S.Y.B.Sc. (IT)

SEMESTER - IV (CBCS)

COMPUTER ORIENTED STATISTICAL TECHNIQUES

SUBJECT CODE : USIT403

Prof. Suhas Pednekar
Vice Chancellor
University of Mumbai, Mumbai.

Prof. Ravindra D. Kulkarni
Pro Vice-Chancellor,
University of Mumbai.

Prof. Prakash Mahanwar
Director
IDOL, University of Mumbai.

Programe Co-ordinator : **Mandar L. Bhanushe**
Head, Faculty of Science and Technology,
IDOL, University of Mumbai – 400098.

Course Co-ordinator : **Ms. Gouri Sawant**
Assistant Professor B.Sc.I.T,
IDOL, University of Mumbai

Editor : **Dr. Aarti Nayak**
Assistant Professor, S.K Somaiya College,
Somaiya University

Course Writers : **Ms. Kumudini Manwar**
Assistant Professor, Sinhgad Institute of Management, Pune

: **Mr. Nimesh Punjani**
Assistant Professor, Lala Lajpatrai College, Mahalakshmi,
Mumbai - 34

: **Mr. Kiran Patil**
Assistant Professor, Anandibai Raorane Arts,
Commerce and Science College Vaibhavwadi

: **Dr. Manisha Pithadia**
Assistant Professor,
Viva College of Arts,Science and Commerce

November 2021, Print I

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

DTP COMPOSED AND PRINTED BY
Mumbai University Press
Vidyanagari, Santacruz (E), Mumbai - 400098.

CONTENTS

*****

UNIT I

1

THE MEAN, MEDIAN, MODE, AND OTHER MEASURES OF CENTRAL TENDENCY

Unit structure

1.0 OBJECTIVES

After going through this chapter, students will able to learn

- To present huge data in a summarized form
- To calculate and interpret the mean, the median and the mode,
- To facilitate comparison
- To calculate geometric mean, harmonic mean
- To trace precise relationship

- To calculate Quartiles, deciles and percentiles
- To help in decision-making

1.1 INTRODUCTION

A measure of central tendency is a single value that describes a set of data by identifying the central position within that set of data. Mean, median and mode are different measures of central tendency in a numerical data. The word average is commonly used in day to day conversation like we often talk about average height of the girls, average student of the class, and average run rate of the match. When we say average means neither too good nor bad. However, in statistics the term average has different meaning. Average is a single value which representing a group of values so such a value easy to understand, easy to compute and based on all observations.

1.2 INDEX OR SUBSCRIPT, NOTATION, SUMMATION NOTATION

Let the symbol X_i (read 'X subscript i) denote any of the N values X_1, X_2, X_3, …….X_N assumed by a variable X. The letter i in X_i which can stand for any of the numbers 1, 2, 3, ……, N is called a subscript or index. Any letter other than i such as j, k, p, q or r could be used also.

Summation Notation:

The symbol $\sum_{i=1}^{N} X_i$ is used to denote the sum of all theX_i's from i = 1 to N.

$\sum_{i=1}^{N} X_i = X_1 + X_2 + X_3 + \ldots\ldots + X_N$

We generally denote this sum simply by $\sum X$, $\sum X_i$.

The symbol $\sum$ is the Greek capital letter *sigma* denoting sum.

Ex. $\sum_{i=1}^{N} aX_i = aX_1 + aX_2 + aX_3 + \ldots\ldots + aX_N$

$= a(X_1 + X_2 + X_3 + \ldots\ldots + X_N) = a\sum_{i=1}^{N} X_i$, where a is a constant.

1.3 AVERAGES OR MEASURES OF CENTRAL TENDENCY

There are different ways of measuring the central tendency of a set of values.
Various authors defined Average differently.

"Average is an attempt to find one single figure to describe whole of figures." **– Clark**

"An average is a single value selected from a group of values to represent them in some way- a value which is supposed to stand for whole group, of which it is a part, as typical of all the values in the group." **– A. E. Waugh**

"An average is a typical value in the sense that it is sometimes employed to represent all the individual values in the series or of a variable." **– Ya-Lun-Chou**

Types of Averages:

Arithmetic Mean: a. Simple, b. Weighted

Median

Mode

Geometric Mean

Harmonic Mean

1.4 ARITHMETIC MEAN

The most popular and widely used measure of representing the entire data by one value is mean or Average.

It simply involves taking the sum of a group of numbers, then dividing that sum by the total number of values in the group.

Arithmetic mean can be of two types.

a. Simple arithmetic mean
b. Weighted arithmetic mean

A. Simple Arithmetic Mean – Individual Observations:

Calculation of mean in case of individual observations [i. e. when frequencies are not given] is very simple. Here, we add all values of the variable and divide the total by the number of items.

$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{N} = \frac{\sum X}{N}$

$\bar{X}$ = Arithmetic Mean; N = number of observations;

$\sum X$= sum of all the values of the variable X i. e. $X_1, X_2, X_3, \ldots, X_n$

Ex 1. Find the Arithmetic mean of following five values 8, 45, 49, 54, 79.

Sol: We know that, $\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{N}$

$\bar{X} = \frac{8+45+49+54+79}{5} = \frac{235}{5} = 47$

Ex 2. Find the Arithmetic mean of following values.
4350, 7200, 6750, 5480, 7940, 3820, 5920, 8450, 4900, 5350.

Sol: We know that, $\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{N}$

$\bar{X} = \frac{8350+ 7200+ 6750+ 5480+ 7940+ 3820+ 5920+ 8450+ 4900+ 5350}{10} = \frac{64160}{10} =$ 6416

Short cut method: $\bar{X} = A + \frac{\sum d}{N}$

Where A is assumed mean and d is deviation of items from assumed mean i. e. $d = (X - A)$.

Ex 3. Calculate arithmetic mean from following data.
2690, 3670, 4580, 5660, 2750, 2830, 4100, 5720, 5040, 4840
Sol:

X	$d = (X-A)$
2690	-2310
3670	-1330
4580	-420
5660	660
2750	-2250
2830	-2170
4100	-900
5720	720
5040	40
4840	-160
	$\sum d$ = -8120

Consider assumed mean, A = 5000
$\bar{X} = A + \frac{\sum d}{N} = 5000 - \frac{8120}{10} = 4188$

1.4.1 Arithmetic Mean Computed from Grouped Data:

Simple Arithmetic Mean – Discrete series:
Calculation of mean in case of frequencies are given,
$\bar{X} = \frac{\sum f X}{N}$
f = Frequency;
X = the variable
N = Total number of observations i.e. $\sum f$

Here, first multiply the frequency of each row with variable and obtain the total $\sum f X$ and then divide the total by number of observations, i.e. total frequency.

Ex 4. Following are the marks obtained by 60 students. Calculate arithmetic mean.

Marks	15	30	45	60	70	80
No. of students	6	14	15	15	4	6

Sol: Let the marks denoted by X and number of students denoted by f.

Marks X	No. of Students f	fX
15	6	90
30	14	420
45	15	675
60	15	900
70	4	280
80	6	480
	N = 60	$\sum f X = 2845$

$\bar{X} = \frac{\sum f X}{N} = \frac{2845}{60} = 47.42$

Short cut method: $\bar{X} = A + \frac{\sum fd}{N}$

Where A is assumed mean and d is deviation of items from assumed mean i. e. $d = (X - A)$,

$N = \sum f$

Ex 5. Calculate arithmetic mean by the short cut method using data from Ex. 4

Sol:

Marks X	**No. of Students f**	**$d = (X - A)$**	**fd**
15	6	-30	-180
30	14	-15	-210
45	15	0	0
60	15	15	225
70	4	25	100
80	6	35	210
	N = 60		$\sum f X = 145$

Assumed mean, A = 45

$\bar{X} = A + \frac{\sum fd}{N} = 45 + \frac{145}{60} = 47.4166$

Simple Arithmetic Mean – Continuous Series:

$\bar{X} = \frac{\sum fm}{N}$

m = mid-point of various classes; f = the frequency of each class;

N = the total frequency

Here, first obtain the mod-point of each class and denote it by m.

Multiply these mid-points by the respective frequency of each class and obtain the total $\sum fm$

Divide the total by the sum of the frequency, i.e. N.

Ex 5. From the following data compute arithmetic mean.

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

Sol:

Marks	**Mid- point m**	**No. of Students f**	**fm**
0-10	5	5	25
10-20	15	10	150
20-30	25	25	625
30-40	35	30	1050
40-50	45	20	900
50-60	55	10	550
		N = 100	$\sum f\, m$ = 3300

$\bar{X} = \frac{\sum f\, m}{N} = \frac{3300}{100} = 33$

Ex 6. From the following data compute arithmetic mean.

Class Intervals	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	4	4	7	10	12	8	5

Sol:

Marks	**Mid- point m**	**No. of Students f**	**f m**
0-10	5	4	20
10-20	15	4	60
20-30	25	7	175
30-40	35	10	350
40-50	45	12	540
50-60	55	8	440
50-60	65	5	325
		N = 50	$\sum f\, m$ = 1910

$\bar{X} = \frac{\sum f\, m}{N} = \frac{1910}{50} = 38.2$

Short cut method: $\bar{X}$ = A + $\frac{\sum fd}{N}$

Where A is assumed mean and d is deviation of items from assumed mean i. e. d = (m − A), m= mid point , N = $\sum f$

Ex 7. Calculate arithmetic mean by the short cut method using data from Ex. 5.

Sol:

Marks	Mid- point m	No. of Students f	d = (m − A)	fd
0-10	5	5	-30	-150
10-20	15	10	-20	-200
20-30	25	25	-10	-250
30-40	35	30	0	0
40-50	45	20	10	200
50-60	55	10	20	200
		N = 100		$\sum f m$ = 200

Assumed mean, A = 35

$\bar{X}$ = A + $\frac{\sum fd}{N}$ = 35 - $\frac{200}{100}$ = 33

1.4.2 Properties of the Arithmetic Mean:

1. The sum of deviation from their arithmetic mean is always equal to zero.

Symbolically, $\sum$(X − $\bar{X}$)= 0

Ex 8:

X	10	20	30	40	50	$\sum X$ = 150
X - $\bar{X}$	-20	-10	0	10	20	$\sum$X − $\bar{X}$= 0

$\bar{X} = \frac{\sum X}{N} = \frac{150}{5}$ = 30

When we calculate the deviations of all the items from their arithmetic mean ($\bar{X}$ =30), we find that the sum of the deviations from the arithmetic mean i. e. $\sum$(X − $\bar{X}$)= 0

2. The sum of squared deviations of the items from arithmetic mean is minimum, that is, less than the sum of squared deviations of the items from any other value.

