origin at mid-point of the year 1980. Convert this into a monthly trend equation. Estimate the average output for June and August, 1980.

Year	:	1976	1977	1978	1979	1980	1981	1982	1983	1984
Output	:	6.3	7.4	9.3	7.4	8.3	10.6	9.0	8.7	7.9

11. Draw a free hand curve showing trend of the following data:

Years	Output(in tonnes)	Years	Output(in tonnes)
1980	115	1985	120
1981	120	1986	130
1982	123	1987	138
1983	125	1988	145
1984	118	1989	150

Answers to Check Your Progress

Check Your Progress 1

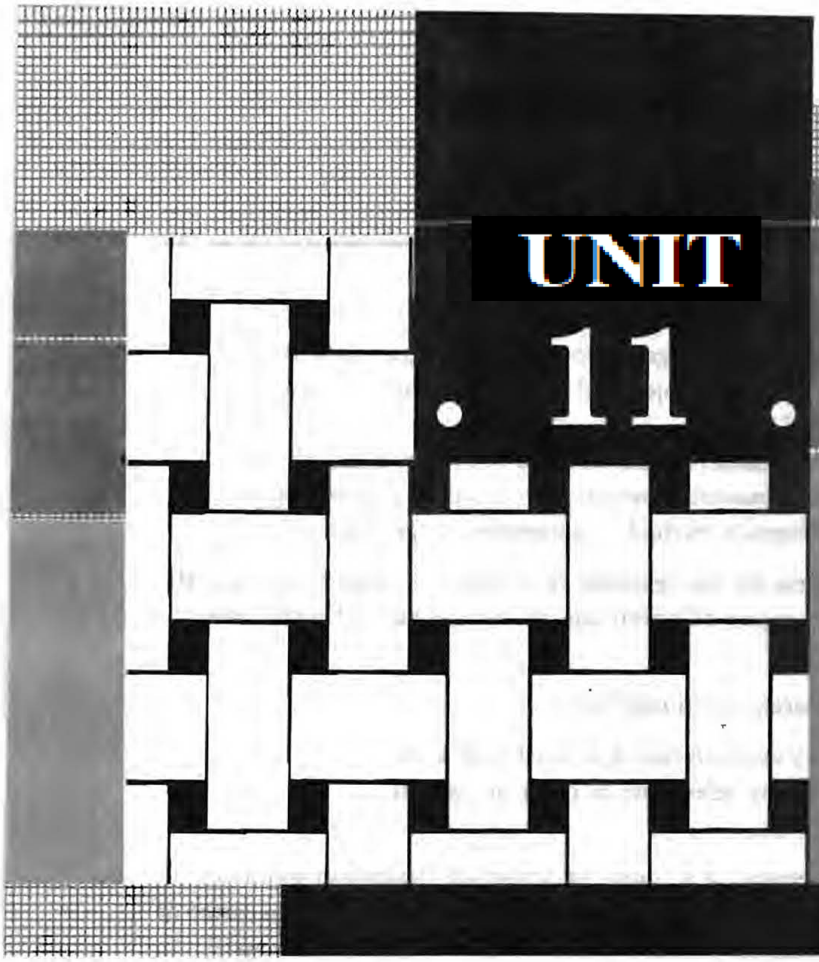
1. Secular Trend
2. $Y_t = a + bt + ct^2$
3. Ratio to Trend Method

Check Your Progress 2

1. Parabolic trend
2. Least squares
3. False

10.8 REFERENCES AND FURTHER READING

- Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2023). Quantitative methods for business (13th ed.). Cengage Learning. ISBN: 9781337909639.
- Newbold, P., Carlson, W. L., & Thorne, B. (2021). Statistics for business and economics (9th ed.). Pearson. ISBN: 9781292315039.
- Vohra, N. D. (2020). Quantitative techniques for management (4th ed.). McGraw-Hill Education. ISBN: 9781259062897.
- Dewhurst, F. (2022). Quantitative methods: An introduction for business management (3rd ed.). Cengage Learning. ISBN: 9781408088481.



UNIT 11

Decision Theories

CHAPTER OUTLINE

11.1 INTRODUCTION

Decision making is needed whenever an individual or an organisation (private or public) is faced with a situation of selecting an optimal (or best in view of certain objectives) course of action from among several available alternatives. For example, an individual may have to decide whether to build a house or to purchase a flat or live in a rented accommodation; whether to join a service or to start own business; which company's car should be purchased, etc. Similarly, a business firm may have to decide the type of technique to be used in production, what is the most appropriate method of advertising its product, etc.

The decision analysis provides certain criteria for the selection of a course of action such that the objective of the decision maker is satisfied. The course of action selected on the basis of such criteria is termed as the optimal course of action.

Every decision problem has four basic features, mentioned below:

1. *Alternative Courses of Action or Acts:* Every decision maker is faced with a set of several alternative courses of action $A_1, A_2, \ldots\ldots A_m$ and he has to select one of them in view of the objectives to be fulfilled.

2. *States of Nature:* The consequences of selection of a course of action are dependent upon certain factors that are beyond the control of the decision maker. These factors are known as states of nature or events. It is assumed that the decision maker is aware of the whole list of events $S_1, S_2, \ldots\ldots S_n$ and exactly one of them is bound to occur. In other words, the events $S_1, S_2, \ldots\ldots S_n$ are assumed to be mutually exclusive and collective exhaustive.

3. *Consequences:* The results or outcomes of selection of a particular course of action are termed as its consequences. The consequence, measured in quantitative or value terms, is called payoff of a course of action. It is assumed that the payoffs of various courses of action are known to the decision maker.

4. *Decision Criterion:* Given the payoffs of various combinations of courses of action and the states of nature, the decision maker has to select an optimal course of action. The criterion for such a selection, however, depends upon the attitude of the decision maker.

If X_{ij} denotes the payoff corresponding to a combination of a course of action and a state of nature, i.e., (A_i, S_j), $i = 1$ to m and $j = 1$ to n, the above elements of a decision problem can be presented in a matrix form, popularly known as the Payoff Matrix.

Payoff Matrix

Events → / *Actions* ↓	S_1	S_2	...	S_j	...	S_n
A_1	X_{11}	X_{12}	...	X_{1j}	...	X_{1n}
A_2	X_{21}	X_{22}	...	X_{2j}	...	X_{2n}
⋮	⋮	⋮		⋮		⋮
A_i	X_{i1}	X_{i2}	...	X_{ij}	...	X_{in}
⋮	⋮	⋮		⋮		⋮
A_m	X_{m1}	X_{m2}	...	X_{mj}	...	X_{mn}

Given the payoff matrix for a decision problem, the process of decision making depends upon the situation under which the decision is being made. These situations can be classified into three broad

categories: (a) Decision making under certainty, (b) Decision making under uncertainty, and (c) Decision making under risk.

The Different Environments in which Decisions are Made

The reason for the existence of a managerial hierarchy, that is, lower, middle and top management, finds itself in different parameters in which an organization operates. There are industry-wide and market-wide decisions that have to be made. Often these decisions must transcend domestic considerations to incorporate international aspects. Such decisions — usually made by top management — occur in a broad-based, complex, ill-defined and non-repetitive problem situation. Middle management usually addresses itself to company-wide problems. It sees to it that the objectives and policies of the organization are properly implemented and that operations are conducted in such a way that optimization may occur.

You may note that while most of the quantitative decision making tools — indeed virtually all of the deterministic tools — were developed to optimize the decision making process, actual managerial practice has sometimes moved away from that objective. The previously mentioned legal or social constraints often at times do not permit optimization and satisfying has been substituted for it. Satisfying refers to the attainment of certain minimum objectives. For example, a company that may have the economic and technological power to smother the competition within its industry but refrains from doing so because of MRTP considerations. Big size per se may be considered in violation of the law or in the international arena, may result in the imposition of quotas.

Lower management is responsible for the conduct of operations — the firing line so to speak — be this in production, marketing, finance or any of the staff functions like personnel or research. This decision environment is usually well-defined and repetitive. Obviously, with reference to a given decision making situation, the distinction between top, middle and lower management may become blurred. In other words, in any on-going business there is always a certain overlapping of the managerial decision making parameters.

The study and analysis of the existence and interaction of these parameters is of great importance to the management systems designer or communication expert. From the quantitative managerial decision making point of view, their importance lies in recognizing their peculiar constraints and then to build the appropriate decision models and to select the best suited quantitative decision tools. A brief discussion of each environment in this light may enhance the understanding of the tools that are discussed later on.

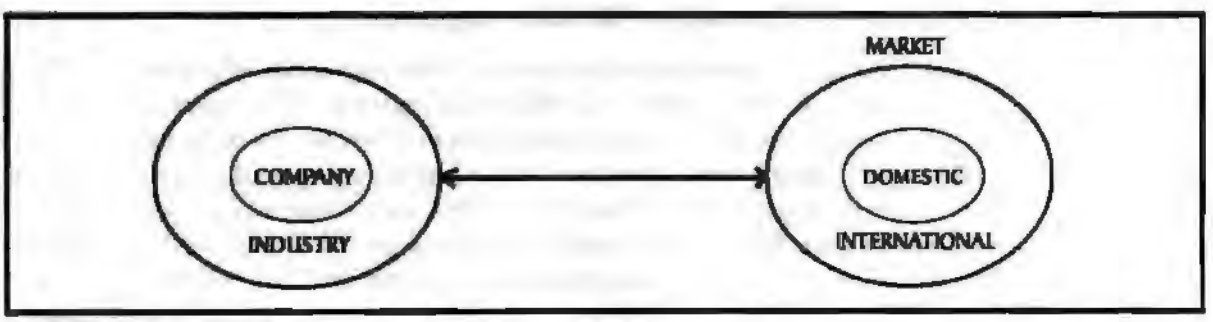


Figure 11.1

Top Management Decision Environment

The top management decision environment is shown in Figure 11.1. The company's approach to the domestic or international market is filtered through industry-wide considerations. What does the market want, what does the competition already supply? Where is our field of attack? Do we have the know-how, do we have the resources? What is the impact of our actions upon the market, our own industry and other industries? These are some of the questions that have to be asked, defined and answered. The problems are unstructured and complex. Thus, often a heuristic decision making process can be utilized to good advantage. Forecasting is of major importance and hence stochastic decision making is widely employed in this uncertain decision environment. But even a deterministic tool — usually intended for decision making situations that assume certainty — input-output analysis, can be effectively used in this environment.

Middle management decisions are primarily company-wide in nature. As mentioned before and shown in Figure 11.2, these decisions steer the organization through its life cycle.

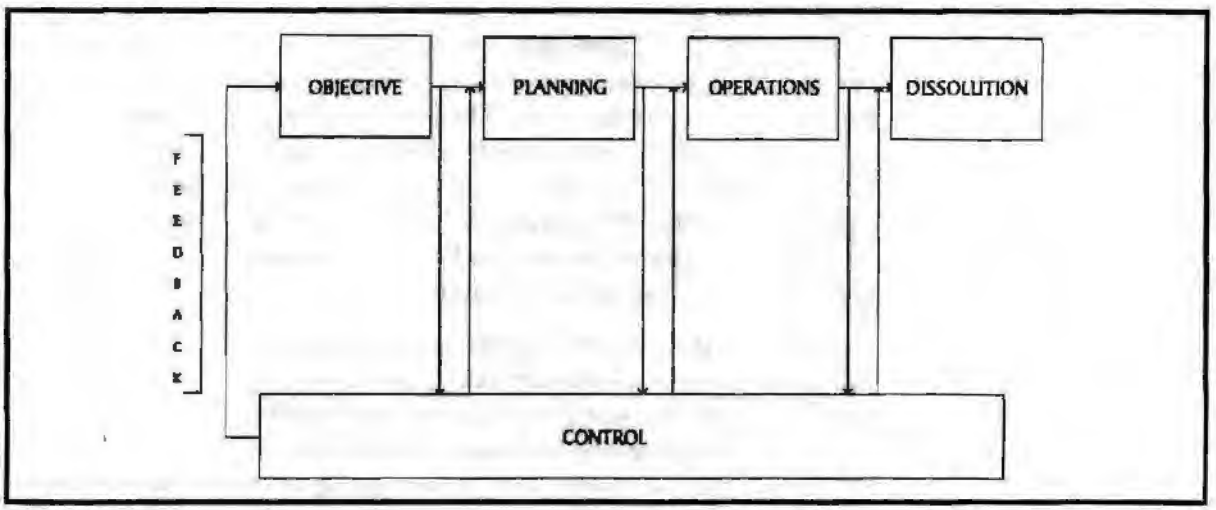


Figure 11.2

Middle Management Decision Environment

Major features of a firm's life are objectives, planning, operation and the ultimate dissolution. The objectives are general and specific in nature. Obviously top management establishes the objectives, but middle management functions as their guardian. Indeed, as Figure11.2 shows, every decision at this level must provide feed-back control for each of the other components.

Planning refers to both policy execution as well as policy development. Scale of production, pricing of the product, product mix, in short the orderly and efficient arrangement of the input factors is to be decided at this point. Making these factors into a product is the job of operations. Some operations have been traditionally called line (financing, production, and marketing) and others staff (personnel, research, etc.); yet, in the quantitative decision systems of the modern firm, such differences are difficult to trace in the decision patterns. Because the same decision making tools are employed. Since the decision environment at this level is somewhat more structured than at the top level but still highly uncertain, stochastic decision tools are frequently employed. In those finance, production and marketing situations that can be well-defined, may be repetitive, deterministic decision tools are found.

It may appear somewhat odd that the decision environment includes attention being paid to the dissolution of the firm. The life cycle concept has been mentioned, and it will be encountered again as

one of the major underlying conceptual aids in forecasting. It is well known that business organisations are born, live and die like natural organisms.

Therefore decision making should always be cognizant of the possibility of dissolution. That moment comes when, to use the vernacular, good money is thrown after bad. While market forces and the application of quantitative analyses normally show the approaching occurrence of that moment — even if the management involved shuts its eyes to the facts or is ignorant about them — at this point the decision is made or superimposed to opt for a turnaround or dissolution. Public agencies unfortunately are rarely subject to such stress producing alternatives.

The lower management decision making environment represents a specialized, narrowly defined area within a company's total decision or operational field. Supervisory personnel of all types are operating in this environment. The decision tasks are normally well defined and repetitive. While the element of uncertainty never leaves the decision environment, here uncertainty can often be programmed into a general or subroutine and stochastic decisions taken as if they were deterministic in nature. A good example is the pricing system of clothing discounters. Merchandise is put on the floor at price A on day one. On, say, day ten the price is automatically reduced to price B and so on until the article is either sold or given to charity after thirty days. This is known as programmed decision making. It should be noted that while the nature of the decision environment remains intact, the decision maker's tasks have been greatly reduced. The complex variables and unstructured decision environment of the merchandising task have been placed first into a model and then into decision making sequence (algorithm). This is the general idea behind model building and the development of algorithms.

Check Your Progress 1

Fill in the blanks

1. Big size per se may be considered in violation of the law or in the international arena, may result in the imposition of.....
2. The ..... decision making environment represents a specialized, narrowly defined area within a company's total decision or operational field.
3. The company's approach to the domestic or international market is filtered through ..... considerations.

11.2 DECISION ANALYSIS

Decision making is needed whenever an individual or an organisation (private or public) is faced with a situation of selecting an optimal (or best in view of certain objectives) course of action from among several available alternatives. For example, an individual may have to decide whether to build a house or to purchase a flat or live in a rented accommodation; whether to join a service or to start own business; which company's car should be purchased, etc. Similarly, a business firm may have to decide the type of technique to be used in production, what is the most appropriate method of advertising its product, etc.

The decision analysis provides certain criteria for the selection of a course of action such that the objective of the decision maker is satisfied. The course of action selected on the basis of such criteria is termed as the optimal course of action. Every decision problem has four basic features, mentioned below:

1. ***Alternative Courses of Action or Acts:*** Every decision maker is faced with a set of several alternative courses of action A_1, A_2, ... A_m and he has to select one of them in view of the objectives to be fulfilled.

2. *States of Nature:* The consequences of selection of a course of action are dependent upon certain factors that are beyond the control of the decision maker. These factors are known as states of nature or events. It is assumed that the decision maker is aware of the whole list of events S_1, S_2, ... S_n and exactly one of them is bound to occur. In other words, the events S_1, S_2, ... S_n are assumed to be mutually exclusive and collective exhaustive.

3. *Consequences:* The results or outcomes of selection of a particular course of action are termed as its consequences. The consequence, measured in quantitative or value terms, is called payoff of a course of action. It is assumed that the payoffs of various courses of action are known to the decision maker.

4. *Decision Criterion:* Given the payoffs of various combinations of courses of action and the states of nature, the decision maker has to select an optimal course of action. The criterion for such a selection, however, depends upon the attitude of the decision maker.

   If X_{ij} denotes the payoff corresponding to a combination of a course of action and a state of nature, i.e., (A_i, S_j), i = 1 to m and j = 1 to n, the above elements of a decision problem can be presented in a matrix form, popularly known as the Payoff Matrix.

Payoff Matrix

Events → *Actions* ↓	S_1	S_2	...	S_j	...	S_n
A_1	X_{11}	X_{12}	...	X_{1j}	...	X_{1n}
A_2	X_{21}	X_{22}	...	X_{2j}	...	X_{2n}
⋮	⋮	⋮		⋮		⋮
A_i	X_{i1}	X_{i2}	...	X_{ij}	...	X_{in}
⋮	⋮	⋮		⋮		⋮
A_m	X_{m1}	X_{m2}	...	X_{mj}	...	X_{mn}

Given the payoff matrix for a decision problem, the process of decision making depends upon the situation under which the decision is being made. These situations can be classified into three broad categories : (a) Decision making under certainty, (b) Decision making under uncertainty and (c) Decision making under risk.

Decision Making under Certainty

The conditions of certainty are very rare particularly when significant decisions are involved. Under conditions of certainty, the decision maker knows which particular state of nature will occur or equivalently, he is aware of the consequences of each course of action with certainty. Under such a situation, the decision maker should focus on the corresponding column in the payoff table and choose a course of action with optimal payoff.

Criteria for Decision Making Under Risk

In case of decision making under uncertainty the probabilities of occurrence of various states of nature are not known. When these probabilities are known or can be estimated, the choice of an optimal action, based on these probabilities, is termed as decision making under risk.

The choice of an optimal action is based on The Bayesian Decision Criterion according to which an action with maximum Expected Monetary Value (EMV) or minimum Expected Opportunity Loss (EOL) or Regret is regarded as optimal.

Example 11.1: The payoffs (in Rs) of three Acts A_1, A_2 and A_3 and the possible states of nature S_1, S_2 and S_3 are given as:

Acts → / States of Nature ↓	A_1	A_2	A_3
S_1	−20	−50	200
S_2	200	−100	−50
S_3	400	600	300

The probabilities of the states of nature are 0.3, 0.4 and 0.3 respectively. Determine the optimal act using the Bayesian Criterion.

Solution:

Computation of Expected Monetary Value

	S_1	S_2	S_3	
$P(S)$	0.3	0.4	0.3	*EMV*
A_1	−20	200	400	$-20\times0.3+200\times0.4+400\times0.3=194$
A_2	−50	−100	600	$-50\times0.3-100\times0.4+600\times0.3=125$
A_3	200	−50	300	$200\times0.3-50\times0.4+300\times0.3=130$

From the above table, we find that the act A_1 is optimal.

The problem can alternatively be attempted by finding minimum EOL, as shown below:

Computation of Expected Opportunity Loss

	S_1	S_2	S_3	
$P(S)$	0.3	0.4	0.3	*EOL*
A_1	220	0	200	$220\times0.3+0\times0.4+200\times0.3=126$
A_2	250	300	0	$250\times0.3+300\times0.4+0\times0.3=195$
A_3	0	250	300	$0\times0.3+250\times0.4+300\times0.3=190$

This indicates that the optimal act is again A_1.

Criteria for Decision Making Under Uncertainty

A situation of uncertainty arises when there can be more than one possible consequences of selecting any course of action. In terms of the payoff matrix, if the decision maker selects A_1, his payoff can be X_{11}, X_{12}, X_{13}, etc., depending upon which state of nature S_1, S_2, S_3, etc., is going to occur. A decision problem, where a decision maker is aware of various possible states of nature but has insufficient information to assign any probabilities of occurrence to them, is termed as decision making under uncertainty.

There are a variety of criteria that have been proposed for the selection of an optimal course of action under the environment of uncertainty. Each of these criteria makes an assumption about the attitude of the decision maker.

1. ***Maximin Criterion:*** This criterion, also known as the criterion of pessimism, is used when the decision maker is pessimistic about future. Maximin implies the maximisation of minimum payoff.

The pessimistic decision maker locates the minimum payoff for each possible course of action. The maximum of these minimum payoffs is identified and the corresponding course of action is selected. This is explained in the following example:

Example 11.2: Let there be a situation in which a decision maker has three possible alternatives A_1, A_2 and A_3, where the outcome of each of them can be affected by the occurrence of any one of the four possible events S_1, S_2, S_3 and S_4. The monetary payoffs of each combination of A_i and S_j are given in the following table:

Payoff Matrix

Events → / Actions ↓	S_1	S_2	S_3	S_4	Min. Payoff	Max. Payoff
A_1	27	12	14	26	12	27
A_2	45	17	35	20	17	45
A_3	52	36	29	15	15	52

Solution: Since 17 is maximum out of the minimum payoffs, the optimal action is A_2.

2. *Maximax Criterion:* This criterion, also known as the criterion of optimism, is used when the decision maker is optimistic about future. Maximax implies the maximisation of maximum payoff. The optimistic decision maker locates the maximum payoff for each possible course of action. The maximum of these payoffs is identified and the corresponding course of action is selected. The optimal course of action in the above example, based on this criterion, is A_3.

3. *Regret Criterion:* This criterion focuses upon the regret that the decision maker might have from selecting a particular course of action. Regret is defined as the difference between the best payoff we could have realised, had we known which state of nature was going to occur and the realised payoff. This difference, which measures the magnitude of the loss incurred by not selecting the best alternative, is also known as opportunity loss or the opportunity cost.

   From the payoff matrix, the payoffs corresponding to the actions A_1, A_2, ... A_n under the state of nature S_j are X_{1j}, X_{2j}, ... X_{nj} respectively. Of these assume that X_{2j} is maximum. Then the regret in selecting A_i, to be denoted by R_{ij} is given by $X_{2j} - X_{ij}$, $i = 1$ to m. We note that the regret in selecting A_2 is zero. The regrets for various actions under different states of nature can also be computed in a similar way.

   The regret criterion is based upon the minimax principle, i.e., the decision maker tries to minimise the maximum regret. Thus, the decision maker selects the maximum regret for each of the actions and out of these the action which corresponds to the minimum regret is regarded as optimal.

   The regret matrix of example can be written as given below:

Regret Matrix

Events → / *Actions ↓*	S_1	S_2	S_3	S_4	*Max. Regret*
A_1	25	24	21	0	25
A_2	7	19	0	6	19
A_3	0	0	6	11	11

From the maximum regret column, we find that the regret corresponding to the course of action is A_3 is minimum. Hence, A_3 is optimal.

4. ***Hurwicz Criterion:*** The maximax and the maximin criteria, discussed above, assumes that the decision maker is either optimistic or pessimistic. A more realistic approach would, however, be to take into account the degree or index of optimism or pessimism of the decision maker in the process of decision making. If a, a constant lying between 0 and 1, denotes the degree of optimism, then the degree of pessimism will be $1 - a$. Then a weighted average of the maximum and minimum payoffs of an action, with a and $1 - a$ as respective weights, is computed. The action with highest average is regarded as optimal.

   We note that a nearer to unity indicates that the decision maker is optimistic while a value nearer to zero indicates that he is pessimistic. If $a = 0.5$, the decision maker is said to be neutralist.

   We apply this criterion to the payoff matrix of example. Assume that the index of optimism $a = 0.7$.

Action	*Max. Payoff*	*Min. Payoff*	*Weighted Average*
A_1	27	12	$27 \times 0.7 + 12 \times 0.3 = 22.5$
A_2	45	17	$45 \times 0.7 + 17 \times 0.3 = 36.6$
A_3	52	15	$52 \times 0.7 + 15 \times 0.3 = 40.9$

   Since the average for A_3 is maximum, it is optimal.

5. ***Laplace Criterion:*** In the absence of any knowledge about the probabilities of occurrence of various states of nature, one possible way out is to assume that all of them are equally likely to occur. Thus, if there are n states of nature, each can be assigned a probability of occurrence = $1/n$. Using these probabilities, we compute the expected payoff for each course of action and the action with maximum expected value is regarded as optimal.

11.3 EXPECTED VALUE CRITERION WITH CONTINUOUSLY DISTRIBUTED RANDOM VARIABLES

Also know as the Expected Value with Perfect Information (EVPI), it is the amount of profit foregone due to uncertain conditions affecting the selection of a course of action.

Given the continuously distributed random variables, a decision maker is supposed to know which particular state of nature will be in effect. Thus, the procedure for the selection of an optimal course of action, for the decision problem given in example, will be as follows :

If the decision maker is certain that the state of nature S_1 will be in effect, he would select the course of action A_3, having maximum payoff equal to Rs. 200.

Similarly, if the decision maker is certain that the state of nature S_2 will be in effect, his course of action would be A_1 and if he is certain that the state of nature S_3 will be in effect, his course of action would be A_2. The maximum payoffs associated with the actions are Rs. 200 and Rs 600 respectively.

The weighted average of these payoffs with weights equal to the probabilities of respective states of nature is termed as Expected Payoff under Certainty (EPC).

Thus, $EPC = 200 \times 0.3 + 200 \times 0.4 + 600 \times 0.3 = 320$

The difference between EPC and EMV of optimal action is the amount of profit foregone due to uncertainty and is equal to EVPI.

Thus, EVPI = EPC – EMV of optimal action = 320 – 194 = 126

It is interesting to note that EVPI is also equal to EOL of the optimal action.

Cost of Uncertainty

This concept is similar to the concept of EVPI. Cost of uncertainty is the difference between the EOL of optimal action and the EOL under perfect information.

Given the perfect information, the decision maker would select an action with minimum opportunity loss under each state of nature. Since minimum opportunity loss under each state of nature is zero, therefore,

EOL under certainty $= 0 \times 0.3 + 0 \times 0.4 + 0 \times 0.3 = 0$

Thus, the cost of uncertainty = EOL of optimal action = EVPI

Example 11.3: A group of students raise money each year by selling souvenirs outside the stadium of a cricket match between teams *A* and *B*. They can buy any of three different types of souvenirs from a supplier. Their sales are mostly dependent on which team wins the match. A conditional payoff (in Rs.) table is as under:

Type of Souvenir →	**I**	**II**	**III**
Team A wins	1200	800	300
Team B wins	250	700	1100

(i) Construct the opportunity loss table.

(ii) Which type of souvenir xshould the students buy if the probability of team A's winning is 0.6?

(iii) Compute the cost of uncertainty.