Ex 9:

X	X - $\bar{X}$	$(X - 4)^2$
2	-2	4
3	-1	1
4	0	0

5	1	1
6	2	4
$\sum X = 20$	$\sum X - \bar{X} = 0$	$\sum(X - \overline{X})^2 = 0$

$\bar{X} = \frac{\sum X}{N} = \frac{20}{5} = 4$

The sum of the squared deviations is equal to 10 in the above example. If the deviations are taken from any other value the sum of the squared deviations are taken from any other value the sum of the squared deviations would be greater than 10.

Let us calculate the squares of the deviations of item from the value less than the arithmetic mean, say 3

X	X - 3	$(X - 3)^2$
2	-1	1
3	0	0
4	1	1
5	2	4
6	3	9
$\sum X = 20$		$\sum(X - 3)^2 = 0$

3. Arithmetic mean is NOT independent of change of origin.

If each observation of a series is increased (or decreased) by a constant, then the mean of these observations is also increased (or decreased) by that constant.

4. Arithmetic mean is NOT independent of change of scale.

If each observation of a series is multiplied (or divided) by constant, then the mean of these observations is also multiplied (or divided) by that constant.

5. If arithmetic mean and number of items of two or more related groups are given, then we can compute the combined mean using the formula given below.

$\bar{X}_{12} = \frac{N1\overline{X1} + N2\overline{X2}}{N1+N2}$,

Where

$\bar{X}_{12}$ = Combined mean of two groups;

N1 = Number of items in the first group; N2 = Number of items in the second group

$\overline{X1}$ = Arithmetic mean of the first group; $\overline{X2}$ = Arithmetic mean of the second group

1.5 WEIGHTED ARITHMETIC MEAN

Arithmetic mean gives equal importance to all the items. When importance of the items are not same, in these cases we compute weighted arithmetic mean. The term weighted represents to the relative importance to the item.

$$\bar{X}_w = \frac{W_1X_1 + WX_2 + W_3X_3 + \cdots + W_nX_n}{W_1 + W_2 + \ldots\ldots + W_n} = \frac{\sum WX}{\sum W}$$

Where
$\bar{X}_w$ represent the weighted arithmetic mean; X represent the variable values i. e. X_1, X_2 …… X_n

W represent the weights attached to the variable values i. e. w_1, w_2 …… w_n respectively.

To calculate weighted arithmetic mean, multiply the weight by the variable X and obtain the total $\sum WX$. Then divide this total by the sum of the weights, i.e. $\sum W$

In case of frequency distribution, if f_1, f_2. ….. . f_n are the frequencies of the variable values X_1, X_2,……X_n respectively then the weighted arithmetic mean is given by

$$\bar{X}_w = \frac{W_1\,(f_1\ X_1) + W_2\,(f_2\ X_2) + \ldots\ldots\ldots + W_n\,(f_n\ X_n)}{W}$$

$$\bar{X}_w = \frac{\sum w(fX)}{\sum w}$$

Note: Simple arithmetic mean shall be equal to the arithmetic mean if the weights are equal.

Ex. 10 Calculate the weighted mean for following data.

X	1	2	5	7
W	2	14	8	32

Sol:

X	W	WX
1	2	2
2	14	28
5	8	40
7	32	224
	$\sum W$ = 56	$\sum WX$ = 294

$\bar{X}_w = \frac{\sum WX}{\sum W} = \frac{294}{56} = 5.25$

Ex. 11 Calculate the weighted mean for following data.

Wages per Day (X)	200	150	85
No. of workers (W)	25	20	10

Sol:

X	W	WX
200	25	5000
150	20	3000
85	10	850
	$\sum W = 55$	$\sum WX = 8850$

$\bar{X}_w = \frac{\sum WX}{\sum W} = \frac{8850}{55} = 160.90$

Ex. 12. Calculate the weighted mean for following data and compare it with arithmetic mean

Subject	Weight	Student		
		X	Y	Z
Physics	2	72	42	52
Chemistry	3	75	52	62
Biology	5	58	88	68

Sol: For Student X,

Arithmetic Mean, $\bar{X}_X = \frac{\sum X}{3} = \frac{72+75+58}{3} = \frac{203}{3} = 67.67$

Weighted Arithmetic Mean, $\bar{X}_{wX} = \frac{\sum WX}{\sum W} = \frac{(2*72)+(3*75)+(5*58)}{2+3+5} = \frac{144+225+290}{10} = \frac{659}{10} = 65.9$

For Student Y,

Arithmetic Mean, $\bar{X}_y = \frac{\sum X}{3} = \frac{42+52+88}{3} = \frac{200}{3} = 60.67$

Weighted Arithmetic Mean, $\bar{X}_{wY} = \frac{\sum WX}{\sum W} = \frac{(2*42)+(3*52)+(5*88)}{2+3+5} = \frac{84+156+440}{10} = \frac{680}{10} = 68$

For Student Z,

Arithmetic Mean, $\bar{X}_Z = \frac{\sum X}{3} = \frac{52+62+68}{3} = \frac{182}{3} = 60.67$

Weighted Arithmetic Mean, $\bar{X}_{wZ} = \frac{\sum WX}{\sum W} = \frac{(2*52)+(3*62)+(5*68)}{2+3+5} = \frac{104+186+340}{10} = \frac{630}{10} = 63$

1.6 MEDIAN

Median is a middle value in the distribution. Median is a numeric value that separates the higher half of a set from the lower half. It is the value that the number of observations above it is equal to the number of observations below it. The median is thus a positional average.

For example, if the salary of five employees is 6100, 7150, 7250, 7500 and 8500 the median would be 7250.

When odd number of observations are there then the calculations of median is simple. When an even number of observations are given, there is no single middle position value and the median is taken to be the arithmetic mean of two middlemost items.

In the above example we are given the salary of six employees as 6100, 7150, 7250, 7500, 8500 and 9000, the median salary would be
Median = $\frac{7250+7500}{2} = \frac{14750}{2} = 7375$

Hence, in case of even number of observations median may be found by averaging two middle position values.

Calculations of Median – Individual Observations:

Arrange the data in ascending or descending order of magnitude.

In a group composed of an odd number of values, add 1 to the total number of values and divide by 2 gives median value.
Median = size of $\frac{N+1}{2}$ th item

Ex. 13 From the following data, compute the median:
15, 9, 7, 23, 25, 25, 42, 25, 16, 14, 58, 25, 31

Sol: Arrange the numbers in ascending order 7, 9, 14, 15, 16, 23, 25, 25, 25, 25, 31, 42, 58
Median = = size of $\frac{N+1}{2}$ th item = $\frac{13+1}{2} = 7^{th}$ item = 25
∴ Median = 25

The procedure for calculating median of an even numbered of items is not as above. The median value for a group composed of an even number of items is the arithmetic mean of the two middle values – i.e. adding two values in the middle and dividing by 2

Ex. 14 From the following data, compute the median:

451, 502, 523, 512, 622, 612, 754, 732, 701, 721

Sol: Arrange the numbers in ascending order

451, 502, 512, 523, 612, 622, 701, 721, 732, 754

Median = = size of $\frac{N+1}{2}$ th item = $\frac{10+1}{2}$ = 5.5^{th} item

Size of 5.5^{th} item = $\frac{612+622}{2}$ = $\frac{1234}{2}$

∴ Median = 617

Calculations of Median – Discrete Series:

Steps:

1. First arrange the data in ascending or descending order.
2. Find out the cumulative frequencies.
3. Apply formula: Median = size of $\frac{N+1}{2}$
4. Find out total in the cumulative frequency column which is equal to $\frac{N+1}{2}$ or next higher to that value and determine the value of the variable corresponding to it. That gives the median value.

Ex. 15 From the following data, find the value of median.

Income (Rs.)	450	500	630	550	710	580
No. of persons	29	31	21	25	11	35

Sol:

Income (Rs.) Ascending order	**No. of persons f**	**Cumulative Frequency c.f.**
450	29	29
500	31	60
550	25	85
580	35	120
630	21	141
710	11	152

Median = = size of $\frac{N+1}{2}$ th item = $\frac{152+1}{2}$ = 76.5^{th} item

Size of 76.5^{th} item = Rs. 550 It is median income.

Calculations of Median – Continuous Series:

The following formula is used to calculate median for continuous series.

Median = L + $\frac{N/_2 - c.f.}{f}$ x i

L = Lower limit of median class; f = Simple freq. of the median class;
c.f. = Cumulative freq. of the preceding the median class;

i= Class interval of the median class

Ex. 16 From the following data, find the value of median.

Marks	70-80	60-70	50-60	40-50	30-40	20-30	10-20
No. of students	10	15	26	30	42	31	24

Sol: Arrange the data in ascending order

Marks	**f**	**c.f.**
10-20	24	24
20-30	31	55
30-40	42	97
40-50	30	127
50-60	26	153
60-70	15	168
70-80	10	178

Median = size of $\frac{N}{2}$ item = $\frac{178}{2}$ = 89th item

Median lies in the class 30-40 (marked in pink)

Median = L + $\frac{N/_2 - c.f.}{f}$ x i

L = 30. $\frac{N}{2}$ = 89, c.f. = 55, f = 42, i = 10

Median =30 + $\frac{89-55}{42}$ x 10

= 30 + 8.09 = 38.09

1.7 MODE

The mode or the modal value is that value in a series of observations which occurs with the greatest frequency.
For example, the mode of the values 4, 6, 9, 6, 5, 6, 9, 4 would be 6.

Calculations of Mode – Discrete Series:
Ex. 17 From the following data, find the value of mode.

Size of cloth	28	29	30	31	32	33
No. of persons wearing	15	25	45	70	55	20

Sol: The mode or modal size is 31 because the value 31 occurred maximum number of times.

Calculations of Mode – Continuous Series:

The following formula is used to calculate mode for continuous series.

Mode = $\boldsymbol{L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2}}$ **x i ,**

L = Lower limit of modal class; f_1 = freq. of the modal class;

f_o= freq. of the class preceding the modal class;

f_2= freq. of the class succeeding the modal class;

i= Class interval of the modal class

Ex. 18 From the following data, find the value of mode.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
No. of students	3	5	7	10	12	15	12	6	2	8

Sol: After observing the table, modal class is 50-60

Mode = $L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2}$ x i ,

= 50 + $\frac{15-12}{2x15-12-12}$ x 10

= 50 + $\frac{3}{6}$ x 10 = 55

1.8 EMPIRICAL RELATION BETWEEN THE MEAN, MEDIAN, AND MODE

Karl Pearson has expressed the relationship between mean, median and mode as follows:

Mode = Mean – 3 [Mean – Median]

Mode = 3 Median – 2 Mean

If we know any of the two values out of the three, we can compute third from these relationships.