Solution:

(i) The Opportunity Loss Table

Actions → / *Events ↓*	*Type of Souvenir bought*		
	I	*II*	*III*
Team A wins	0	400	900
Team B wins	850	400	0

(ii) EOL of buying type I Souvenir $= 0 \times 0.6 + 850 \times 0.4 = 340$

EOL of buying type II Souvenir $= 400 \times 0.6 + 400 \times 0.4 = 400.$

EOL of buying type III Souvenir $= 900 \times 0.6 + 0 \times 0.4 = 540.$

Since the EOL of buying Type I Souvenir is minimum, the optimal decision is to buy Type I Souvenir.

(iii) Cost of uncertainty = EOL of optimal action = Rs. 340

Example 11.4: The following is the information concerning a product X:

(i) Per unit profit is Rs 3.

(ii) Salvage loss per unit is Rs 2.

(iii) Demand recorded over 300 days is as under :

Units demanded	:	5	6	7	8	9
No. of days	:	30	60	90	75	45

Find: (i) EMV of optimal order.

(ii) Expected profit presuming certainty of demand.

Solution:

(i) The given data can be rewritten in terms of relative frequencies, as shown below:

Units demanded	:	5	6	7	8	9
No. of days	:	0.1	0.2	0.3	0.25	0.15

From the above probability distribution, it is obvious that the optimum order would lie between and including 5 to 9.

Let A denote the number of units ordered and D denote the number of units demanded per day.

If $D \geq A$, profit per day = $3A$, and if $D < A$, profit per day = $3D - 2(A - D) = 5D - 2A$.

Thus, the profit matrix can be written as

Units Demanded	5	6	7	8	9	
Probability → Action (units ordered) ↓	0.10	0.20	0.30	0.25	0.15	EMV
5	15	15	15	15	15	15.00
6	13	18	18	18	18	17.50
7	11	16	21	21	21	19.00
8	9	14	19	24	24	19.00
9	7	12	17	22	27	17.75

From the above table, we note that the maximum EMV = 19.00, which corresponds to the order of 7 or 8 units. Since the order of the 8th unit adds nothing to the EMV, i.e., marginal EMV is zero, therefore, order of 8 units per day is optimal.

(ii) Expected profit under certainty

$= (5 \times 0.10 + 6 \times 0.20 + 7 \times 0.30 + 8 \times 0.25 + 9 \times 0.15) \times 3 =$ Rs 21.45

Alternative Method

The work of computations of EMV's, in the above example, can be reduced considerably by the use of the concept of expected marginal profit. Let π be the marginal profit and λ be the marginal loss of ordering an additional unit of the product. Then, the expected marginal profit of ordering the Ath unit, is given by

$$= \pi.P(D \geq A) - \lambda.P(D < A) = \pi.P(D \geq A) - \lambda.[1 - P(D \geq A)]$$

$$= (\pi + \lambda).P(D \geq A) - \lambda \quad \ldots. (1)$$

The computations of EMV, for alternative possible values of A, are shown in the following table:

In our example, $\pi = 3$ and $\lambda = 2$.

Thus, the expression for the expected marginal profit of the Ath unit

$$= (3+2)P(D \geq A) - 2 = 5P(D \geq A) - 2.$$

Table for computations

Action(A)	$P(D \geq A)^*$	$EMP = 5P(D \geq A) - 2$	*Total profit or EMV*
5	1.00	$5 \times 1.00 - 2 = 3.00$	$5 \times 3.00 = 15.00$
6	0.90	$5 \times 0.90 - 2 = 2.50$	$15.00 + 2.50 = 17.50$
7	0.70	$5 \times 0.70 - 2 = 1.50$	$17.50 + 1.50 = 19.00$
8	0.40	$5 \times 0.40 - 2 = 0.00$	$19.00 + 0.00 = 19.00$
9	0.15	$5 \times 0.15 - 2 = -1.25$	$19.00 - 1.25 = 17.75$

* This column represents the 'more than type' cumulative probabilities.

Since the Expected Marginal Profit (EMP) of the 8th unit is zero, therefore, optimal order is 8 units.

Marginal Analysis

Marginal analysis is used when the number of states of nature is considerably large. Using this analysis, it is possible to locate the optimal course of action without the computation of EMV's of various actions.

An order of A units is said to be optimal if the expected marginal profit of the Ath unit is non-negative and the expected marginal profit of the $(A + 1)$th unit is negative. Using equation (1), we can write

$$(\pi + \lambda)P(D \geq A) - \lambda \geq 0 \text{ and} \quad \ldots. (2)$$

$$(\pi + \lambda)P(D \geq A+1) - \lambda < 0 \quad \ldots. (3)$$

From equation (2), we get

$$P(D \geq A) \geq \frac{\lambda}{\pi + \lambda} \text{ or } 1 - P(D < A) \geq \frac{\lambda}{\pi + \lambda}$$

$$\text{or } P(D<A)\le 1-\frac{\lambda}{\pi+\lambda} \text{ or } P(D\le A-1)\le \frac{\pi}{\pi+\lambda} \quad \ldots (4)$$

[$P(D \le A - 1) = P(D < A)$, since A is an integer]

Further, equation (3) gives

$$P(D\ge A+1)<\frac{\lambda}{\pi+\lambda} \text{ or } 1-P(D<A+1)<\frac{\lambda}{\pi+\lambda}$$

$$\text{or } P(D<A+1)>1-\frac{\lambda}{\pi+\lambda} \text{ or } P(D\le A)>\frac{\pi}{\pi+\lambda} \quad \ldots (5)$$

Combining (4) and (5), we get

$$P(D\le A-1)\le\frac{\pi}{\pi+\lambda}<P(D\le A).$$

Writing the probability distribution, given in example, in the form of less than type cumulative probabilities which is also known as the distribution function $F(D)$, we get

Units demanded(D)	:	5	6	7	8	9
$F(D)$	:	0.1	0.3	0.6	0.85	1.00

We are given $\pi = 3$ and $\lambda = 2$, $\therefore \quad \frac{\pi}{\pi+\lambda}=\frac{3}{5}=0.6$

Since the next cumulative probability, i.e., 0.85 is 8 units. , corresponds to 8 units, hence, the optimal order

11.4 DECISION TREE ANALYSIS

The decision tree diagrams are often used to understand and solve a decision problem. Using such diagrams, it is possible to describe the sequence of actions and chance events.

A decision node is represented by a square and various action branches stem from it. Similarly, a chance node is represented by a circle and various event branches stem from it. Various steps in the construction of a decision tree can be summarized as follows:

(i) Show the appropriate action-event sequence beginning from left to right of the page.

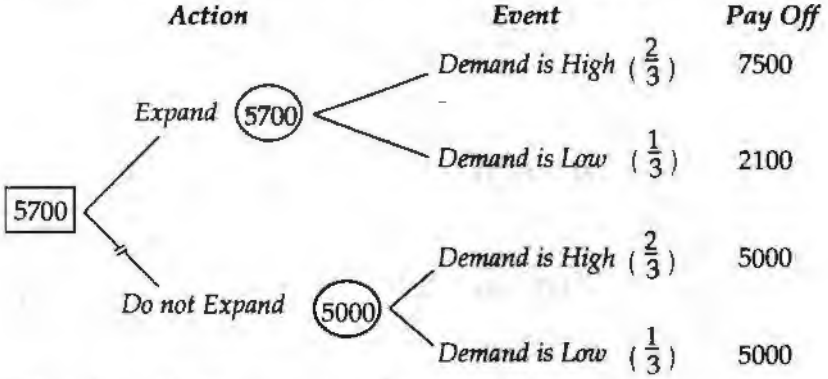
(ii) Write the probabilities of various events along their respective branches stemming from each chance node.

(iii) Write the payoffs at the end of each of the right-most branch.

(iv) Moving backward, from right to left, compute EMV of each chance node, wherever encountered. Enter this EMV in the chance node. When a decision node is encountered, choose the action branch having the highest EMV. Enter this EMV in the decision node and cutoff the other action branches.

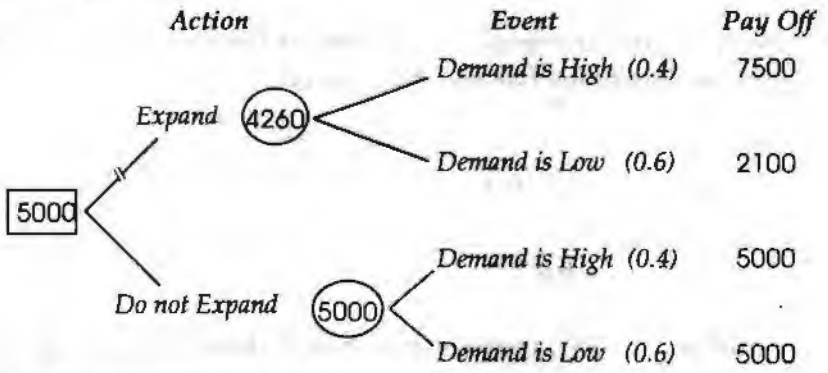
Following this approach, we can describe the decision problem of the above example as given:

Case I: When the survey predicts that the demand is going to be high



Thus, the optimal act to expand capacity.

Case II: In the absence of survey



Thus, the optimal act is not to expand capacity.

Graphic Displays of the Decision Making Process

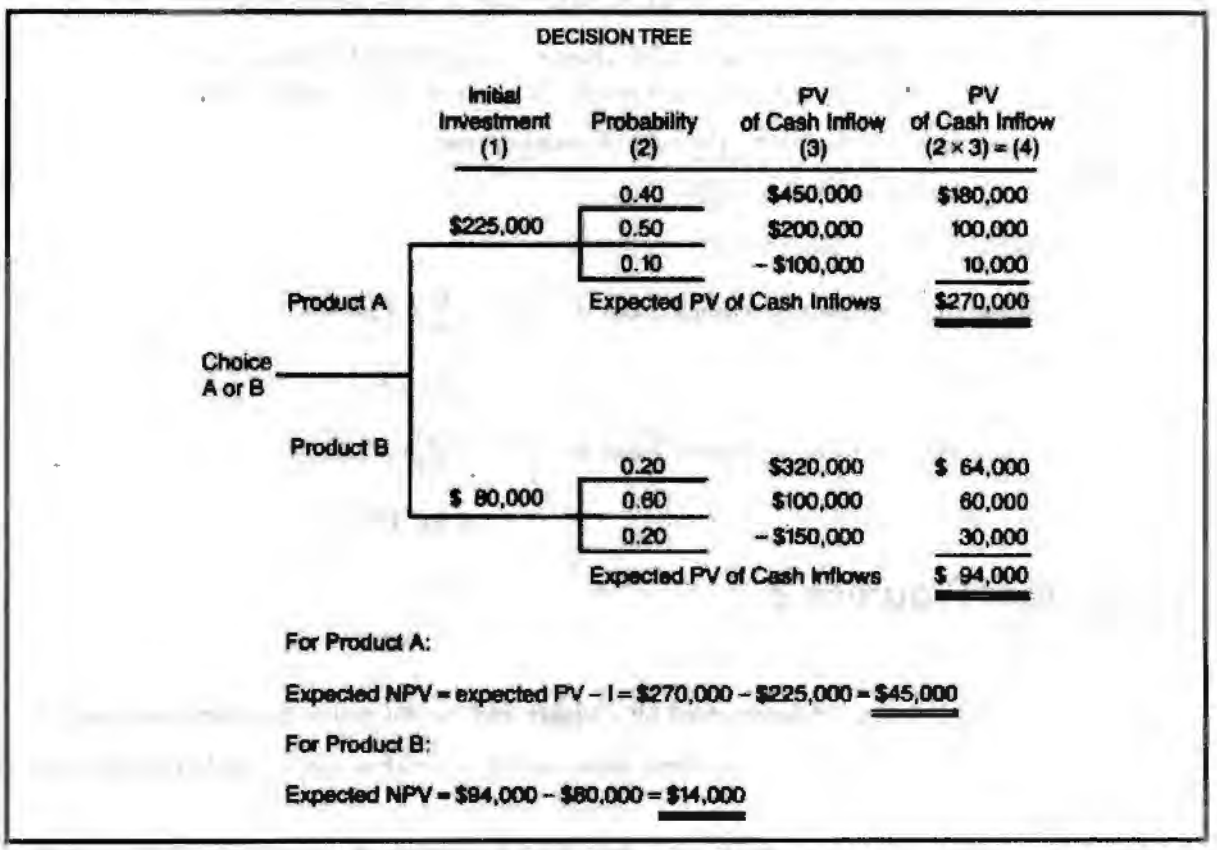
Pictorial representation of a decision situation, normally found in discussions of decision-making under uncertainty or risk. It shows decision alternatives, states of nature, probabilities attached to the state of nature, and conditional benefits and losses.

The tree approach is most useful in a sequential decision situation.

Example 11.5: Assume XYZ Corporation wishes to introduce one of two products to the market this year. The probabilities and Present Values (PV) of projected cash inflows follow:

Products	Initial Investment	PV of Cash Inflows	Probabilities
A	$225,000		1.00
		$450,000	0.40
		200,000	0.50
		– 100,000	0.10
B	80,000		1.00
		320,000	0.20
		100,000	0.60
		– 150,000	0.20

A decision tree analyzing the two products follows:



Based on the expected net present value, the company should choose product A over product B.

Constructing the Decision Tree

Start a decision tree with a decision that needs to be made. This decision is represented by a small square towards the left of a large piece of paper. From this box draw out lines towards the right for each possible solution, and write that solution along the line. Keep the lines apart as far as possible so that you can expand your thoughts.

At the end of each solution line, consider the results. If the result of taking that decision is uncertain, draw a small circle. If the result is another decision that needs to be made, draw another square. Squares represent decisions; circles represent uncertainty or random factors. Write the decision or factor to be considered above the square or circle. If you have completed the solution at the end of the line, just leave it blank.