1.9 GEOMETRIC MEAN

Geometric mean of a set of n observations is the nth root of their product.

G. M. = $\sqrt[n]{(X_1)(X_2)(X_3) \dots \dots (X_n)}$.

G. M. of 3 values 2, 4, 8 would be

G. M. = $\sqrt[3]{2\ X\ 4\ X\ 8}$ = $\sqrt[3]{64}$ = 4

For calculation purpose, take the logarithm of both sides

log G. M.= $\frac{log\ X_1 + log\ X_2 + \cdots \dots\dots\dots + log\ X_n}{N}$

log G. M.= $\frac{\sum \log X}{N}$ ∴ G. M. = Antilog [$\frac{\sum \log X}{N}$]

In Discrete series, G. M. = Antilog [$\frac{\sum f\ \log X}{N}$]

In Continuous series, G. M. = Antilog [$\frac{\sum f\ \log m}{N}$]

Calculations of Geometric Mean – Discrete/Individual Series:

Ex. 19 Daily income of ten families of a particular place is below. Calculate Geometric Mean.

85	70	15	75	500	8	45	250	40	36

Sol:

X	log X
85	1.9294
70	1.8451
15	1.1761
75	1.8751
500	2.6990
8	0.9031
45	1.6532
250	2.3979
40	1.6021
36	1.5563
	$\sum$ log X 17.6373

G. M. = Antilog $[\frac{\sum \log X}{N}]$

= Antilog $[\frac{\sum 17.6373}{10}]$ = Antilog (1.7637) = 58.03

Calculations of Geometric Mean – Continuous Series:

Ex 20. Calculate Geometric Mean from following data.

Marks	4-8	8-12	12-16	16-20	20-24	24-28	28-32	32-36	36-40
Frequency	8	12	20	30	15	12	10	6	2

Sol:

Marks	**m.p (m)**	**f**	**log m**	**f log m**
4-8	6	8	0.7782	6.2256
8-12	10	12	1.0000	12.0000
12-16	14	20	1.1461	22.922
16-20	18	30	1.2553	37.6590
20-24	22	15	1.3424	20.1360
24-28	26	12	1.4150	16.9800
28-32	30	10	1.4771	14.7710
32-36	34	6	1.5315	9.1890
36-40	38	2	1.5798	3.156
		N= 115		$\sum$ f log m= 143.0386

G. M. = Antilog $[\frac{\sum f \log m}{N}]$

= Antilog $[\frac{\sum 143.0386}{115}]$

= Antilog (1.2438) = 17. 53

1.10 HARMONIC MEAN

Harmonic mean of a number of observations, none of which is zero, is the reciprocal of the arithmetic mean of the reciprocals of the given values. Thus, harmonic mean (H. M.) of n observations x_i, i = 1, 2, ….,n is given by,

H. M. = $\frac{1}{\frac{1}{n}\sum_1^n \frac{1}{x_i}} = \frac{N}{\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}.....+\frac{1}{x_n}}$

Calculations of Harmonic Mean – Individual Observations:

Ex. 21 Find the harmonic mean of 4, 36, 45, 50, 75.

Sol: H. M. = $\frac{N}{\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}.....+\frac{1}{x_n}} = \frac{5}{\frac{1}{4}+\frac{1}{36}+\frac{1}{45}+\frac{1}{50}+\frac{1}{75}} = \frac{5}{\frac{1}{3}} = 15$

Calculations of Harmonic Mean – Discrete Series:

Formula for harmonic mean in Discrete series,

H. M. = $\frac{N}{\sum [f X \frac{1}{X}]} = \frac{N}{\sum [\frac{f}{X}]}$

Ex. 22 From the following data, Find the harmonic mean.

Marks	10	20	30	40	50
No. of students	20	40	60	30	10

Sol:

Marks X	f	f/X
10	20	2
20	40	2
30	60	2
40	30	0.75
50	10	0.20
	N = 160	$\sum f/X = 6.95$

H. M. = $\frac{N}{\sum [\frac{f}{X}]} = \frac{160}{6.95} = 23.0215$

Calculations of Harmonic Mean – Continuous Series:

Formula for harmonic mean in continuous series,

H. M. = $\frac{N}{\sum [\frac{f}{m}]}$

Ex. 23 From the following data, compute the value of harmonic mean.

Class interval	10-20	20-30	30-40	40-50	50-60
Frequency	6	8	12	9	5

Sol:

Class Interval	**Mid point (m)**	**f**	**f/m**
10 – 20	15	6	0.40
20 – 30	25	8	0.32
30 – 40	35	12	0.3428
40 – 50	45	9	0.2
50 - 60	55	5	0.0909
		N = 40	$\sum f/m = 1.3537$

H. M. = $\frac{N}{\sum [\frac{f}{m}]} = \frac{40}{1.3537} = 29.54$

1.11 RELATION BETWEEN THE ARITHMETIC, GEOMETRIC AND HARMONIC MEAN

Arithmetic mean is greater than geometric mean and geometric mean is greater than harmonic mean.

A.M. $\geq$ G. M $\geq$ H. M.

The quality signs hold only if all the numbers X_1, X_2, X_3,…. X_n are identical.

1.12 ROOT MEAN SQUARE

The root mean square (RMS) is defined as the square root of the mean square (the arithmetic mean of the squares of a set of numbers). It is also called as the Quadratic average. Sometimes it is denoted by $\sqrt{\bar{X}^2}$ and given by,

RMS = $\sqrt{\bar{X}^2} = \sqrt{\frac{\sum_{i=1}^{n} {X_i}^2}{N}} = \sqrt{\frac{\sum X^2}{N}}$

It is very useful in fields that study sine waves like electrical engineering.

Ex. 24 Find RMS of 1, 3, 5, 7 and 9

Sol: RMS = $\sqrt{\frac{\sum X^2}{N}}$

$= \sqrt{\frac{1^2+3^2+5^2+7^2+9^2}{5}} = \sqrt{\frac{1^2+3^2+5^2+7^2+9^2}{5}} = \sqrt{53} = 7.28$

1.13 QUARTILES, DECILES AND PERCENTILES

From the definition of median that it's the middle point which divides the set of ordered data into two equal parts. In the same way we can divide the set into four equal parts and this called quartiles. These values denoted by Q_1, Q_2 and Q_3, are called the first, second and the third quartile respectively. In the same way the values that divide the data into 10 equal parts are called deciles and are denoted by D_1, D_2, …., D_9 whereas the values dividing the data into 100 equal parts are called percentiles and are denoted by P_1, P_2, …., P_{99}. The fifth decile and the 50^{th} percentile corresponds to median.

Formulae:

Quartile:

For individual observations, $Q_i = (\frac{i}{4})$. No. of observation, i= 1, 2, 3

For discrete series, $Q_i = (\frac{i}{4})$. N, N= $\sum f$ and i= 1, 2, 3

For continuous series, $Q_i = L + \frac{\frac{iN}{4} - cf}{f}$. c,

Where, i= 1, 2, 3, c = size of class interval.

L = Lower limit of the class interval in which lower quartile lies,

f = freq. of the interval in which lower quartile lies,

cf = cumulative freq. of the class preceding the quartile class,

Deciles:

For individual observations, $D_i = (\frac{i}{10})$. No. of observation, i= 1, 2, …, 9

For discrete series, $D_i = (\frac{i}{10})$. N, N= $\sum f$ and i= 1, 2,…., 9

For continuous series, $D_i = L + \frac{\frac{iN}{10} - cf}{f}$. c, i= 1, 2, …, 9

Percentiles:

For individual observations, $P_i = (\frac{i}{100})$. value of observation, i= 1,2,…, 99

For discrete series, $P_i = (\frac{i}{100})$. N, N= $\sum f$ and i= 1, 2,…., 99

For continuous series, $P_i = L + \frac{\frac{iN}{100} - cf}{f}$. c, i= 1, 2, …, 99

Ex. 25 Find the quartiles Q_1, Q3, D_1, D_5, D_8, P_8, P_{50} and P_{85} of the following data 20, 30, 25, 23, 22, 32, 36.

Sol: Arrange data in ascending order, n = 7 i.e. odd number
20, 22, 23, 25, 30, 32, 36

$q_1 = (\frac{1}{4}).7 = 1.75$ $\therefore q_1 = 2$ $\therefore$ **Q_1= 22**

$q_3 = (\frac{3}{4}).7 = 5.75$ $\therefore q_2 = 6$ $\therefore$ **Q_3= 32**

$d_1 = (\frac{1}{10}).7 = 0.7$ $\therefore d_1 = 1$ $\therefore$ **D_1= 20**

$d_5 = (\frac{5}{10}).7 = 3.5$ $\therefore d_5 = 4$ $\therefore$ **D_5= 25**

$d_8 = (\frac{8}{10}).7 = 5.6$ $\therefore d_8 = 6$ $\therefore$ **D_8= 32**

$p_8 = (\frac{8}{100}).7 = 0.56$ $\therefore p_8 = 1$ $\therefore$ **P_8= 20**

$p_{50} = (\frac{50}{100}).7 = 3.5$ $\therefore p_{50} = 4$ $\therefore$ **P_{50}= 25**

$p_{85} = (\frac{85}{100}).7 = 5.95$ $\therefore p_{85} = 6$ $\therefore$ **P_{85}= 32**

Ex. 26 Find Q_1, Q_3, D_4, P_{27} for the following data.

X	0	1	2	3	4	5	6	7	8
f	1	9	26	59	72	52	29	7	1
c.f.	1	10	36	95	167	219	248	255	256

Sol. We know that, $Q_i = (\frac{i}{4})$. N

Q_1 = $(\frac{1}{4}).256 = 64$ and c.f. just greater than 64 is 95. Hence **Q_1 = 3**

Q_3 = $(\frac{3}{4}).256 = 192$ and c.f. just greater than 192 is 219. Hence **Q_3 = 5**

D_4 = $(\frac{4}{10}).256 = 102.4$ and c.f. just greater than 102.4 is 167. Hence **D_4 = 4**

P_{27} = $(\frac{27}{100}).256 = 69.12$ and c.f. just greater than 69.12 is 95. Hence **P_{27} = 3**

Ex. 27 Find Q_1, Q_3, D_2, P_{90} for the following data.

Marks	Below 10	10-20	20- 40	40-60	60-80	Above 80
No. of students	8	10	22	25	10	5

Sol: We know that, $Q_i = L + \frac{\frac{iN}{4} - cf}{f}$. c,

Marks	Below 10	10-20	20- 40	40-60	60-80	Above 80
f	8	10	22	25	10	5
cf	8	18	40	65	75	80

Q_1 = Size of (N/4)th item = size of (80/4)= 20th item. Q_1 lies in the class 20-40.

L=20, N/4 = 20, cf = 18, f = 22 and c = 20

$Q_1 = 20 + \{(20 – 18)/22\}* 20 = 20 + 1.82 = 21.82$

Q_3 = Size of (3N/4)th item = size of (3*80/4)= 60th item. Q_3 lies in the class 40-60.

L=40, 3N/4 = 60, cf = 40, f = 25 and c = 20

$Q_3 = 40 + \{(60 – 40)/25\}* 20 = 56$

D_2 = Size of (2N/10)th item = size of (2*80/10)= 16th item. D_2 lies in the class 10-20

L=10, 2N/10 = 16, cf=8, f = 10 and c=10

$D_2 = 10 + \{(16 –8)/10\}*10 = 18$

P_{90} = Size of (90N/100)th item = size of (90*80/10)= 72th item. P_{90} lies in the class 60-80.

L=60, 90N/100 = 72, cf=65, f = 10 and c=20

$P_{90} = 60 + \{(72 –65)/10\}*20 = 74.$

1.14 SOFTWARE AND MEASURES OF CENTRAL TENDENCY

There are many software available to calculate measures of central tendency. We can use Excel to calculate the standard measures of central tendency (mean, median and mode). In Microsoft Excel, the mean can calculated by using one of the functions like AVERAGE, AVERAGEA, AVERAGEIF, AVERAGEIFS. The mean can be calculated by using the MEDIAN function. We can calculate a mode by using the MODE function, GEOMEAN to calculate geometric mean and HARMEAN to calculate harmonic mean.

We can use SPSS to calculate the standard measures of central tendency (mean, median and mode). We can get SPSS to compute mean, median and mode in the command submenu. Go to the Statistics menu, select the Analyse submenu, and then the Descriptive Statistics submenu and then the Frequencies option. We can use MINITAB to calculate the standard measures of central tendency using the functions Mean, Median, Mode ang GMEAN. To compute these go to Stat-Tables-Descriptive statistics.
Using R software one could easily obtain the value of the mean using summary function.

We could find median value using summary function in R. The randomForest library can be used to impute the missing values using Median for numeric variables. Mode is used for missing value imputation for categorical variables using randomForest library in R. Model can be easily located graphically. You shouldn’t be surprised that the R’s mode

function (mode ()) does not provide a model value. It shows the datatype of the particular variable which does not comply with our standard expectation. So how one would find mode using R software? We need to use table function for finding mode. As you know the table function in R provides frequency distribution of the variable. Thus the value with highest frequency is a modal value.

Geometric mean is the only average that is recommended for finding average growth (decline) rates. It is defined as the nth root of the product of n terms. Since it is defined in product terms so the observation shouldn't be having zero or negative values. We don't have a built-in function in R for its computation but one could find it by using its formula directly in R platform.

1.15 SUMMARY

A measure of central tendency is a measure that tells us where the middle of a group of data lies. Mean, median and mode are the most important measures of central tendency. The complete dataset may be represented by these values. It is not necessary for mean, median and mode to have the same values. Mean is sensitive to extreme data values. Median is a better way to understand skewed distribution than mean. It is possible that there is no mode in the data. Mean and median cannot be zero unless all data values are zero.

1.16 EXERCISE

1. Find the arithmetic mean of the following distribution:

X	10	30	50	70	89
f	7	8	10	15	10

2. Find the arithmetic mean of the following distribution:

X	3	9	12	14	15	17
f	1	3	4	1	4	2

3. Find the arithmetic mean of the following data.

Class Interval	15-25	25-35	35-45	45-55	55-65	65-75	75-85
Frequency	6	11	7	4	4	2	1

4. Find the arithmetic mean of the following data.

Class Interval	10-20	20-30	30-40	40-50	50-60
Frequency	30	27	14	17	2

5. Obtain the median for the following frequency distribution:

X	1	2	3	4	5	6	7	8	9
f	8	10	11	16	20	25	15	9	6

[Ans: Median = 5]

6. Obtain the median from the following data.

X	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60
f	35	45	70	105	90	74	51	30

7. Find the mode for the following distribution.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of students	5	8	7	12	28	20	10	10

[Ans: Mode = 46.67]

8. Calculate Geometric Mean from following data.

125	1462	38	7	0.22	0.08	12.75	0.5

[Ans: 6.952]

9. Find the geometric mean, harmonic mean and root mean square of the numbers 3, 5, 6, 6, 7, 10 and 12.
   [Ans: G. M. = 6.43, H. M. = 5.87, RMS = 7.55]

10. Find the arithmetic mean, geometric mean, harmonic mean of numbers 2, 4 and 8. Check the relation between them.

11. Calculate Quartile 3, Deciles -7 and Percentiles 20 from following data.

Class	2 - 4	4 – 6	6 – 8	8 - 10
Frequency	3	4	2	1

12. Calculate Q_1, Q_2, Q_3 D_1, D_5, D_9, P_{11}, P_{65} from following data.

Wages	**No. of employees**
250.00 – 259.99	8
260.00 – 269.99	10

270.00 – 279.99	16
280.00 – 289.99	14
290.00 – 299.99	10
300.0 – 309.99	5
310.00 – 319.99	2

1.16 REFERENCES

- FUNDAMENTAL OF MATHEMATICAL STATISTICS by S. C. Gupta and V. K. Kapoor
- Statistical Methods by S. P. Gupta
- STATISTICS by Murray R. Spiegel, Larry J. Stephens

*****

2

THE STANDARD DEVIATION AND OTHER MEASURES OF DISPERSION

Unit structure

2.0 OBJECTIVES

After going through this chapter, students will able to learn

- To provide the importance of the concept of dispersion
- To calculate range, semi-Interquartile range, mean deviation
- To explain why measures of dispersion must be reported in addition to measures of central tendency
- To calculate standard deviation, variance, standard scores
- To trace precise relationship
- To compare two or more series with regard to their variability

2.1 INTRODUCTION

The measures of central tendency or Averages give us an idea of the concentration of the observations about the central part of distribution. But the average alone cannot adequately describe a set of observations. They must be supported and supplemented by some other measures, called Dispersion.

2.2 DISPERSION OR VARIATION

Literal meaning of dispersion is 'scatteredness'. In two or more distributions the central value may be the same but still there can be wide differences in the formation of distribution. Measures of dispersion help us in studying this important characteristic of a distribution.

Definitions of Dispersion:

1. "Dispersion is the measure of the variation of the items." – A. L. Bowley
2. "Dispersion is the measure of extent to which individual item vary." – L. R. Connor
3. "The degree to which numerical data tend to spread about an average value is called variation or dispersion of the data". – Spiegel

2.3 RANGE

Range is the difference between two extreme observations of the distribution. Symbolically,

Range = L – S, where L = Largest item, S = smallest item

The relative measure corresponding to range, called the coefficient of range.

Coefficient of range = $\frac{L-S}{L+S}$

Since range is based on two extreme observations, it is not at all a reliable measure of dispersion.

Ex 1. From the following data, calculate range and coefficient of range.

Day	Mon	Tues	Wed	Thurs	Fri	Sat
Price	20	21	18	16	22	25

Sol: Range = L – S = 25 – 16 = 9

Coefficient of range = $\frac{L-S}{L+S}$

$= \frac{25-16}{25+16} = \frac{9}{41} = 0.21$

For continuous series, find the difference between the upper limit of the highest class and the lower limit of the lowest class.

Ex 2. . From the following data coefficient of range.

Marks	10– 20	20 -30	30-40	40-50	50-60
No. of Students	10	12	14	8	6

Sol: Coefficient of range = $\frac{L-S}{L+S}$

$= \frac{60-10}{60+10} = \frac{50}{70} = 0.21$

2.4 SEMI-INTERQUARTILE RANGE OR QUARTILE DEVIATION

Semi-Interquartile Range Or Quartile Deviation is given by,

Q. D. $= \frac{Q_3 - Q_1}{2}$

Quartile Deviation is a better measure than a range as it makes use of 50% of the data. But since it ignores the other 50% of the data, it cannot be considered as a reliable measure.

Q. D. $= \frac{Q_3 - Q_1}{2}$

The relative measure corresponding to Q. D., called the coefficient of Q. D.

Coefficient of Q. D. $= \frac{Q_3 - Q_1/2}{Q_3 + Q_1/2} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$

Coefficient of Q. D. can be used to compare the degree of variation in different distributions.

Computation of Quartile Deviation - Individual Observations:

Ex. 3 Find out Quartile Deviation and Coefficient of Quartile Deviation from following data.

25 33 45 17 35 20 55

Sol: Arrange the data in ascending order:

17 20 25 33 35 45 55

Q_1 = size of $[\frac{N+1}{4}]$ th item = size of $[\frac{7+1}{4}]$ th item = 2^{nd} item

$\therefore Q_1 = 20$

Q_3 = size of $3[\frac{N+1}{4}]$ th item = size of $3[\frac{7+1}{4}]$ th item = 6^{th} item

$\therefore Q_3 = 45$

Q. D. $= \frac{Q_3 - Q_1}{2} = \frac{45-20}{2} = 12.5$

Coefficient of Q. D. $= \frac{Q_3 - Q_1}{Q_{3+} Q_1} = \frac{45-20}{45+20} = \frac{25}{65} = 0.455$

Computation of Quartile Deviation -Discrete Series:

Ex. 4 Find out Quartile Deviation and Coefficient of Quartile Deviation from following data.

Marks	10	20	30	40	50	60
No. of Students	7	10	18	12	10	6

Sol:

Marks	10	20	30	40	50	60
Frequency f	7	10	18	12	10	6
cf	7	17	35	47	57	63

Q_1 = size of $[\frac{N+1}{4}]$ th item = size of $[\frac{63+1}{4}]$ th item = 16^{th} item

$\therefore Q_1 = 20$

Q_3 = size of $3[\frac{N+1}{4}]$ th item = size of $3[\frac{63+1}{4}]$ th item = 48^{th} item

$\therefore Q_3 = 50$

Q. D. $= \frac{Q_3 - Q_1}{2} = \frac{50-20}{2} = 15$

Coefficient of Q. D. $= \frac{Q_3 - Q_1}{Q_{3+} Q_1}$

$= \frac{50-20}{50+20} = \frac{30}{70} = 0.4285$

Computation of Quartile Deviation - Continuous Series:

Ex. 5 Find out Quartile Deviation and Coefficient of Quartile Deviation from following data.