Starting from the new decision squares on your diagram, draw out lines representing the options that could be taken. From the circles draw out lines representing possible outcomes. Again mark a brief note on the line saying what it means. Keep on doing this until you have drawn down as many of the possible outcomes and decisions as you can see leading on from your original decision.

Example 11.6: A private investment firm has Rs. 10 crores available in cash. It can invest the money in a bank at 10% yielding a return of Rs. 15 crore over five years (ignore compound interest).

Alternatively it can invest in mutual funds, of which there are currently two available.

If it invests in Mutual Fund A there is a 0.5 chance of it being a success yielding Rs. 20 crore, and a 0.5 chance of it failing leading to a loss of Rs. 5 crore. (over the five year period)

If it invests in Mutual Fund B there is a 0.6 chance of the project being a success yielding Rs. 30 crore and a 0.4 chance of it failing leading to a loss of Rs. 2 crore. (over the five year period)

Show the most feasible solution by the help of decision tree.

Solution: Working out the likely outcomes:

Invest in bank – return = Rs. 15 cr

Expected Value of investment in Mutual Fund $A = E(X) = \sum_j x_j P(X = x_j)$

= Rs 7.5 cr

Expected Value of investment in Mutual Fund $B = E(x) = \sum_j x_j P(X = x_j)$

= Rs. 17.2 cr

Check Your Progress 2

Fill in the Blanks:

1. A decision ..... is represented by a square and various action branches stem from it.
2. A situation of ..... arises when there can be more than one possible consequences of selecting any course of action.
3. By using ..... it is possible to describe the sequence of actions and chance events.

11.5 SUMMARY

Decision making is needed whenever an individual or an organization is faced with a situation of selecting an optimal course of action from among several available alternatives. The decision analysis provides certain criteria for the selection of a course of action such that the objective of the decision maker is satisfied. The reason for the existence of a managerial hierarchy, that is, lower, middle and top

management, finds itself in different parameters in which an organization operates. The study and analysis of the existence and interaction of these parameters is of great importance to the management systems designer or communication expert. The problems are unstructured and complex. Thus, often a heuristic decision making process can be utilized to good advantage. Forecasting is of major importance and hence stochastic decision making is widely employed in this uncertain decision environment. Decision making should always be cognizant of the possibility of dissolution.

The choice of an optimal action is based on The Bayesian Decision Criterion according to which an action with maximum Expected Monetary Value (EMV) or minimum Expected Opportunity Loss (EOL) or Regret is regarded as optimal. A situation of uncertainty arises when there can be more than one possible consequences of selecting any course of action. The decision tree diagrams are often used to understand and solve a decision problem. The tree approach is most useful in a sequential decision situation.

11.6 KEYWORDS

- Expected Value
- Conditions
- Inertia
- Gambling
- Falsification
- Actions
- Outcomes
- Acquiescence
- Semantics

11.7 REVIEW QUESTIONS

1. Mention the four basic features decision problem.
2. What are the different environments in which decisions are made?
3. Explain the different criteria for decision making under uncertainty.
4. A shopkeeper at a local stadium must determine whether to sell ice cream or coffee at today's game. The shopkeeper believes that the profit will depend upon the weather.

   Based upon his past experience at this time of the year, the shopkeeper estimates the probability of warm weather as 0.60. Prior to making his decision, the shopkeeper decides to hear forecast of the local weatherman. In the past, when it has been cool, the weatherman has forecast cool weather 80% times. When it has been warm, the weatherman has forecast warm weather 70% times. If today's forecast is for cool weather, using Bayesian decision theory and EMV criterion, determine whether the shopkeeper should sell ice cream or coffee?
5. A producer of boats has estimated the following distribution of demand for a particular kind of boat:

   Each boat costs him Rs. 7,000 and he sells them for Rs. 10,000 each. Any boats that are left unsold at the end of the season must be disposed off for Rs. 6,000 each. How many boats should be kept in stock to maximize his expected profit?
6. Consider the decision problem with the profit payoff table with four decision alternatives and three states of nature.

	States of nature		
Alternative	X	Y	Z
A	4	3	3
B	5	2	5
C	5	6	2
D	6	1	4

(a) If the decision maker knows nothing about the probabilities of the three states of nature, what is the recommended decision using the minimax rule?

(b) Assume that the payoff table provides cost rather than profit payoffs. What is the recommended decision now?

7. Consider the decision problem with the payoff table with four decision alternatives and three states of nature.

	States of nature		
Alternative	X	Y	Z
A	4	3	3
B	5	2	5
C	5	6	2
D	6	1	4

(a) Draw a decision tree for the above payoff table.

(b) If the probability of the states of nature is 0.3, 0.4 and 0.3 for the states X, Y and Z respectively, find out the optimum decision.

(c) Assume that the payoff table provides cost rather than profit payoffs. What is the recommended decision now if the probability given in the option (b) still holds?

8. Which action would an optimal Decision Theoretic Agent take in the following situation?

Utility of Resulting State	Probability	
Action 1	10	0.2
Action 2	10000	0.001
Action 3	5	0.799

9. After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and that the test is 99% accurate (i.e. the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given you don't have the disease). The good news is that this is a rare disease, striking only 1 in 10,000 people. Why is it good news that the disease is rare? What are the chances that you actually have the disease?

10. John's monthly consumption:

- $X_1 \leq 4000$ if he does not get ill
- $X_2 \leq 500$ if he gets ill (so he cannot work)
- Probability of illness 0.25
- Consequently, probability of no illness = 1 – 0.25 = 0.75

What will be the expected value?

Answers to Check Your Progress

Check Your Progress 1

1. Quotas
2. Lower management
3. Industry-wide

Check Your Progress 2

1. Node
2. Uncertainty
3. Decision tree diagrams

11.8 REFERENCES AND FURTHER READING

- Vohra, N. D. (2020). Quantitative techniques for management (4th ed.). McGraw-Hill Education. ISBN: 9781259062897.
- Dewhurst, F. (2022). Quantitative methods: An introduction for business management (3rd ed.). Cengage Learning. ISBN: 9781408088481
- Johnson, R. A., & Wichern, D. W. (2019). Applied multivariate statistical analysis (6th ed.). Pearson. ISBN: 9780134995381.

UNIT

12

Linear Programming, Transportation and Assignment Problems

CHAPTER OUTLINE

12.1 INTRODUCTION

Linear programming is a widely used mathematical modeling technique to determine the optimum allocation of scarce resources among competing demands. Resources typically include raw materials, manpower, machinery, time, money and space. The technique is very powerful and found especially useful because of its application to many different types of real business problems in areas like finance, production, sales and distribution, personnel, marketing and many more areas of management. As its name implies, the linear programming model consists of linear objectives and linear constraints, which means that the variables in a model have a proportionate relationship. For example, an increase in manpower resource will result in an increase in work output.

Essentials of Linear Programming Model

For a given problem situation, there are certain essential conditions that need to be solved by using linear programming.

3. ***Linearity:*** increase in labour input will have a proportionate increase in output.
4. ***Homogeneity :*** the products, workers' efficiency, and machines are assumed to be identical.
5. ***Divisibility :*** it is assumed that resources and products can be divided into fractions. (in case the fractions are not possible, like production of one-third of a computer, a modification of linear programming called integer programming can be used).

Properties of Linear Programming Model

The following properties form the linear programming model:

1. Relationship among decision variables must be linear in nature.
2. A model must have an objective function.
3. Resource constraints are essential.
4. A model must have a non-negativity constraint.

12.2 FORMULATION OF LINEAR PROGRAMMING PROBLEM

Formulation of Linear Programming Problem (LPP) is the representation of problem situation in a mathematical form. It involves well defined decision variables, with an objective function and set of constraints.

Objective Function

The objective of the problem is identified and converted into a suitable objective function. The objective function represents the aim or goal of the system (i.e., decision variables) which has to be determined from the problem. Generally, the objective in most cases will be either to maximize resources or profits or, to minimize the cost or time.

Limited resources: limited number of labour, material equipment and finance.

For example, assume that a furniture manufacturer produces tables and chairs. If the manufacturer wants to maximize his profits, he has to determine the optimal quantity of tables and chairs to be produced.

Let x_1 = Optimal production of tables

p_1 = Profit from each table sold

x_2 = Optimal production of chairs

p_2 = Profit from each chair sold.

Hence, Total profit from tables = $p_1 x_1$

Total profit from chairs = $p_2 x_2$

The objective function is formulated as below,

Maximize Z or $Z\max = p_1 x_1 + p_2 x_2$

Constraints

When the availability of resources are in surplus, there will be no problem in making decisions. But in real life, organizations normally have scarce resources within which the job has to be performed in the most effective way. Therefore, problem situations are within confined limits in which the optimal solution to the problem must be found.

Considering the previous example of furniture manufacturer, let w be the amount of wood available to produce tables and chairs. Each unit of table consumes w_1 unit of wood and each unit of chair consumes w_2 units of wood.

For the constraint of raw material availability, the mathematical expression is,

$$w_1 x_1 + w_2 x_2 \leq w$$

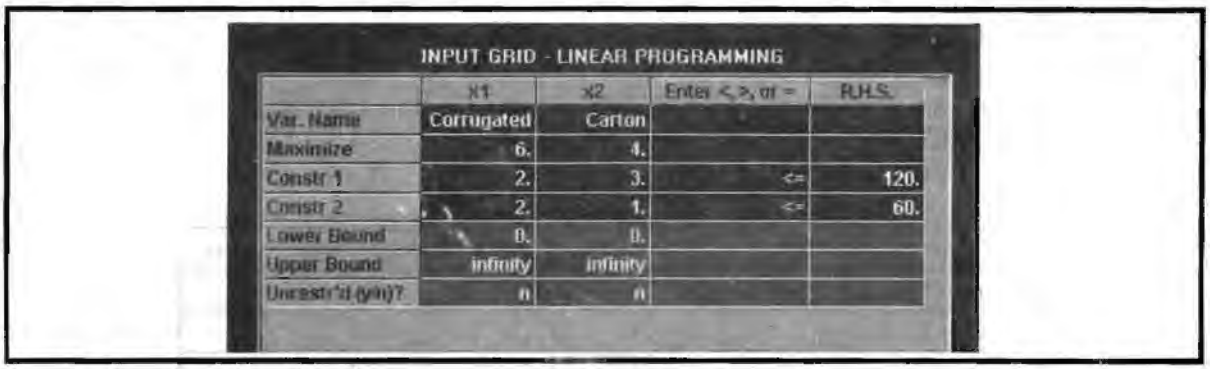
In addition to raw material, if other resources such as labor, machinery and time are also considered as constraint equations.

Non-negativity Constraint

Negative values of physical quantities are impossible, like producing negative number of chairs, tables, etc., so it is necessary to include the element of non-negativity as a constraint i.e., $x_1, x_2 \geq 0$.

Solving Linear Programming Graphically Using Computer

The above problem is solved using computer with the help of TORA. Open the TORA package and select LINEAR PROGRAMMING option. Then press Go to Input and enter the input data as given in the input screen shown in Figure 12.1.



INPUT GRID - LINEAR PROGRAMMING

	x1	x2	Enter <, >, or =	RHS
Var. Name	Corrugated	Carton		
Maximize	6.	4.		
Constr 1	2.	3.	<=	120.
Constr 2	2.	1.	<=	60.
Lower Bound	0.	0.		
Upper Bound	infinity	infinity		
Unrestr'd (y/n)?	n	n		

Figure 12.1

Linear Programming, TORA Package (Input Screen)

Now, go to Solve Menu and click Graphical in the 'solve problem' options. Then click Graphical, and then press Go to Output . The output screen is displayed with the graph grid on the right hand side and equations in the left hand side. To plot the graphs one by one, click the first constraint equation. Now the line for the first constraint is drawn connecting the points (40, 60). Now, click the second equation to draw the second line on the graph. You can notice that a portion of the graph is cut while the second constraint is also taken into consideration. This means the feasible area is reduced further. Click on the objective function equation. The objective function line locates the furthermost point (maximization) in the feasible area which is (15,30) shown in Figure 12.2.

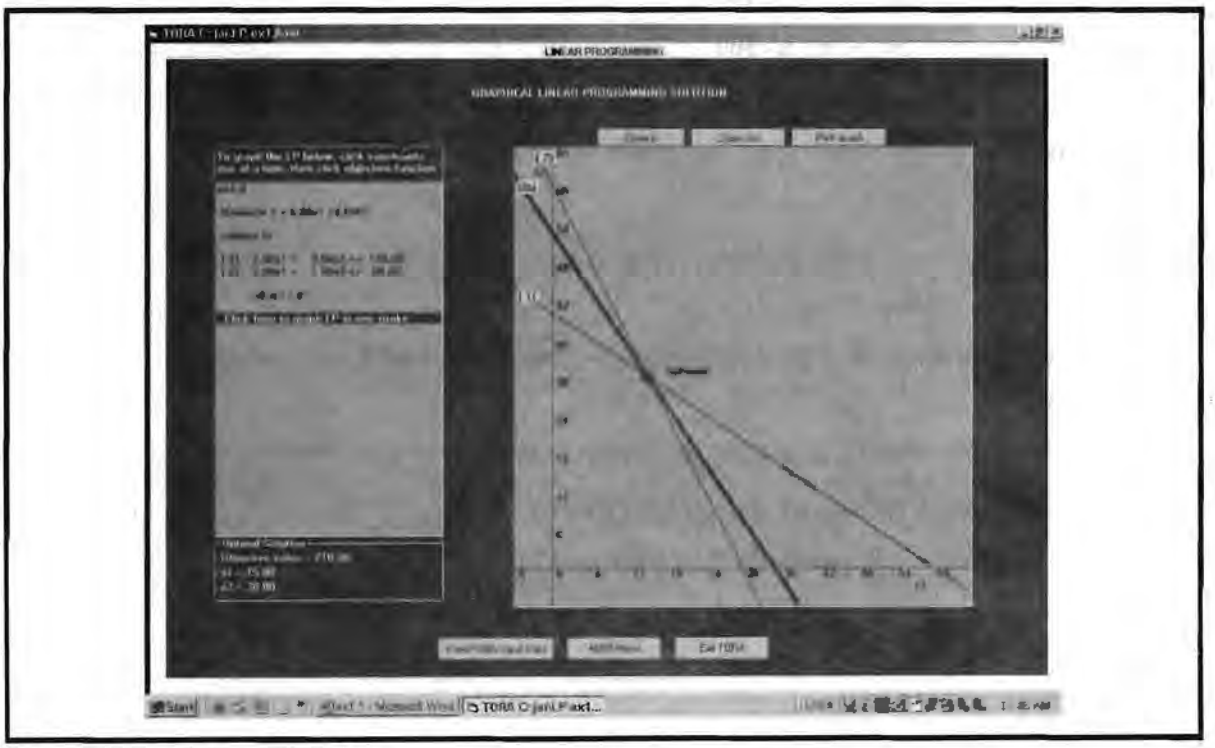


Figure 12.2

Graph Showing Feasible Area

Example 12.1: A soft drink manufacturing company has 300 ml and 150 ml canned cola as its products with profit margin of Rs. 4 and Rs. 2 per unit respectively. Both the products have to undergo process in three types of machine. The following Table 12.1, indicates the time required on each machine and the available machine-hours per week.