Marks	35-44	45 - 54	55- 64	65 - 74	75 - 84
No. of Students	12	40	33	13	12

Sol:

Marks	35-44	45 - 54	55- 64	65 - 74	75 - 84
Frequency f	12	40	33	13	12
cf	12	52	75	88	100

Q_1 = size of $[\frac{N}{4}]$ th item = size of $[\frac{100}{4}]$ th item = 25^{th} item

∴ Q_1 lies in the class 45 – 54

$Q_1 = L + \frac{N/4 - c.f.}{f} * i$

L = 45, $N/4$ = 25, c.f. = 12 [c.f. of previous class], f= 40, i = 9

$Q_1 = 45 + \frac{25 - 12}{40} * 9 = 47.925$

Q_3 = size of $3[\frac{N}{4}]$ th item = size of $3[\frac{100}{4}]$ th item = 75^{th} item

∴ Q_3 lies in the class 55-64

$Q_3 = L + \frac{3N/4 - c.f.}{f} * i$

L = 55, $3N/4$ = 75, c.f. = 52 [c.f. of previous class], f= 33, i = 9

$Q_3 = 55 + \frac{75 - 52}{33} * 9 = 61.2727$

Q. D. $= \frac{Q_3 - Q_1}{2}$

$= \frac{61.2727 - 47.925}{2} = 6.67$

Coefficient of Q. D. $= \frac{Q_3 - Q_1}{Q_3 + Q_1}$

$= \frac{61.2727 - 47.925}{61.2727 + 47.925} = \frac{6.67}{109.1977} = 0.061$

2.5 MEAN DEVIATION

Mean deviation is also known as the average deviation.

If xi | fi , i = 1, 2, …, n is the frequency distribution, then mean deviation from the average A (usually mean, median or mode).

Since mean deviation is based on all the observations, it is a better measure of dispersion than range and quartile deviation.

Note: Mean deviation is least when taken from median

The relative measure corresponding to the mean deviation called the coefficient of mean deviation and is obtained by,

Coefficient of M. D. = $\frac{M.D.}{Median}$

Computation of Mean deviation – Individual observations

M. D. = $\frac{1}{n}\sum|X-A|$

= $\frac{1}{N}\sum|D|$, where $|D| = |X-A|$ is the modulus value or absolute value of the deviation ignoring plus and minus signs.

Ex. 6 Calculate mean deviation and coefficient of mean deviation from following data:

600, 620, 640, 660, 680

Sol: From above data, Median = 640

Data	**Deviation from median 640 $\|D\|$**
600	40
620	20
640	0
660	20
680	40
N= 5	$\sum\|D\|$ =120

M. D. = $\frac{1}{N}\sum|D| = \frac{120}{5} = 24$

Coefficient of M. D. = $\frac{M.D.}{Median}$

= $\frac{24}{640}$

= 0.0375

Computation of Mean deviation – Discrete series:

M. D. = $\frac{1}{N}\sum f|D|$, where $|D| = |X-A|$

Ex. 7 Calculate mean deviation from following data.

X	20	21	22	23	24
f	6	15	21	15	6

Sol:

X	f	c.f	$\|D\|$	f $\|D\|$
20	6	6	2	12

21	15	21	1	15
22	21	42	0	0
23	15	57	1	15
24	6	63	2	12
	N = 63			$\sum f\|D\|$=54

Median = size of $\frac{N+1}{2}$ th item = size of $\frac{63+1}{2}$ th item = 32th item

Size of 32th item is 22, hence Median = 22

M. D. = $\frac{1}{N}\sum f|D|$

= $\frac{54}{63}$ = 0.857

Computation of Mean deviation – Continuous series:

Here we have to obtain the mid-point of the various classes and take deviations of these points from median. Formula is same.

M. D. = $\frac{1}{N}\sum f|D|$

Ex. 8 Calculate mean and mean deviation from following data.

Size	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	7	12	18	25	16	14	8

Sol:

Size	f	c.f.	m.p (m)	$\|m - 35.2\|$ $\|D\|$	f $\|D\|$
0-10	7	7	5	30.2	211.4
10-20	12	19	15	20.2	242.4
20-30	18	37	25	10.2	183.6
30-40	25	62	35	0.2	5.0
40-50	16	78	45	9.8	156.8
50-60	14	92	55	19.8	277.2
60-70	8	100	65	29.8	238.4
	N= 100				$\sum f\|D\|$=1314.8

Median = size of $\frac{N}{2}$ th item = size of $\frac{100}{2}$ th item = 50th item

Median lies in the class 30 – 40

Median = L + $\frac{N/2 - c.f.}{f}$ * i

L= 30, $N/2$= 50, c.f. = 37, f = 25, i = 10

Median = 30 + $\frac{50-37}{25}$ *10 = 35.2

M. D. = $\frac{1}{N}\sum f|D|$

= $\frac{1314.8}{100}$ =13.148

2.6 10–90 PERCENTILE RANGE:

The 10 – 90 percentile range of a set of data is defined by,
10 – 90 percentile range = $P_{90} - P_{10}$
Where P_{10} and P_{90} are the 10^{th} and 90^{th} for the data.
Semi 10-90 percentile range = $\frac{P_{90}-P_{10}}{2}$

2.7 STANDARD DEVIATION:

Standard deviation is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean.

Standard deviation is also known as root mean square deviation as it is the square root of the mean of the standard deviation from arithmetic mean. Standard deviation is denoted by the small Greek letter σ (read as sigma).

Calculation of Standard Deviation - Individual Observations:

$\sigma = \sqrt{\frac{\sum x^2}{N}}$, where x = (X -$\bar{X}$)

Calculation of Standard Deviation - Discrete Series:

$\sigma = \sqrt{\frac{\sum fx^2}{N}}$, where x = (X -$\bar{X}$)

Calculation of Standard Deviation: Continuous Series:

$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} * i$, where d =$\frac{(m-A)}{i}$, i = class interval

Ex. 9 Calculate mean and standard deviation from the following data.

Size	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	7	10	32	43	50	35	23

Sol:

Marks	m. p. (m)	f	d = (m-35)/10	d^2	fd	fd^2
0-10	5	7	-3	9	-21	63
10-20	15	10	-2	4	-20	40
20-30	25	32	-1	1	-32	32
30-40	35	43	0	0	0	0
40-50	45	50	1	1	50	50
50-60	55	35	2	4	70	140

60-70	65	23	3	9	69	207
		N = 200			$\sum fd$ =116	$\sum fd^2$= 532

Assumed mean, A= 35

$\bar{X}$ = A + $\frac{\sum fd}{N} * i$

=35 + $\frac{116}{200}$ *10 = 40.8

$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} * i$

$= \sqrt{\frac{532}{200} - \left(\frac{116}{200}\right)^2} * 10$

$= \sqrt{2.66 - 0.3364}$ *10 = 1.5243*10 = 15.243

2.8 SHORT METHODS FOR COMPUTING THE STANDARD DEVIATION:

Calculation of Standard Deviation- Individual Observations:
When actual mean is in fractions eg 568.245, it would be too bulky to do calculations. In such case either the mean may be approximated or the deviations be taken from assumed mean A. Following is formula if we take deviations from assumed mean A:

$\sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2}$, where d = (X – A)

Ex. 10 Calculate standard deviation with the help of assumed mean.

340, 360, 390, 345, 355, 388, 372, 363, 277, 351

Sol: Consider assumed mean = 364

X	d = (X – 364)	d^2
340	-24	576
360	-4	16
390	26	676
345	-19	361
355	-9	81
388	24	576
372	8	64
363	-1	1
377	13	169
351	-13	169
	$\sum$ d = 1	$\sum d^2$ = 2689

$\sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2}$

$= \sqrt{\frac{2689}{10} - \left(\frac{1}{10}\right)^2} = 16.398$

Calculation of Standard Deviation - Discrete Series:

Assumed mean method: $\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$, where d = (X – A)

Ex. 11 Calculate standard deviation from the following data.

Size	3.5	4.5	5.5	6.5	7.5	8.5	9.5
Frequency	4	8	21	60	85	30	9

Sol:

Size	f	d = (X-6.5)	d^2	fd	fd^2
3.5	4	-3	9	-12	36
4.5	8	-2	4	-16	32
5.5	21	-1	1	-21	21
6.5	60	0	0	0	0
7.5	85	1	1	85	85
8.5	30	2	4	60	120
9.5	9	3	9	27	81
	N = 217			$\sum$fd = 123	$\sum fd^2$= 375

Assumed mean, A= 6.5

: $\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$

$= \sqrt{\frac{375}{217} - \left(\frac{123}{217}\right)^2} = \sqrt{1.7281 - 0.3212} = 1.1861$

2.9 PROPERTIES OF THE STANDARD DEVIATION

1. Combined standard deviation: We can compute combined standard deviation of two or more groups. It is denoted by σ_{12} and given by

$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2\sigma_2^2}{N_1 + N_2}}$

Where σ_{12} = combined standard deviation;

σ_1 = standard deviation of first group;

σ_2 = standard deviation of second group;

$d_1 = |\overline{X_1} - \overline{X_{12}}|$;

$d_2 = |\overline{X_2} - \overline{X_{12}}|$

2. The standard deviation of the first n natural numbers can obtained by,

$\sigma = \sqrt{\frac{1}{12}\ (N^2 - 1)}$

Thus the standard deviation of natural numbers 1 to 20 will be

$\sigma = \sqrt{\frac{1}{12}\ (20^2 - 1)} = \sqrt{\frac{1}{12}\ 399} = 5.76$

3. Standard deviation is always computed from the arithmetic mean because the sum of the squares of the deviations of items from their arithmetic mean is minimum.

4. For normal distribution,

Mean ± 1 σ covers 68.27% of the items.

Mean ± 2 σ covers 95.45% of the items.

Mean ± 3 σ covers 99.73% of the items.

2.10 VARIANCE

The square of standard deviation is called the variance and is given by,

Variance = $\frac{\sum(X - \overline{X})^2}{N}$

i.e. Variance = σ^2 or $\sigma = \sqrt{Variance}$

In the frequency distribution where deviations are taken from assumed mean,

Variance = $\left\{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2\right\}$* i^2, where d = $\frac{(X-A)}{i}$ and i = class interval

Ex. 12 Calculate standard deviation from the following data.

Marks	10-20	20- 30	30-40	40-50	50-60	60-70
No. of students	2	6	8	12	7	5

Sol:

Marks	m.p (m)	f	d=(m-35)/10	d^2	fd	fd^2
10-20	15	2	-2	4	-4	8
20-30	25	6	-1	1	-6	6
30-40	35	8	0	0	0	0
40-50	45	12	1	1	12	12
50-60	55	7	2	4	14	28
60-70	65	5	3	9	15	45
		N = 40			$\sum fd$ = 31	$\sum fd^2$= 99

Variance = $\left\{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2\right\} * i^2$

$= \left\{\frac{99}{40} - \left(\frac{31}{40}\right)^2\right\} * 10^2$

$= (2.475 - 0.6006) * 100 = 187.44$

2.11 CHARLIE'S CHECK

Some error may be made while calculating the value of mean and standard deviations using different method. The accuracy of calculations can be checked by using following formulae.