Table 12.1: Available Data

Requirement	Cola 300 ml	Cola 150 ml	Available machine-hours per week
Machine 1	3	2	300
Machine 2	2	4	480
Machine 3	5	7	560

Formulate the linear programming problem specifying the product mix which will maximize the profits within the limited resources. Also solve the problem using computer.

Solution: Let x_1 be the number of units of 300 ml cola and x_2 be the number of units of 150 ml cola to be produced respectively. Formulating the given problem, we get

Objective function:

$$Z_{max} = 4x_1 + 2x_2$$

Subject to constraints,

$$3x_1 + 2x_2 \leq 300 \quad \text{.....(i)}$$

$$2x_1 + 4x_2 \leq 480 \quad \text{.....(ii)}$$

$$5x_1 + 7x_2 \leq 560 \quad \text{.....(iii)}$$

where $x_1, x_2 \geq 0$

The inequalities are removed to give the following equations:

$$3x_1 + 2x_2 = 300 \quad \text{.....(iv)}$$

$$2x_1 + 4x_2 = 480 \quad \text{.....(v)}$$

$$5x_1 + 7x_2 = 560 \quad \text{.....(vi)}$$

Find the co-ordinates of lines by substituting $x_1 = 0$ to find x_2 and $x_2 = 0$ to find x_1.

Therefore,

Line $3x_2 + 2x_2 = 300$ passes through (0,150),(100,0)

Line $2x_1 + 4x_2 = 480$ passes through (0,120),(240,0)

Line $5x_1 + 7x_2 = 650$ passes through (0,80),(112,0)

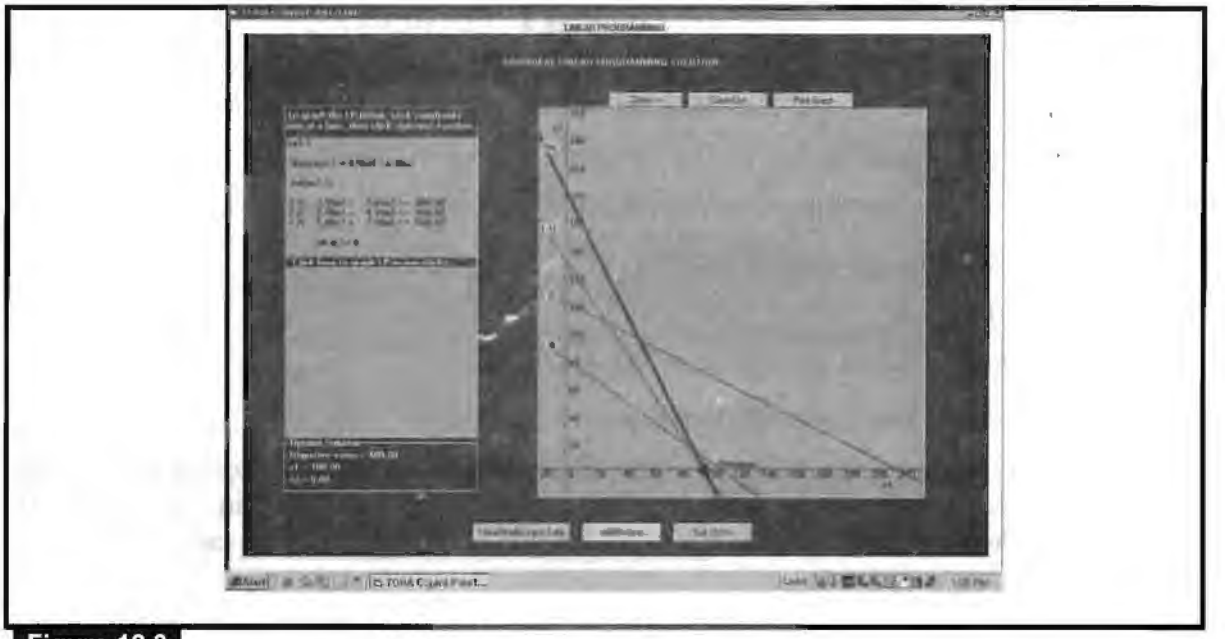


Figure 12.3

Graphical Presentation of lines
(TORA, Output Screen)

For objective function,

The Line $4x_1 + 2x_2 = 0$ passes through (–10, 20), (10, –20)

Plot the lines on the graph as shown in the computer output Figure 12.3.

The objective is to maximize the profit. Move the objective function line away from the origin by drawing parallel lines. The line that touches the furthermost point of the feasible area is (100, 0). Therefore, the values of x_1 and x_2 are 100 and 0 respectively.

Maximum Profit, $Z_{max} = 4x_1 + 2x_2$

$= 4(100) + 2(0)$

$=$ Rs. 400.00

Example 12.2: Solve the following LPP by graphical method.

Minimize $Z = 18x_1 + 12x_2$

Subject to constraints,

$2x_1 + 4x_2 \leq 60$ .....(i)

$3x_1 + x_2 \geq 30$ .....(ii)

$8x_1 + 4x_2 \geq 120$ .....(iii)

where $x_1, x_2 \geq 0$

Solution: The inequality constraints are removed to give the equations,

$2x_1 + 4x_2 = 60$ .....(iv)

$3x_1 + x_2 = 30$ .....(v)

$$8x_1 + 4x_2 = 120 \qquad \text{.....(vi)}$$

The equation lines pass through the co-ordinates as follows:

For constraints,

$2x_1 + 4x_2 = 60$ passes through (0,15), (30,0).

$3x_1 + x_2 = 30$ passes through (0,30), (10,0).

$8x_1 + 4x_2 = 120$ passes through (0,30), (15,0).

The objective function,

$18x_1 + 12x_2 = 0$ passes through (–10, 15), (10, –15).

Plot the lines on the graph as shown in Figure 12.4

Here the objective is minimization. Move the objective function line and locate a point in the feasible region which is nearest to the origin, i.e., the shortest distance from the origin. Locate the point P, which lies on the x – axis. The co-ordinates of the point P are (15, 0) or $x_1 = 15$ and $x_2 = 0$.

The minimum value of Z

$$\begin{aligned} Z_{min} &= 18\ x_1 + 12x_2 \\ &= 18\ (15) + 12\ (0) \\ &= \text{Rs. } 270.00 \end{aligned}$$

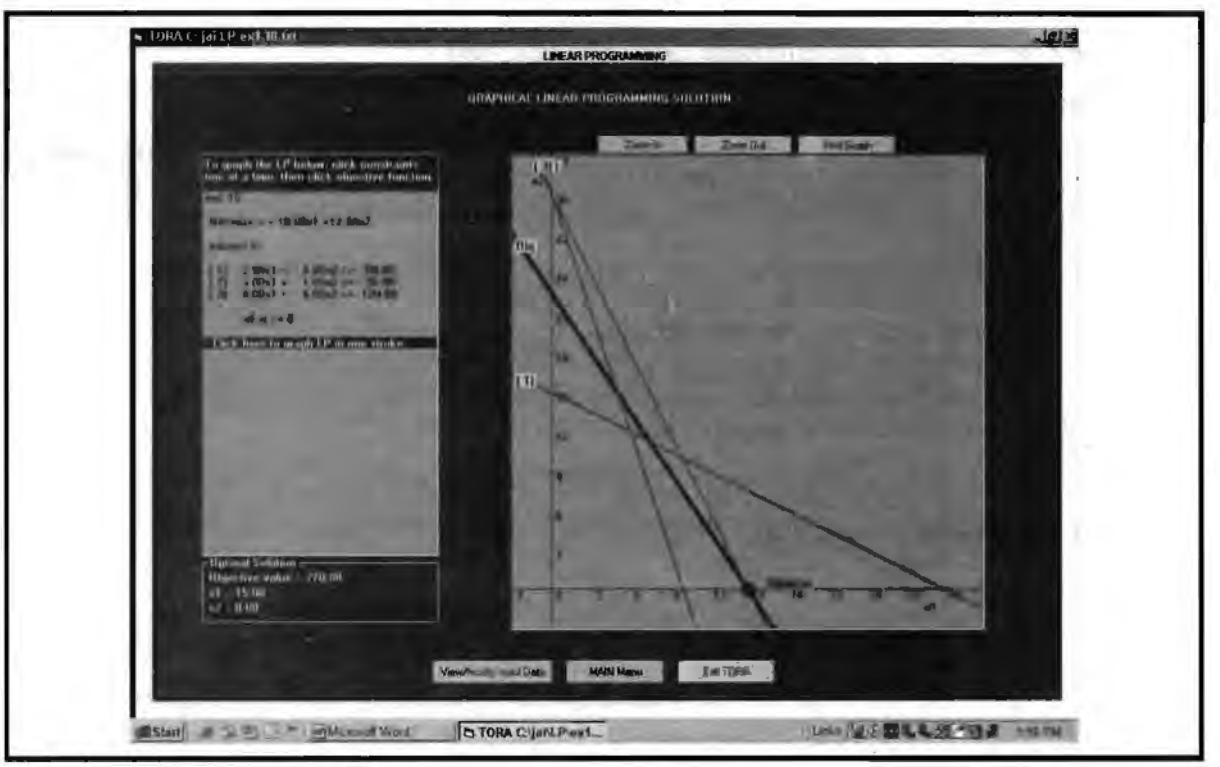


Figure 12.4

Graphical Presentation (Output Screen, TORA)

12.3 SUMMARY OF GRAPHICAL METHOD

Step 1: Convert the inequality constraint as equations and find co-ordinates of the line.

Step 2: Plot the lines on the graph.

(*Note:* If the constraint is ≥ type, then the solution zone lies away from the centre. If the constraint is ≤ type, then solution zone is towards the centre.)

Step 3: Obtain the feasible zone.

Step 4: Find the co-ordinates of the objectives function (profit line) and plot it on the graph representing it with a dotted line.

Step 5: Locate the solution point.

(*Note:* If the given problem is maximization, z_{max} then locate the solution point at the far most point of the feasible zone from the origin and if minimization, Z_{min} then locate the solution at the shortest point of the solution zone from the origin).

Step 6: Solution type

(i) If the solution point is a single point on the line, take the corresponding values of x_1 and x_2.

(ii) If the solution point lies at the intersection of two equations, then solve for x_1 and x_2 using the two equations.

(iii) If the solution appears as a small line, then a multiple solution exists.

(iv) If the solution has no confined boundary, the solution is said to be an unbound solution.

Example 12.3: Solve the formulated LP model graphically using computer.

$$Z_{max} = 7x_1 + 5x_2$$

Subject to constraints,

$$8x_1 + 4x_2 \leq 20 \quad \text{.....(i)}$$

$$2x_1 + 3x_2 \leq 8 \quad \text{.....(ii)}$$

$$-x_1 + x_2 \leq 2 \quad \text{.....(iii)}$$

$$x_2 \leq 2 \quad \text{.....(iv)}$$

where $x_1, x_2 \geq 0$

Solution: The input values of the problem are given to obtain the output screen as shown in Figure 12.5.

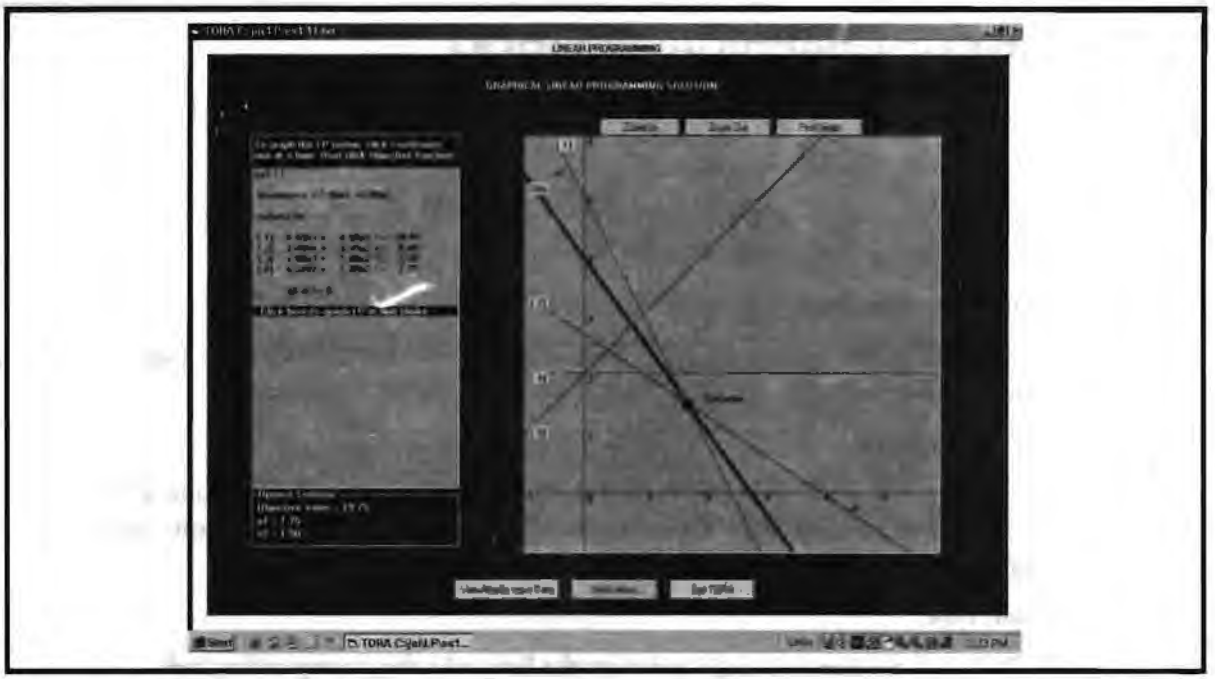


Figure 12.5

Graphical Presentation (Output Screen, TORA)

Results:

Perfumes to be produced, x_1 = 1.75 litres or 17.5 say 18 bottles of 100 ml each

Body sprays to be produced, x_2 = 1.50 litres or 15 bottles of 100 ml each

Maximum profit, Z_{max} = Rs. 19.75

General Linear Programming Model

A general representation of LP model is given as follows:

Maximize or Minimize, $Z = p_1 x_1 + p_2 x_2 \ldots\ldots\ldots p_n x_n$

Subject to constraints,

$$w_{11} x_1 + w_{12} x_2 + \ldots\ldots\ldots w_{1n} x_n \leq \text{or} = \text{or} \geq w_1 \quad \ldots(i)$$

$$w_{21} x_1 + w_{22} x_2 + \ldots\ldots\ldots w_{2n} x_n \leq \text{or} = \text{or} \geq w_2 \quad \ldots(ii)$$

. . .

. . .

. . .

$$w_{m1} x_1 + w_{m2} x_2 + \ldots\ldots\ldots w_{mn} x_n \leq = \text{or} \geq w_m \quad \ldots(iii)$$

Non-negativity constraint,

$x_i \geq 0$ (where $i = 1, 2, 3 \ldots\ldots n$)

Check Your Progress 1

State whether the following is true or false.

1. LP is a widely used mathematical modeling technique.
2. LP consists of linear objectives and linear constraints.
3. Divisibility refers to the aim to optimize.
4. Limited resources mean limited number of labour, material equipment and finance.
5. The objective function represents the aim or goal of the system, which has to be determined from the solution.

12.4 FORMULATION OF TRANSPORTATION

Vogel's Approximation Method (VAM)

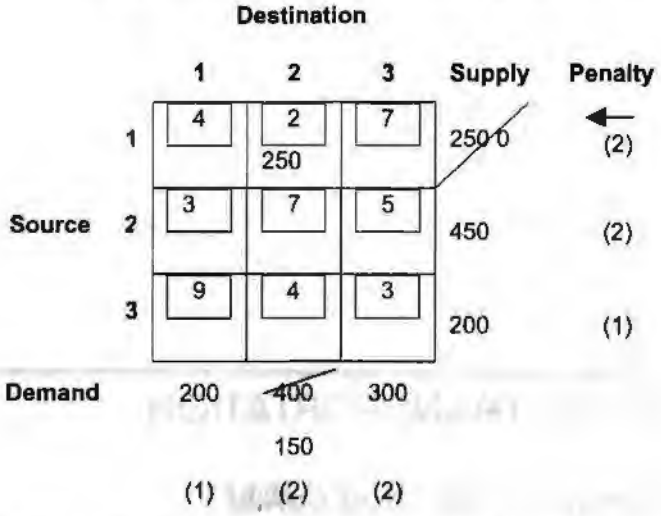
The penalties for each row and column are calculated (steps given on pages 176-77) Choose the row/ column, which has the maximum value for allocation. In this case there are five penalties, which have the maximum value 2. The cell with least cost is Row 3 and hence select cell (3,4) for allocation. The supply and demand are 500 and 300 respectively and hence allocate 300 in cell (3,4) as shown in Table 12.2.

Table 12.2: Penalty Calculation for each Row and Column

		Destination					
		1	2	3	4	Supply	Penalty
Source	1	4	2	7	3	250	(1)
	2	3	7	5	8	450	(2)
	3	9	4	3	1	~~500~~ 200	← (2)
					300		
Demand		200	400	300	~~300~~ 0		
		(1)	(2)	(2)	(2)		

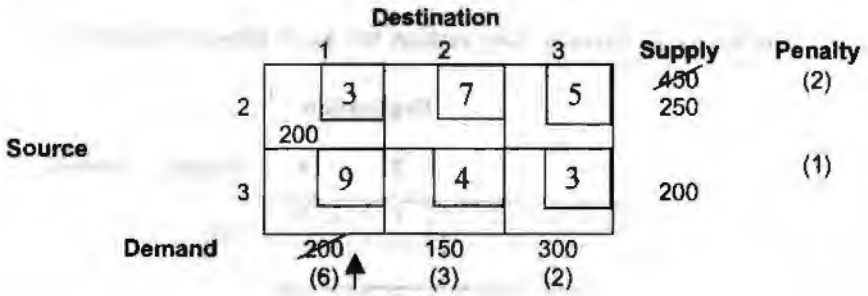
Since the demand is satisfied for destination 4, delete column 4 . Now again calculate the penalties for the remaining rows and columns.

Table 12.3: Exhausted Column 4 Deleted



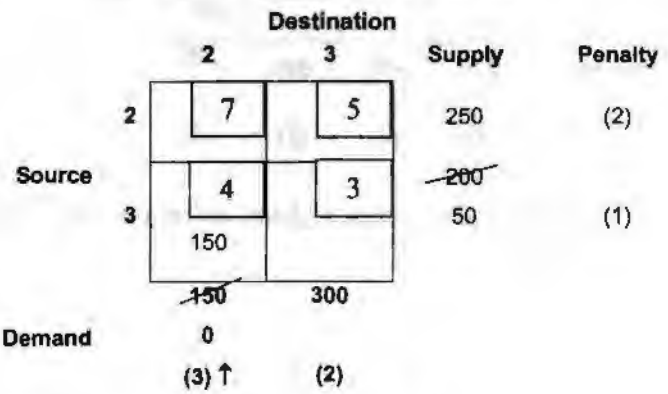
In the Table 12.4 shown, there are four maximum penalties of values which is 2. Selecting the least cost cell, (1,2) which has the least unit transportation cost 2. The cell (1, 2) is selected for allocation as shown in Table 12.4. Table 12.4 shows the reduced table after deleting row 1.

Table 12.4: Row 1 Deleted



After deleting column 1 we get the table as shown in the Table 12.5 below.

Table 12.5: Column 1 Deleted



Finally we get the reduced table as shown in Table 12.6

Table 12.6: Final Reduced Table

		Destination 3	Supply
Source	2	5 250	250 0
	3	3 50	50 0
Demand		300 0	

The initial basic feasible solution is shown in Table 12.7.

Table 12.7: Initial Basic Feasible Solution

	Destination W_1	W_2	W_3	W_4	Supply
F_1	140 17	5	9	65	140 (4) (4) (8) (48) (48)
F_2	50 20	10	12	210 65	260 (2) (2) (8) (45) (45)
F_3	10 15	100 0	250 5	65	360 (5) (5) (10) (50) __
F_4	13	220 1	10	65	220 (9) _____
Demand	200	320	250	210	
	(2)	(1)	(4)	(0)	
	(2)	(5)	(4)	(0)	
	(2)	—	—	(0)	
	(2)	—	—	(0)	
	(3)?	—	—	(0)	

Transportation cost = $(2 \times 250) + (3 \times 200) + (5 \times 250) + (4 \times 150) + (3 \times 50) + (1 \times 300)$

$= 500 + 600 + 1250 + 600 + 150 + 300$

= Rs. 3,400.00

Example 12.4: Find the initial basic solution for the transportation problem and hence solve it.

Table 12.8: Transportation Problem

		Destination 1	2	3	4	Supply
Source	1	4	2	7	3	250
	2	3	7	5	8	450
	3	9	4	3	1	500
Demand		200	400	300	300	

Solution: Vogel's Approximation Method (VAM) is preferred to find initial feasible solution. The advantage of this method is that it gives an initial solution which is nearer to an optimal solution or the optimal solution itself.

Step 1: The given transportation problem is a balanced one as the sum of supply equals to sum of demand.

Step 2: The initial basic solution is found by applying the Vogel's Approximation method and the result is shown in Table 12.9.

Table 12.9: Initial Basic Solution Found by Applying VAM

		Destination 1	2	3	4	Supply
Source	1	4	2 250	7	3	250
	2	3 200	7	5 250	8	450
	3	9	4 150	3 50	1 300	500
Demand		200	400	300	300	

Step 3: Calculate the Total Transportation Cost.

Initial Transportation cost = $(2 \times 250) + (3 \times 200) + (5 \times 250) + (4 \times 150) + (3 \times 50) + (1 \times 300)$

$= 500 + 600 + 1250 + 600 + 150 + 300$

= Rs. 3,400

Step 4: Check for degeneracy. For this, verify the condition,

Number of allocations, $N = m + n - 1$

$6 = 3 + 4 - 1$

$6 = 6$

Since the condition is satisfied, degeneracy does not exist.

Step 5: Test for optimality using modified distribution method. Compute the values of U_i and V_j for rows and columns respectively by applying the formula for occupied cells.

$$C_{ij} + U_i + V_j = 0$$

Then, the opportunity cost for each unoccupied cell is calculated using the formula $\overline{C_{ij}} = C_{ij} + U_i + V_j$ and denoted at the left hand bottom corner of each unoccupied cell. The computed valued of u_j and v_j and are shown in Table 12.10.

Table 12.10: Calculation of the Opportunity Cost

		Destination 1	2	3	4	Supply
Source	1	4 / 5	2 / 250	7 / 6	3 / 4	250 $U_1 = 2$
	2	3 / 200	7 / 1	5 / 250	8 / 5	450 $U_2 = -2$
	3	9 / 8	4 / 150	3 / 50	1 / 300	500 $U_3 = 0$
Demand		200	400	300	300	
		$V_1 = 1$	$V_2 = 4$	$V_3 = -3$	$V_4 = 1$	

Calculate the values of U_i and V_j, using the formula for occupied cells. Assume any one of U_i and V_j value as zero (U_3 is taken as 0)

$$C_{ij} + U_i + V_j = 0$$

$$4 + 0 + V_2 = 0, \quad V_2 = -4$$

$$5 + V_2 - 3 = 0, \quad U_2 = -2$$

$$3 - 2 + V_1 = 0, \quad V_1 = -1$$

$$2 - 4 + U_1 = 0, \quad U_1 = 2$$

Calculate the values of $\overline{C_{ij}}$, using the formula for unoccupied cells

$$\overline{C_{ij}} = C_{ij} + U_i + V_j$$

$$C_{11} = 4 + 2 - 1 = 5$$

$$C_{13} = 7 + 2 - 3 = 6$$

$$C_{14} = 3 + 2 - 1 = 4$$

$$C_{22} = 7 - 2 - 4 = 1$$

$$C_{24} = 8 - 2 - 1 = 5$$

$$C_{31} = 9 + 0 - 1 = 8$$

Since all the opportunity cost, $\overline{C_{ij}}$ values are positive the solution is optimum.

Total transportation cost $= (2 \times 25) + (3 \times 200) + (5 \times 250) + (4 \times 150) + (3 \times 50) + (1 \times 300)$

$= 50 + 600 + 1250 + 600 + 150 + 300$

$=$ Rs 2,950/-

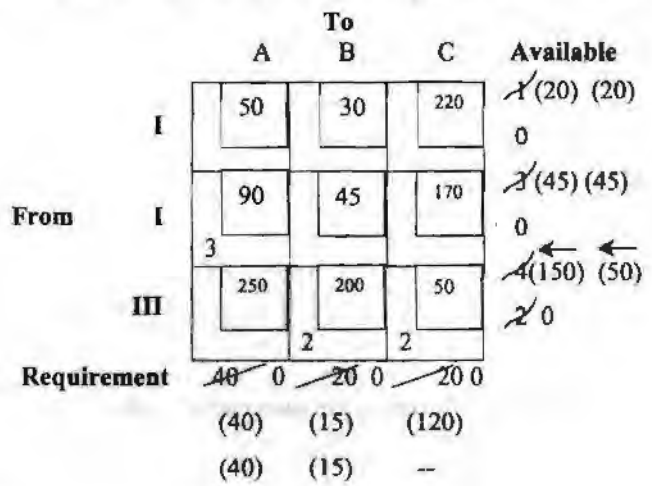
Example 12.5: Find the initial basic feasible solution for the transportation problem given in Table 12.11.

Table 12.11: Transportation Problem

From	To			Available
	A	B	C	
I	50	30	220	1
II	90	45	170	3
III	250	200	50	4
Requirement	4	2	2	

Solution: The initial basic feasible solution using VAM is shown in Table 12.12.