$\sum f\ (u + 1) = \sum fu + \sum f = \sum fu + N$

$\sum f\ (u + 1)^2 = \sum f(u^2 + 2u + 1) = \sum fu^2 + 2\sum fu + \sum f = \sum fu^2 + 2\sum fu + N$

$\sum f\ (u + 1)^3 = \sum fu^3 + 3\sum fu^2 + 3\sum f\ u\ + N$

Ex. 13 Use Charlier's check to verify mean and the standard deviation.

Size	20-30	30-40	40-50	50-60	60-70	70-80	80-90
Freq	9	12	8	10	11	35	15

Sol:

X	f	m. p. (m)	u= (m-55)/i	u+1	f(u+1)	u^2	fu	$\mathbf{fu^2}$
20-30	9	25	-3	-2	-18	9	-27	81
30-40	12	35	-2	-1	-12	4	-24	48
40-50	8	45	-1	0	0	1	-8	8
50-60	10	55	0	1	10	0	0	0
60-70	11	65	1	2	22	1	11	11
70-80	35	75	2	3	105	4	70	140
80-90	15	85	3	4	60	9	45	135
	N=∑f =100				∑ f(u+1)= 167		∑fu =67	$\sum fu^2$ =423

$\sum f\ (u + 1) = 167$
$\sum fu + N = 67 + 100 = 167$
$\therefore \sum f\ (u + 1) = \sum fu + N$
This provides the required check on the mean.

X	f	m. p. (m)	u= (m-55)/i	u+1	f(u+1)	$f(u+1)^2$
20-30	9	25	-3	-2	-18	36
30-40	12	35	-2	-1	-12	12
40-50	8	45	-1	0	0	0

50-60	10	55	0	1	10	10
60-70	11	65	1	2	22	44
70-80	35	75	2	3	105	315
80-90	15	85	3	4	60	240
	N= ∑f =100				∑ f(u+1) = 167	∑ $f(u+1)^2$ ==657

$\sum f\ (u + 1)^2 = 657$

$\sum f u^2 + 2 \sum f u + N = 423 + 2*67 + 100 = 657$

$\therefore \sum f\ (u + 1)^2 = \sum f u^2 + 2 \sum f u + N$

This provides the required check on the standard deviation**.**

2.12 SHEPPARD'S CORRECTION FOR VARIANCE

The computation of the standard deviation is somewhat in error as a result of grouping the data into classes (grouping error). To adjust for grouping error, we use the formula,

Corrected variance = variance from grouped data - $\frac{i^2}{12}$

Where i is the class interval size. The correction $\frac{i^2}{12}$ is called Sheppard's correction. It is used for distribution of continuous variables where the tails tends to zero in both direction.

Ex. 14 Apply Sheppard's Correction to determine the standard deviation of the data in Ex. 8

Sol: $\sigma = 15.243 \therefore \sigma^2 = 232.349$ and i= 10.

Corrected variance = variance from grouped data - $\frac{i^2}{12}$

$= 232.349 - \frac{10^2}{12} = 224.016$

Corrected Standard deviation = $\sqrt{224.016} = 14.9671$

2.13 EMPIRICAL RELATIONS BETWEEN MEASURES OF DISPERSION:

There is a fixed relationship between the three measures of dispersion in normal distribution.

Q. D. = $\frac{2}{3}\sigma$ or $\sigma = \frac{3}{2}$ Q. D and

M. D. = $\frac{4}{5}\sigma$ or $\sigma = \frac{4}{5}$ M. D

The quartile deviation is smallest, the mean deviation next and the standard deviation is largest.

2.14 ABSOLUTE AND RELATIVE DISPERSION:

Measures of dispersion may be either absolute or relative. Absolute measures of dispersion are expressed in the same statistical unit in which the original data are given such as kilograms, tons, rupees etc. These values may be used to compare the variations in two distributions provided the variables are expressed in the same units and of the same average size. In case the two sets of data are expressed in different units such as quintals of sugar versus tons of sugarcane, the absolute measures of dispersion are not comparable. In such cases measures of relative dispersion is used.
A measure of relative dispersion is the ratio of a measure of absolute dispersion to an appropriate average. It is sometimes called coefficient of dispersion.

Relative dispersion = $\frac{absolute\ dispersion}{average}$

2.15 COEFFICIENT OF VARIATION:

Coefficient of is used in problems where we want to compare the variability of two or more than two series. That series or group for which the coefficient of variation is greater is said to be more variable or less consistent, less uniform, less stable or less homogeneous. The series for which the coefficient of variation is less is said to be less variable or more consistent, more uniform, more stable or more homogeneous.

If the absolute dispersion is standard deviation σ and if average is the mean $\bar{X}$, then relative dispersion is called coefficient of variation, it is denoted by C. V. and is given by,

Coefficient of variation (C.V.) = $\frac{\sigma}{\bar{X}}$ x100

Ex. 15 Calculate arithmetic mean, standard deviation and coefficient of variation.

Class	23-27	28-32	33-37	38-42	43-47	48-52	53-57	58-62	63-67	68-72
Freq	2	6	7	12	18	13	9	7	4	2

Sol:

Class	m. p. (m)	f	d = (m-50)/5	d^2	fd	fd^2
23-27	25	2	-5	25	-10	50
28-32	30	6	-4	16	-24	96
33-37	35	7	-3	9	-21	63
38-42	40	12	-2	4	-24	48
43-47	45	18	-1	1	-18	18
48-52	50	13	0	0	0	0
53-57	55	9	1	1	9	9
58-62	60	7	2	4	14	28
63-67	65	4	3	9	12	36
68-72	70	2	4	16	8	32
		N = 80			$\sum$fd = -44	$\sum fd^2$= 380

Mean $(\bar{X})$ = A + $\frac{\sum fd}{N}$ x i

= 50 + $\frac{-44}{80}$ x 5 = 47.25

S. D. (σ) = $\sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$ x i

= $\sqrt{\frac{380}{80} - \left(\frac{-44}{80}\right)^2}$ x 5

= $\sqrt{4.75 - 0.3025}$ x 5= 22.2375

C. V. = $\frac{\sigma}{\bar{X}}$ x100

= $\frac{22.2375}{47.25}$ x100 = 47.06%

2.16 STANDARDIZED VARIABLE AND STANDARD SCORES

The variable that measures the deviation from the mean in units of the standard deviation is called standardized variable, is independent of the units used and is given by,

z = $\frac{X-\bar{X}}{\sigma}$

If the deviations from the mean are given in units of the standard deviation, they are said to be expressed in standard units or standard scores. These are of great value in the comparison of distribution. The variable z is often used in educational testing, where it is called as a standard score.

Ex. 16 Your test score is 160 while the test has a mean of 120 and standard deviation of 15. If the distribution is normal, what is your z score? Explain the meaning of the result.

Sol: $z = \frac{X-\bar{X}}{\sigma} = \frac{160-120}{15} = 2.7$

The score is 2.7 standard deviations above the mean.

Ex. 17 A student received a grade of 84 on a final examination in English for which mean grade was 76 and the standard deviation was 10. On a final examination in Science for which mean grade was 82 and the standard deviation was 16, she received a grade of 90. In which subject was her relative standing higher?

Sol: Standardized variable $z = \frac{X-\bar{X}}{\sigma}$

For English, $z = \frac{84-76}{10} = 0.8$

For Science, $z = \frac{90-82}{16} = 0.5$

Thus, the student had a grade of 0.8 of a standard deviation above the mean in English but only 0.5 of a standard deviation above the mean in science. Thus her relative standing was higher in English.

2.17 SOFTWARE AND MEASURES OF DISPERSION:

The statistical software gives a variety of measures for dispersion. The dispersion measures are usually given in descriptive statistics. EXCEL and MINITAB allows for the computation of all the measures discussed above. The output from MINITAB and STATISTIX has helped clarify some of the statistical concepts which are hard to understand without some help from the graphics involved.

Calculating Range In Excel: Excel does not offer a function to compute range. However, we can easily compute it by subtracting the minimum value from the maximum value. The formula would be =MAX()-MIN() where the dataset would be the referenced in both the parentheses. The =MAX() and =MIN() functions would find the maximum and the minimum points in the data. The difference between the two is the range. Microsoft Excel has two functions to compute quartiles. The inter-quartile range has to be calculated as the difference between the quartile 3 and quartile 1 values Quartiles can be calculated using =QUARTILE.INC() or =QUARTILE.EXC(). Both functions calculate the quartiles by calculating the percentiles on the data. Excel offers two functions, =STDEV.S() for sample standard deviation, and =STDEV.P() for population standard deviation.Excel with two different functions: =VAR.P() for population variance, and =VAR.S() for sample variance. Minitab may be used to

compute descriptive statistics for numeric variables, including the mean, median, mode, standard deviation, variance and coefficient of variance. To compute these go to Stat-Tables-Descriptive statistics.

You can use SPSS to calculate the measures of dispersion such as range, semi-interquartile range, standard deviation and variance. We can get SPSS to compute these in the command submenu. Go to the Statistics menu, select the Analyse submenu, and then the Descriptive Statistics submenu and then the Frequencies option. We can use MINITAB to calculate the measures of dispersion the functions Q1, Q3, Range StDev, Variance and CorfVar

2.18 SUMMARY

A measure of dispersion indicates the scattering of data. Dispersion is the extent to which values in a distribution differ from the average of the distribution. The measure of dispersion displays and gives us an idea about the variation and the central value of an individual item. The range and interquartile range are generally ineffective to measure the dispersion of set of data. The useful measure that describes the dispersion of all the values is standard deviation or variance. Dispersion can prove very effective in association with central tendency in making any statistical decision.