Table 12.12: Initial Basic Feasible Solution Using VAM



Check for degeneracy,

The number of allocations, N must be equal to $m + n - 1$.

i.e. $N = m + n - 1$

$5 = 3 + 3 - 1$

since $4 \neq 5$, therefore degeneracy exists.

To overcome degeneracy, the condition $N = m + n - 1$ is satisfied by allocating a very small quantity, close to zero in an occupied independent cell. (i.e., it should not form a closed loop) or the cell having the lowest transportation cost. This quantity is denoted by ϵ.

This quantity would not affect the total cost as well as the supply and demand values. Table 12.13 shows the resolved degenerate table.

Table 12.13: Resolved Degenerate Table

From	To A	To B	To C	Available
I	50	30	220	1
I	90 (3)	45	170	3
III	250 (ε)	200 (2)	50 (2)	4
Requirement	4	2	2	

Total transportation cost $= (50 \times 1) + (90 \times 3) + (200 \times 2) + (50 \times 2) + (250 \times e)$

$= 50 + 270 + 400 + 100 + 250\ e$

$= 820 + 250\ e$ = Rs. 820 since $e \rightarrow 0$

Example 12.6: Obtain an optimal solution for the transportation problem by MODI method given in Table 12.14.

Table 12.14: Transportation Problem

Source	Destination D_1	D_2	D_3	D_4	Supply
S_1	19	30	50	10	7
S_2	70	30	40	60	9
S_3	40	8	70	20	18
Demand	5	8	7	14	

Solution:

Step 1: The initial basic feasible solution is found using Vogel's Approximation Method as shown in Table 12.15.

Table 12.15: Initial Basic Feasible Solution Using VAM

		Destination				
		D_1	D_2	D_3	D_4	Supply
ource	S_1	19	30	50	10	(9) (9) (40) (40 0
	S_2	70	30	40 7	60 2	(10) (20) (20) (2 0
	S_3	40	8 8	70	20	(12) (20) (50) - 0
	Demand	0	0	0		0
		(21)	(22)	(10)	(10)	
		(21)	--	(10)	(10)	
		--	--	(10)	(10)	
		--	--	(10)	(50)↑	

Total transportation cost = $(19 \times 5) + (10 \times 2) + (40 \times 7) + (60 \times 2) + (8 \times 8) + (20 \times 10)$

$= 95 + 20 + 280 + 120 + 64 + 200$

= Rs. 779.00

Step 2: To check for degeneracy, verify the number of allocations, $N = m + n - 1$. In this problem, number of allocation is 6 which is equal $m + n - 1$.

$\therefore \quad N = m + n - 1$

$6 = 3 + 4 - 1$

$6 = 6$ therefore degeneracy does not exist.

Step 3: Test for optimality using MODI method. In Table 12.16 the values of U_i and V_j are calculated by applying the formula $C_{ij} + U_i + V_j = 0$ for occupied cells , and $\overline{C_{ij}} = C_{ij} + U_i + V_j$ for unoccupied cells respectively.

Table 12.16: Optimality Test Using MODI Method

		Destination				
		D_1	D_2	D_3	D_4	Supply
Source	S_1	19 5	30 32	50 60	10 2	$7\ U_1 = 0$
	S_2	70 1	30 -18	40 7	60 2	$9U_2 = -50$
	S_3	40 11	8 8	70 70	20 10	$18\ U_3 = -10$
	Demand	5	8	7	14	

$V_1 = -19 \qquad V_2 = 2 \qquad V_3 = 10 \qquad V_4 = -10$

Find the values of the dual variables U_i and V_j for occupied cells.

Initially assume $U_i = 0$,

$$C_{ij} + U_i + V_j = 0,$$

$$19 + 0 + V_1 = 0, \qquad V_1 = -19$$

$$10 + 0 + V_4 = 0, \qquad V_4 = -10$$

$$60 + U_2 - 10 = 0, \qquad U_2 = -50$$

$$20 + U_3 - 10 = 0, \qquad U_3 = -10$$

$$8 - 10 + V_2 = 0, \qquad V_2 = 2$$

$$40 - 50 + V_3 = 0, \qquad V_3 = 10$$

Find the values of the opportunity cost, $\overline{C_{ij}}$ for unoccupied cells,

$$\overline{C_{ij}} = C_{ij} + U_i + V_j$$

$$C_{12} = 30 + 0 + 2 = 32$$

$$C_{13} = 50 + 0 + 10 = 60$$

$$C_{21} = 70 - 50 - 19 = 1$$

$$C_{22} = 30 - 50 + 2 = -18$$

$$C_{31} = 40 - 10 - 19 = 11$$

$$C_{33} = 70 - 10 + 10 = 70$$

In Table the cell (2,2) has the most negative opportunity cost. This negative cost has to be converted to a positive cost without altering the supply and demand value.

Step 4: Construct a closed loop. Introduce a quantity + q in the most negative cell (S_2, D_2) and a put $-q$ in cell (S_3, D_2) in order to balance the column D_2. Now, take a right angle turn and locate an occupied cell in column D_4. The occupied cell is (S_3, D_4) and put $a + q$ in that cell. Now, put $a - q$ in cell (S_2, D_4) to balance the column D_4. Join all the cells to have a complete closed path. The closed path is shown in Figure 12.6.

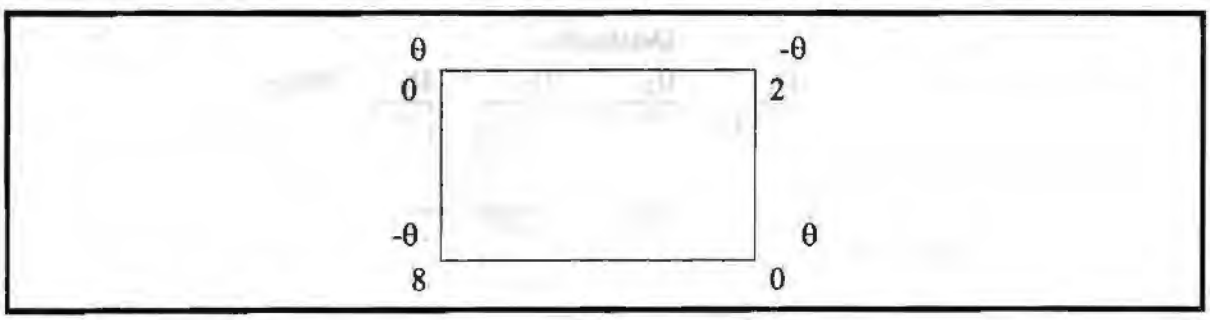


Figure 12.6

Closed Path

Now, identify the – q values, which are 2 and 8. Take the minimum value, 2 which is the allocating value. This value is then added to cells (S_2, D_2) and (S_3, D_4) which have '+' signs and subtract from cells (S_2, D_4) and (S_3, D_2) which have '–' signs. The process is shown in Figure 12.7

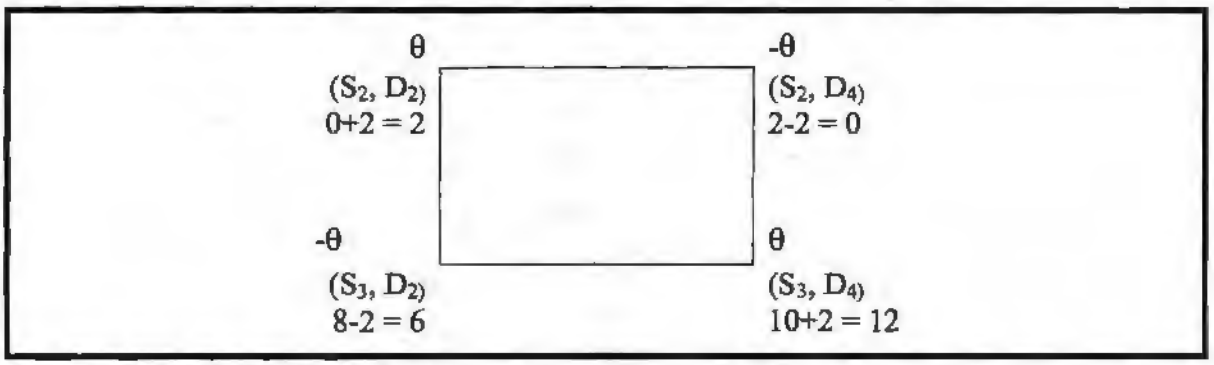


Figure 12.7

Table 12.17: Closed Path

		Destination				
		D_1	D_2	D_3	D_4	Supply
Source	S_1	19 (5)	30 (32)	50 (60)	10 (2)	7
	S_2	70 (1)	30 (θ)	40 (7)	60 (2 -θ)	9
	S_3	40 (11)	8 (-θ 8)	70 (70)	20 (10 θ)	18
Demand		5	8	7	14	

The table after reallocation is shown in Table 12.18

Table 12.18: After Reallocation

		Destination				
		D_1	D_2	D_3	D_4	Supply
Source	S_1	19 (5)	30	50	10 (2)	7
	S_2	70	30 (2)	40 (7)	60	9
	S_3	40 (6)	8	70	20 (12)	18
Demand		5	8	7	14	

Now, again check for degeneracy. Here allocation number is 6.

Verify whether number of allocations,

$$N = m + n - 1$$

$$6 = 3 + 4 - 1$$

$$6 = 6$$

therefore degeneracy does not exits.

Again find the values of U_i, V_j and $\overline{C_{ij}}$ for the Table 12.19 shown earlier.

For occupied cells, $C_{ij} + U_i + V_j = 0$

$$19 + 0 + V_1 = 0, \qquad V_1 = -19$$

$$10 + 0 + V_4 = 0, \qquad V_4 = -10$$

$$20 + U_3 - 10 = 0, \qquad U_3 = -10$$

$$8 - 10 + V_2 = 0, \qquad V_2 = 2$$

$$30 + U_2 + 2 = 0, \qquad U_2 = -32$$

$$40 - 50 + V_3 = 0, \qquad V_3 = -10$$

For unoccupied cells, $\overline{C_{ij}} = C_{ij} + U_i + V_j$

$$C_{12} = 30 + 0 + 20 = 50$$

$$C_{13} = 50 + 0 - 8 = 42$$

$$C_{21} = 70 - 32 - 19 = 19$$

$$C_{24} = 60 - 32 - 10 = 18$$

$$C_{31} = 40 - 10 - 19 = 11$$

$$C_{33} = 70 - 10 - 8 = 52$$

The values of the opportunity cost $\overline{C_{ij}}$ are positive. Hence the optimality is reached. The final allocations are shown in Table 12.19.

Table 12.19: Final Allocation

		Destination D_1	D_2	D_3	D_4	Supply	
Source	S_1	19	30	50	10		$7\ U_1 = 0$
		5			2		
	S_2	70	30	40	60		$9\ U_2 = -32$
			2	7			
	S_3	40	8	70	20		$18\ U_3 = -10$
			6		12		
Demand		5	8	7	14		

$V_1 = -19$ $V_2 = 2$ $V_3 = -8$ $V_4 = -10$

Total transportation cost $= (19 \times 5) + (10 \times 2) + (30 \times 2) + (40 \times 7) + (8 \times 6) + (20 \times 12)$

$= 95 + 20 + 60 + 280 + 48 + 240$

$=$ Rs. 743

Example 12.7: Solve the transportation problem

		Destination 1	2	3	Supply
Source	1	3	5	7	10
	2	11	8	9	8
	3	13	3	9	5
Demand		5	9	11	23 / 25

The problem is unbalanced if $S a_i = S b_j$ that is, when the total supply is not equal to the total demand. Convert the unbalanced problem into a balanced one by adding a dummy row or dummy column as required and solve.

Here the supply does not meet the demand and is short of 2 units. To convert it to a balanced transportation problem add a dummy row and assume the unit cost for the dummy cells as zero as shown in Table 12.20 and solve.

Table 12.20: Dummy Row Added to TP

		Destination 1	2	3	Supply
Source	1	3	5	7	10
	2	11	8	9	8
	3	13	3	9	5
	4	0	0	0	2
	Demand	5	9	11	25 / 25

MODI Method of Solving Transportation Problem

The MODI (*modified distribution*) method allows us to compute improvement indices quickly for each unused square without drawing all of the closed paths. Because of this, it can often provide considerable time savings over other methods for solving transportation problems.

MODI provides a new means of finding the unused route with the largest negative improvement index. Once the largest index is identified, we are required to trace only one closed path. This path helps determine the maximum number of units that can be shipped via the best unused route.

How to Use the MODI Method

In applying the MODI method, we begin with an initial solution obtained by using the northwest corner rule or any other rule. But now we must compute a value for each row (call the values R_1, R_2, R_3 if there are three rows) and for each column (K_1, K_2, K_3) in the transportation table. In general, we let

R_i = value assigned to row i

K_j = value assigned to column j

C_{ij} = cost in square ij (cost of shipping from source i to destination) j

The MODI method then requires five steps:

1. To compute the values for each row and column, set

   $$R_i + K_j = C_{ij}$$

   but *only for those squares that are currently used or occupied.* For example, if the square at the intersection of row 2 and column 1 is occupied, we set $R_2 + K_1 = C_{21}$.

2. After all equations have been written, set $R_1 = 0$.

3. Solve the system of equations for all R and K values.

4. Compute the improvement index for each unused square by the formula improvement index (I_{ij}) $= C_{ij} R_i K_j$

5. Select the largest negative index and proceed to solve the problem as you did using the stepping-stone method.

Solving the Arizona Plumbing Problem with MODI

Let us try out these rules on the Arizona Plumbing problem. The initial northwest corner solution is shown in Table 12.21. MODI will be used to compute an improvement index for each unused square. Note that the only change in the transportation table is the border labeling the R_is (rows) and K_js (columns).

We first set up an equation for each occupied square:

1. $R_1 + K_1 = 5$
2. $R_2 + K_1 = 8$
3. $R_2 + K_2 = 4$
4. $R_3 + K_2 = 7$
5. $R_3 + K_3 = 5$

Letting $R_1 = 0$, we can easily solve, step by step, for K_1, R_2, K_2, R_3, and K_3.

1. $R_1 + K_1 = 5$

   $0 + K_1 = 5 \qquad K_1 = 5$
2. $R_2 + K_1 = 8$

   $R_2 + 5 = 8 \qquad R_2 = 3$
3. $R_2 + K_2 = 4$

   $3 + K_2 = 4 \qquad K_2 = 1$

Table 12.21: Initial Solution to Arizona Plumbing Problem in the MODI Format

	K_j	K_1	K_2	K_3	
R_i	TO / FROM	ALBUQUERQUE	BOSTON	CLEVELAND	FACTORY CAPACITY
R_1	DES MOINES	5 / 100	4	3	100
R_2	EVANSVILLE	8 / 200	4 / 100	3	300
R_3	FORT LAUDERDALE	9	7 / 100	5 / 200	300
	WAREHOUSE REQUIREMENTS	300	200	200	700

4. $R_3 + K_2 = 7$

   $R_3 + 1 = 7 \qquad R_3 = 6$

5. $R_3 + K_3 = 5$

   $6 + K_3 = 5 \qquad K_3 = 1$

You can observe that these R and K values will not always be positive; it is common for zero and negative values to occur as well. After solving for the Rs and Ks in a few practice problems, you may become so proficient that the calculations can be done in your head instead of by writing the equations out.

The next step is to compute the improvement index for each unused cell. That formula is

$$\text{improvement index} = I_{ij} = C_{ij} - R_i - K_j$$

We have:

$$\begin{aligned}
\text{Des Moines–Boston index} &= I_{DB} \text{ (or } I_{12}) = C_{12} - R_1 - K_2 = 4 - 0 - 1 \\
&= +\$3 \\
\text{Des Moines–Cleveland index} &= I_{DC} \text{ (or } I_{13}) = C_{13} - R_1 - K_3 = 3 - 0 - (-1) \\
&= +\$4 \\
\text{Evansville–Cleveland index} &= I_{EC} \text{ (or } I_{23}) = C_{23} - R_2 - K_3 = 3 - 3 - (-1) \\
&= +\$1 \\
\text{Fort Lauderdale–Albuquerque index} &= I_{FA} \text{ (or } I_{31}) = C_{31} - R_3 - K_1 = 9 - 6 - 5 \\
&= -\$2
\end{aligned}$$

Because one of the indices is negative, the current solution is not optimal. Now it is necessary to trace only the one closed path, for Fort Lauderdale–Albuquerque, in order to proceed with the solution procedures.

The steps we follow to develop an improved solution after the improvement indices have been computed are outlined briefly:

1. Beginning at the square with the best improvement index (Fort Lauderdale–Albuquerque), trace a closed path back to the original square via squares that are currently being used.
2. Beginning with a plus (+) sign at the unused square, place alternate minus () signs and plus signs on each corner square of the closed path just traced.
3. Select the smallest quantity found in those squares containing minus signs. *Add* that number to all squares on the closed path with plus signs; *subtract* the number from all squares assigned minus signs.
4. Compute new improvement indices for this new solution using the MODI method.

Table 12.22: Second Solution to the Arizona Plumbing Problem

FROM \ TO	A	B	C	FACTORY
D	100 $5	$4	$3	100
E	100 $8	200 $4	$3	300
F	100 $9	$7	200 $5	300
WAREHOUSE	300	200	200	700

Table 12.23: Third and Optimal Solution to Arizona Plumbing Problem

FROM \ TO	A	B	C	FACTORY
D	100 $5	$4	$3	100
E	$8	200 $4	100 $3	300
F	200 $9	$7	100 $5	300
WAREHOUSE	300	200	200	700

Following this procedure, the second and third solutions to the Arizona Plumbing Corporation problem can be found. See Tables 12.22 and 12.23. With each new MODI solution, we must recalculate the R and K values. These values then are used to compute new improvement indices in order to determine whether further shipping cost reduction is possible.

12.5 ASSIGNMENT PROBLEMS

The basic objective of an assignment problem is to assign n number of resources to n number of activities so as to minimize the total cost or to maximize the total profit of allocation in such a way that the measure of effectiveness is optimized. The problem of assignment arises because available resources such as men, machines, etc., have varying degree of efficiency for performing different activities such as job. Therefore cost, profit or time for performing the different activities is different. Hence the problem is, how should the assignments be made so as to optimize (maximize or minimize) the given objective.

The assignment model can be applied in many decision-making processes like determining optimum processing time in machine operators and jobs, effectiveness of teachers and subjects, designing of good plant layout, etc. This technique is found suitable for routing travelling salesmen to minimize the total travelling cost, or to maximize the sales.

Hungarian Method for Solving Assignment Problem

Step 1: In a given problem, if the number of rows is not equal to the number of columns and vice versa, then add a dummy row or a dummy column. The assignment costs for dummy cells are always assigned as zero.

Step 2: Reduce the matrix by selecting the smallest element in each row and subtract with other elements in that row.

Step 3: Reduce the new matrix column-wise using the same method as given in step 2.

Step 4: Draw minimum number of lines to cover all zeros.

Step 5: If Number of lines drawn = order of matrix, then optimally is reached, so proceed to step 7. If optimally is not reached, then go to step 6.

Step 6: Select the smallest element of the whole matrix, which is NOT COVERED by lines. Subtract this smallest element with all other remaining elements that are NOT COVERED by lines and add the element at the intersection of lines. Leave the elements covered by single line as it is. Now go to step 4.

Step 7: Take any row or column which has a single zero and assign by squaring it. Strike off the remaining zeros, if any, in that row and column (X). Repeat the process until all the assignments have been made.

Step 8: Write down the assignment results and find the minimum cost/time.

Note: While assigning, if there is no single zero exists in the row or column, choose any one zero and assign it. Strike off the remaining zeros in that column or row, and repeat the same for other assignments also. If there is no single zero allocation, it means multiple number of solutions exist. But the cost will remain the same for different sets of allocations.

Example 12.8: Solve the following assignment problem shown in Table 12.24 using Hungarian method. The matrix entries are processing time of each man in hours.

Table 12.24: Assignment Problem

		Men				
		1	2	3	4	5
Job	I	20	15	18	20	25
	II	18	20	12	14	15
	III	21	23	25	27	25
	IV	17	18	21	23	20
	V	18	18	16	19	20

Solution: The row-wise reductions are shown in Table 12.25

Table 12.25: Row-wise Reduction Matrix

Job	Men 1	2	3	4	5
I	5	0	3	5	10
II	6	8	0	2	3
III	0	2	4	6	4
IV	0	1	4	6	3
V	2	2	0	3	4

The column wise reductions are shown in Table 12.26.

Table 12.26: Column-wise Reduction Matrix

Job	Men 1	2	3	4	5
I	5	0	3	3	7
II	6	8	0	0	0
III	0	2	4	4	1
IV	0	1	4	4	0
V	2	2	0	1	1

Matrix with minimum number of lines drawn to cover all zeros is shown in Table 12.27.

Table 12.27: Matrix will all Zeros Covered

Job	Men 1	2	3	4	5
I	~~5~~	~~0~~	~~3~~	~~3~~	~~7~~
II	~~6~~	~~8~~	~~0~~	~~0~~	~~0~~
III	0	2	4	4	1
IV	~~0~~	~~1~~	~~4~~	~~4~~	~~0~~
V	~~2~~	~~2~~	~~0~~	~~1~~	~~1~~

The number of lines drawn is 5, which is equal to the order of matrix. Hence optimality is reached. The optimal assignments are shown in Table 12.28.

Table 12.28: Optimal Assignment

		Men				
		1	2	3	4	5
Job	I	5	0	3	3	7
	II	6	8	~~0~~	0	~~0~~
	III	0	2	4	4	1
	IV	~~0~~	1	4	4	0
	V	2	2	0	1	1

Therefore, the optimal solution is:

Job	Men	Time
I	2	15
II	4	14
III	1	21
IV	5	20
V	3	16
Total time = 86 hours		

Check Your Progress 2

True or false

1. The basic objective of an assignment problem is to assign n number of resources to n number of activities so as to minimize the total cost or to maximize the total profit.
2. The assignment costs for dummy cells are always assigned as zero.

12.6 SUMMARY

Thus we can say that LP is a method of planning whereby objective function is maximized or minimized while at the same time satisfying the various restrictions placed on the potential solution. In technical words, linear programming is defined as a methodology whereby a linear function in optimized (minimized or maximized) subject to a set of linear constraints in the form of equalities or inequalities. Thus LP is a planning technique of selecting the best possible (optimal) strategy among number of alternatives.

Transportation problem is a particular class of linear programming, which is associated with day-to-day activities in our real life and mainly deals with logistics. It helps in solving problems on distribution and transportation of resources from one place to another.

AP brings into play the allocation of a number of jobs to a number of persons in order to minimize the completion time. Although an AP can be formulated as LPP, it solved by a special method known a Hungarian method. The Hungarian method of assignment provides us with an efficient means of

finding the optimal solution without having to make a direct comparison of every option. Further we will take into consideration the opportunity cost. This is a next best alternative cost.

12.7 KEYWORDS

- Linear Programming
- Constraints
- Optimality
- Hungarian Method
- Graphical Method
- Profit
- Transportation problem

12.8 REVIEW QUESTIONS

1. Define Linear Programming.
2. What are the essentials of LP Model?
3. Why linear programming is used?
4. What is the transportation problem?
5. Determine the feasible space for each of the following constraints:
   (a) $2x_1 - 2x_2 \leq 5$
   (b) $5x_1 + 10x_2 \leq 60$
   (c) $x_1 - x_2 \leq 0$
   (d) $4x_1 + 3x_2 \geq 15$
6. A company manufactures two types of products, *A* and *B*. Each product uses two processes, I and II. The processing time per unit of product *A* on process I is 6 hours and on the process II is 5 hours. The processing time per unit of product *B* on process I is 12 hours and on process II is 4 hours. The maximum number of hours available per week on process I and II are 75 and 55 hours respectively. The profit per unit of selling *A* and *B* are Rs.12 and Rs.10 respectively.
   (i) Formulate a linear programming model so that the profit is maximized.
   (ii) Solve the problem graphically and determine the optimum values of product A and B.
7. Solve the following LP graphically:
   Maximize $Z = 8x_1 + 10x_2$
   Subject to constraints,
   $2x_1 + 3x_2 \geq 20$
   $4x_1 + 2x_2 \geq 25$
   Where $x_1, x_2 \geq 0$
8. A company has plants at locations A, B and C with the daily capacity to produce chemicals to a maximum of 3000 kg, 1000 kg and 2000 kg respectively. The cost of production (per kg) are Rs. 800 Rs. 900 and Rs. 7.50 respectively. Customer's requirement of chemicals per day is as follows:

Customer	Chemical Required	Price offered
1	2000	200
2	1000	215
3	2500	225
4	1000	200

Transportation cost (in rupees) per kg from plant locations to customer's place is given in table.

		Customer			
		1	2	3	4
Plant	A	5	7	10	12
	B	7	3	4	2
	C	4	6	3	9

Find the transportation schedule that minimizes the total transportation cost.

9. A transportation model has four supplies and five destinations. The following table shows the cost of shipping one unit from a particular supply to a particular destination.

Source	Destination					Supply
	1	2	3	4	5	
1	13	6	9	6	10	13
2	8	2	7	7	9	15
3	2	12	5	8	7	13
Demand	10	15	7	10	2	

The following feasible transportation pattern is proposed:

$x_{11} = 10, x_{12} = 3, x_{22} = 9, x_{23} = 6, x_{33} = 9, x_{34} = 4, x_{44} = 9, x_{45} = 5.$

Test whether these allocations involve least transportation cost. If not, determine the optimal solution.

10. Consider the following assignment problem:

Destination	Unit cost (Rs.)				Supply
	1	2	3	4	
Source					
1	30	61	45	50	1
2	25	54	49	52	1
3	27	60	45	54	1
4	31	57	49	55	1
Demand	1	1	1	1	

(a) Draw the network representation of the assignment problem.

(b) Formulate a linear programming model for the assignment problem.

11. A consumer durables manufacturing company has plans to increase its product line, namely, washing machine, refrigerator, television and music system. The company is setting up new plants and considering four locations. The demand forecast per month for washing machine, refrigerator, television and music system are 1000, 750, 850 and 1200, respectively. The company decides to produce the forecasted demand. The fixed and variable cost per unit for each location and item is given in the following table. The management has decided not to set-up more than one unit in one location.

Location	Fixed cost (lakhs)				Variable cost / unit			
	WM	RF	TV	MS	WM	RF	TV	MS
Chennai	30	35	18	16	4	3	6	2
Coimbatore	25	40	16	12	3	2	4	4
Madurai	35	32	15	10	4	2	7	6
Selam	20	25	14	12	2	1	3	7

Determine the location and product combinations so that the total cost is minimized.

Answers to Check Your Progress

Check Your Progress 1

1. True
2. True
3. False
4. True
5. False

Check Your Progress 2

1. True
2. True

12.9 REFERENCES AND FURTHER READING

- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2021). Multivariate data analysis (8th ed.). Cengage Learning.
- Pallant, J. (2022). SPSS survival manual: A step by step guide to data analysis using SPSS (7th ed.). Open University Press.
- Boslaugh, S. (2023). Statistics in a nutshell (2nd ed.). O'Reilly Media.
- Cohen, L., Manion, L., & Morrison, K. (2024). Research methods in education (9th ed.). Routledge.