2.19 EXERCISE

1. Calculate Quartile deviation (Q. D.), Mean Deviation (M. D.) from mean for the following data.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of Students	6	5	8	15	7	6	3

[Ans: Q.D. = 11.23, Mean = 33.4, M.D from mean = 13.184]

2. Calculate Mean Deviation (M. D.) from mean for the following data

Size	2	4	6	8	10	12	14	16
f	2	2	4	5	3	2	1	1

[Ans: Mean = 8, M.D from mean = 2.8]

3 Calculate Mean Deviation and its coefficient from mean for the following data.

Size	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70 -80
Freq	5	8	12	15	20	14	12	6

[Ans: Median = 43, M.D = 15.37, Coe. Of M. D. = 0.357]

4. Find the standard deviation of the following data.

i. 12, 6, 7,3,15, 10, 18, 5

ii. 9, 3, 8, 8, 9, 8, 9,18

[Ans: i. St. dev. σ= 4.87, ii. St. dev. σ= 3.87]

5. Find the standard deviation of the following data.

Age	20-25	25-30	30-35	35-40	40-45	45-50
No. of persons	170	110	80	45	40	35

Take assumed average = 32.5

[Ans: Standard deviation σ= 7.936]

6. Calculate the standard deviation from the following data by short method.

240.12, 240.13, 240.15, 240.12, 240.17, 240.15, 240.17, 240.16, 240.22, 240.21

7. Calculate standard deviation from the following data by short method.

Salary	45	50	55	60	65	70	75	80
No. of persons	3	5	8	7	9	7	4	7

[Ans: Standard deviation = 10.35]

8. Calculate arithmetic mean, standard deviation and coefficient of variation.

Class	20-25	25-30	30-35	35-40	40-45
Frequency	1	22	64	10	3

[Ans: $\bar{X}$ = 32.1, S. D. (σ) =3.441, C. V. = 10.72]

2.20 REFERENCE

- FUNDAMENTAL OF MATHEMATICAL STATISTICS by S. C Gupta and V. K. Kapoor
- Statistical Methods by S. P. Gupta
- STATISTICS by Murray R. Spiegel, Larry J. Stephens

*****

3

INTRODUCTION TO R

Unit structure

3.0 OBJECTIVES

After going through this chapter, students will able to learn

1. Understand the different data types, variables in R.
2. Understand the basics in R programming in terms of operators, control statements
3. Use of built-in and user defined function
4. Understand the different data structures in R.

3.1 INTRODUCTION

R is programming language and software environment for statistical computing and graphics. It is an open source programming language. It was designed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand.in 1993. It was released on 31-Oct-2014 by the R Development Core Team. It is widely used by researchers from diverse disciplines to estimate and display results and by teachers of statistics and research methods. Today, millions of analysts, researchers, and brands such as Facebook, Google, Bing, Accenture, and Wipro are using R to solve complex issues. The applications of R are not limited to just one sector, we can see the use of R in banking, e-commerce, finance, and many more sectors

It is freely available on www.r-project.org or can also download from CRAN (Comprehensive R Archive Network) website http://CRAN.R-project.org.

R Command Prompt:

We will be using RStudio. Once we have R environment setup, then it's easy to start R command prompt by just typing the following command at command prompt −
$ R

This will launch R interpreter and you will get a prompt > where you can start typing your program

> "Hello, World!"

[1] "Hello, world!"

Usually, we will write our code inside scripts which are called RScripts in R.

Write the below given code in a file

1 print ("Hello, World!")

and save it as myfirstprogrm.R and then run it in console by writing:

Rscript myfirstprogram.R

It will produce following output

[1] "Hello, World!"

3.2 BASIC SYNTAX

Any program in R is made up of three things: Variables, Comments, and Keywords. Variables are used to store the data.

Comments are used to improve code readability. They are like helping text in your R program. Single comment is written using # in the beginning of the statement.

Eg. # This is my first R program

Keywords are reserved words that hold a specific meaning to the compiler. Keyword cannot be used as a variable name, function name.

Following are the Reserved words in R: if, else, while, repeat, for, function, in, next, break, TRUE, FALSE, NULL, Inf, NaN, NA, NA_integer_, NA_real_, NA_complex_, NA_character etc

We can view these keywords by using either help(reserved) or ?reserved
R is case sensitive language.

3.3 DATA TYPES

Variables can store data of different types and different types can do different things. Variables are the reserved memory location to store values. As we create a variable in our program some space is reserved in memory.

Following are the data types used in R programming.

Data type	Example	Description
Numeric	50, 25.65, 999	Decimal values
Logical	True, False	Data with only two possible values which can be constructed as true/false
Character	'A', "Excellent", '50.50'	A character is used represent string values.
Integer	5L, 70L, 9876L	L tells R to store the value as an integer.
Complex	X= 5+4i	A complex value in R defined as the pure imaginary value i.
Raw		A raw data type is used to holds raw bytes.

We can use the class ( ) function to check the data type of a variable.

```
# numeric
a < - 25.5
class (a)
# complex
a < - 10+5i
class (a)
# integer
a < - 100L
class (a)
# logical/boolean
a < - TRUE
class (a)
# character/string
a < - "I am doing R programming"
class (a)
```

output:

[1] "numeric"

[1] “complex”

[1] “integer”

[1] “logical”

[1] “character”

3.4 VARIABLES

Variables are used to store the information to be manipulated in the R program. A variable in R can store an atomic vector, group of atomic vectors or a combination of many R-objects. A valid variable name consists of letters, numbers and the dot or underline characters. The variable name must start with a letter or the dot not followed by a number.
Ex – valid - a , a_b, a.b , a1 , a1. , a.c
Invalid - 2a, _a

R does not have a command for declaring a variable. A variable is created the moment you first assign a value to it.

In R, the assignment can be denoted in three ways:

1. = (Simple Assignment)
2. <- (Leftward Assignment)
3. -> (Rightward Assignment)

name = “Ajay”

gender < - “Male”

age < - 25

Here, name, gender and age are variables and “Ajay”, “Male”, 25 are values.

To print/output variable, you do not need any function. You can just type the name of the variable.
name = “Ajay”
O/P:
[1] “Ajay” #auto print the value of name variable

However, R have a print() and cat() functions which are used to print the value of the variable. The cat() function combines multiple values into a continuous print output.

Cat (“My name is” , name, “\n”)

Cat (“my age is”, age, “\n”)

O/P: My name is Ajay

My age is 25

ls() function: To know all the variables currently available in the workspace, use the ls() function.

```
# using equal to operator
a = “Good morning”
# using leftward operator
b < - “Good morning”
# using leftward operator
“Good morning - > c
 print(ls())
```

O/P: “a” “b” “c”

```
# List the variables starting with the pattern "var".
>  print(ls(pattern="var"))
```

The variables starting with dot (.) are hidden, they can be listed using "all.names=TRUE" argument to ls() function.

```
>  print(ls(all.name=TRUE))
```

rm() function: This is a built in function used to delete an wanted variables.

```
> rm(variable)
# using equal to operator
a = “Good morning”
# using leftward operator
b < - “Good morning”
# using leftward operator
“Good morning - > c
# Removing variable
rm(a)
print(a)
```

O/P: Error in print(a) : object ‘a’ not found

All the variables can be deleted by using the **rm()** and **ls()** function together.

```
> rm(list=ls())
> print(ls())
```

3.5 OPERATORS

Operators are the symbols directing the compiler to perform various kinds of operations between the operands. There are different types of

operator, and each operator performs a different task. Operators simulate the various mathematical, logical, and decision operations performed on a set of Complex Numbers, Integers, and Numericals as input operands.

Types of Operators used in R programming:

- Arithmetic Operators
- Relational Operators
- Logical Operators
- Assignment Operators
- Miscellaneous Operators

Arithmetic Operators:

Arithmetic operators are used with numeric values to perform common mathematical operations

< - , =	Assignment	A < - 5 ; b=10
+	Addition	x <- c(2,5.5,6) ; y <- c(8, 3, 4); print(x+ y) # O/P [1] 10.0 8.5 10.0
-	Subtraction	x <- c(2, 5.5,6); y <- c(8, 3, 4); print(x - y)
*	Multiplication	x<- c(2,5.5,6); y <- c(8, 3, 4); print(x*y)
/	division	x <- c(2,5.5,6); y <- c(8, 3, 4); print(x/y)
%%	remainder	x<- c(2,5.5,6); y <- c(8, 3, 4); print(x%%y) #O/P [1] 2.0 2.5 2.0
%/%	gives quotient	x <- c(2,5.5,6); y <- c(8, 3, 4) ;print(x%/%y) # O/P 0 1 1
^ **	exponent	x <- c(2,5.5,6) ; y <- c(8, 3, 4); print(x^y) #O/P 256.000 166.375 1296.000

Relational Operators: Relational/Comparison operators are used to compare two values

>	Greater than	x <- c(2,5.5,6,9) ; y <- c(8,2.5,14,9); print(x>y) # O/P [1] FALSE TRUE FALSE FALSE
<	Less than	x <- c(2,5.5,6,9) ; y <- c(8,2.5,14,9) ; print(x < y) ; #O/P [1] TRUE FALSE TRUE FALSE

<=	Less than equal to	x <- c(2,5.5,6,9) ;y<- c(8,2.5,14,9) print(x<=y) #O/P [1] TRUE FALSE TRUE TRUE
>=	Greater than equal to	x <- c(2,5.5,6,9) ;y <- c(8,2.5,14,9) print(x>=y) #O/P [1] FALSE TRUE FALSE TRUE
==	Equal	x < - c(2,5.5,6,9); y < - c(8,2.5,14,9) print(x==y) #O/P [1] FALSE FALSE FALSE TRUE
!=	Not equal	x <- c(2,5.5,6,9) ; y <- c(8,2.5,14,9) print(x!=y) # O/P [1] TRUE TRUE TRUE FALSE

Logical Operators: Logical operators are used to combine conditional statements.

&	Element wise Logical AND	x <- c(3, 1, TRUE, 2+3i); y <- c(4, 1, FALSE, 2+3i) print(x&y); # O/P [1] TRUE TRUE FALSE TRUE
|	Element wise Logical OR	x <- c(3, 0, TRUE, 2+2i); y <- c(4, 5, FALSE, 2+3i) print(x|y) # O/P [1] TRUE TRUE TRUE TRUE
!	Element wise Logical NOT	x<- c(8, 0, FALSE, 4+4i); print(!x) # O/P [1] FALSE TRUE TRUE FALSE
&&	Takes first element of both the vectors and gives the TRUE only if both are TRUE.	x <- c(3,0,TRUE, 8+9i); y<- c(1,3,TRUE, 3+4i) print(x&&y) # O/P [1] TRUE
||	Logical OR operator. It returns TRUE if one of the statement is TRUE.	x <- c(4, 0,TRUE, 8+9i); y<- c(3, 5, TRUE, 2+3i) print(x||y) # O/P [1] TRUE

Miscellaneous Operators: Miscellaneous operators are used to anipulate data:

:	Create a series of numbers in sequence	x <- 2:9 print(x) # [1] 2 3 4 5 6 7 8 9
%in%	Find out if an element belongs to a vector	x <- 8 ; y <- 12 ; z <- 1:10 print(x %in% z) ; print(y %in% z) # O/P [1] TRUE [1] FALSE

%*%	It is used to multiply a matrix with its transpose	

3.6 CONTROL STATEMENTS

Control statements are expressions used to control the execution and flow of the program based on the conditions provided in the statements.

if condition: if statement checks the expression provided in the parenthesis is true or not true. The block of code inside if statement will be executed only when the expression evaluates to be true.

Syntax:

```
if (expression ) {
// statements will execute if expression is true.
}
```

```
a < - 500
if (a > 100) {
print((x, “is greater than 100”))
}
```

O/P: [1] “500 is greater than 100”

If ….. else condition: If expression evaluates to be true, then the if block of code will be executed, otherwise else block of code will be executed.

Syntax:

```
if(expression){
   // statements will execute if expression is true.
}
else{
   // statements will execute if expression is false.
}
```

```
a < - 500
if (a > 100) {
print(a, “is greater than 100”)
} else {
print(a, “is smaller than 100”)
}
```

O/P: [1] “500 is greater than 100”

Repeat loop: Repeat loop executes the same code again and again until stop condition met.

Syntax:

```
repeat { commands
if  (condition){ break
}
}
```

```
a <- 1
repeat {
        print(a)
        a =a+1
        if (a>5){ break
                }
         }
```

O/P:

```
[1] 1
[1] 2
[1] 3
[1] 4
[1] 5
```

return statement: return statement is used to return the result of an executed function and returns control to the calling function.

Syntax:

```
return(expression)
```

Example:

```
func < - function(a) {
if(a > 0){
                return (“POSITIVE”)
}else if (a < 0){
                        return(“NEGATIVE”)
}else{
return( “ZERO”)
}
}
fun(1)
fun(0)
```

fun(-1)

O/P :

“POSITIVE”

“NEGATIVE”

“ZERO”

next statement: next statement is useful when we want to skip the current iteration of a loop without terminating it.

Syntax:

next

Example:

```
a < - 1:8
#Print even numbers
for( i in a){
        if(i%%2 !=0){
                next
        }
        print(i)
}
```

O/P:

[1] 2

[1] 4

[1] 6

[1] 8

break statement: The break keyword is a jump statement that is used to terminate the loop at a particular iteration

Syntax:

```
if (test_ expression) {
break
}
```

switch Statement: A switch statement is a selection control mechanism. Switch case is a multiway branch statement. It allows a variable to be tested for equality against a list of values. If there is more than one match for a specific value, then the switch statement will return the first match found of the value matched with the expression.

Syntax:

switch(expression, case1, case2, case3, ……)

Example:

```
a < - switch( 2, “Nagpur”, “Mumbai”, “Delhi”,“Raipur”)
print(a)
```

O/P:

[1] “Mumbai”

while loop: The while loop executes the same code again and again until stop condition is met.

Syntax:

```
While (test_expression) {
Statement
}
```

Example:

```
a < - c(“Hello”, “World”)
count < - 1
while (count < 5) {
print(a)
count = count + 1
}
```

O/P

[1] “Hello” “World”

[1] “Hello” “World”

[1] “Hello” “World”

[1] “Hello” “World”

for loop: The for loop can be used to execute a group of statements repeatedly depending upon the number of elements in the object. It is an entry controlled loop, in this loop the test condition is tested first, then the body of the loop executed, the loop body would not be executed if the test condition is false.

Syntax:

```
for (value in vector) {
statements
}
```

Example:

```
v < - LETTERS[1:5]
for ( x in v
}) {
```

```
print(x)
```

O/P:

[1] “A”

[1] “B”

[1] “C”

[1] “D”

[1] “E”

Example:

```
for (x in c(-5, 8, 9, 11))
{ print(x)
}
```

O/P:

[1] -5

[1] 8

[1] 9

[1] 11

Nested for-loop: Nested loops are similar to simple loops. Nested means loops inside loop. R programming allows using one loop inside another loop. In loop nesting, we can put any type of loop inside of any other type of loop. For example, a if loop can be inside a for loop or vice versa. Moreover, nested loops are used to manipulate the matrix.

```
for ( i in 1:3)
{
        for ( j in 1:i)
        {
                print( i * j)
        }
}
```

O/P:

[1] 1

[1] 2

[1] 4

[1] 3

[1] 6

[1] 9

3.7 R-FUNCTIONS

A set of statements which are organized together to perform a specific task is known as a function. A function is a set of statements organized together to perform a specific task. R has a large number of in-built functions and the user can create their own functions. An R function is created by using the keyword **function**.

Syntax :

```
Function_name < - function(arg1, arg2, …..)
        { function body
                                    }
```

The different components of function are -

- Function name is the actual name of the function.
- An argument is placeholder. In function, argument are optional means a function may or may not contain arguments, and these arguments have default values also.
- The function body contains a set of statements which defines what the function does.
- Return value is the last expression in the function body which is to be evaluated.
- R also has two types of function, i.e. Built in function and user defined function.

Built-in function: The functions which are already defined in the programming framework are known as built in functions. Simple examples of built-in functions are seq(), mean(), amx(), sum(), paste(…) etc. They are directly called by user written programs.

```
print(seq(50, 60))
O/P: [1] 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60
print(mean(30, 40))
O/P: [1] 35
```

User defined Function: R allows us to create our own function in our program. They are specific to what a user wants and once created they can be used like built-in functions.

Example:

```
areaofCircle < - function(radius){
area = pi*radius^2
return(area)
}
print(areaofCircle(2))
O/P: [1] 12.56637
```

Example:

```
# create a function without an argument.
a.function < - function (){
for(i in 1:5) {

        b < - i^2
print(b)
}
}
# call the function a.function without supplying an argument
a.function()
O/P
[1] 1
[1] 4
[1] 9
[1] 16
[1] 25
# create a function with an argument.
a.function < - function (a){
for(i in 1:a) {

        b < - i^2
print(b)
}
}
# call the function a.function without supplying 5 as an argument
a.function(5)
O/P
[1] 1
[1] 4
[1] 9
[1] 16
[1] 25
```

Calling a function with argument values:

```
#Create a function with argument
a.function < - function(x,y,z) {
esult < - x * y + z
print( result)
}
# call the function by position of arguments
a.function(4, 2, 10)
# call the function by names of the arguments
a.function(x=10, y=4, z=2)
O/P:
[1] 18
```

[1] 42

Calling a function with default argument:

```
#Create a function with argument
a.function < - function(x = 5, y= 7) {
result < - x * y
print( result)
}
# call the function without giving any arguments
a.function()
# call the function with giving new values of the argument.
a.function(10, 6)
```

O/P:

[1] 35

[1] 60

3.8 R –VECTORS

A vector is a basic data structure. In R, a sequence of elements which share the same data type is known as vector. A vector supports logical, integer, double, character, complex, or raw data type. A vector length is basically the number of elements in the vector, and it is calculated with the help of the length() function.

Vector is classified into two parts, i.e., Atomic vectors and Lists. There is only one difference between atomic vectors and lists. In an atomic vector, all the elements are of the same type, but in the list, the elements are of different data types. The elements which are contained in vector known as components of the vector. We can check the type of vector with the help of the typeof() function.

Creation of atomic vector

Single Element Vector: when you write just one value, it becomes a vector of length 1.

```
print("xyz")
print(25.5)
print(TRUE)
```

O/P

[1] "xyz"

[1] 25.5

[1] TRUE

Multiple Elements vector:

1. Using the colon(:) operator:

```
# Creating a sequence from 1 to 8
v< - 1:8
```

```
print(v)
# Creating a sequence from 1.5 to 8.5
v< - 1.5:8.5
print(v)
O/P:
[1] 1    2       3       4       5       6       7       8
[1] 1.5 2.5      3.5     4.5     5.5     6.5     7.5     8.5
```

2. **Using sequence** (seq) operator:

```
# Create a vector from 1 to 5 incrementing by 0.6
print(seq(1, 5, by = 0.6))
O/P [1] 1.0     1.6     2.2     2.8     3.4     4.0     4.6
```

3. **Using the c() function**: The non character values are converted to character type if one of the elements is a character.

```
x < - c('mango', 'yellow', 10, TRUE)
print(x)
O/P
[1] "mango"   "yellow",     "10",          "TRUE"
```

Accessing Vector Elements: Elements of a Vector are accessed using indexing. The [ ] brackets are used for indexing. Indexing starts with position 1. Giving a negative value in the index drops that element from result. TRUE, FALSE or 0 and 1 can also be used for indexing.

```
# Accessing vector elements using position
x < - c("Jan", "Feb", "Mar", "April", "May", "Jun", "July", "Aug", "Sept", "Oct", "Nov", "Dec")
a < - x[c(2,4,8)]
print(a)
# Accessing vector elements using logical indexing.
b < - x[c(TRUE, TRUE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, TRUE, TRUE, FALSE, FALSE)]
print (b)
 # Accessing vector elements using negative indexing.
c < - x[c(-1, -3, -4, -8, -9, -10, -11)]
print(c )
# Accessing vector elements using 0/1indexing.
d < - x[c(0,1,0,0,1,0,0,0,0,0,0,1)]
print(d )
O/P
[1] "Feb"       "April"         "Aug"
[1] "Jan"       "Feb"  "May" "Sept" "Oct"
[1] "Feb"       "May" "Jun"  "July"   "Dec"
[1] "Feb"       "May" "Dec"
```

Vector Manipulation:

Vector arithmetic: Two vectors of same length can be added, subtracted, multiplied or divided giving the result as a vector output.

```
#create two vectors.
x < - c(2,7,3,4,0,10)
y < - c(3,10,0,7,1,1)
add.result < - x + y
print(add.result)
multi.result < - x * y
print(multi.result)
```

O/P:

```
[1] 5   17    3    11    1    11
[1] 6   70    0    28    0    10
```

Vector Element Recycling: If we apply arithmetic operations to two vectors of unequal length, then the elements of the shorter vector are recycled to complete the operations.

```
x < - c(2,7,3,4,0,10)
y < - c(3,10)
 # y becomes c(3,10,3,10,3,10)
add.result < - x + y
print(add.result)
```

O/P: [1] 5 17 6 14 3 20

Vector Element Sorting: Elements in a vector can be sorted using the sort() function.

```
x < - c(2,7,3,-11, 4,0,210)
sort.result < - sort(x)
print(sort.result)
resort.result < - sort(x, decreasing - TRUE)
print(resort.result)
```

O/P:

```
[1]   -11   0   2   3   4   7   210
[1]   210   7   4   3   2   0   -11
```

3.9 R – LISTS

Lists are heterogeneous data structures. Lists are the R objects which contain elements of different types. These are also one-dimensional data structures. A list can be a list of vectors, list of matrices, a list of characters and a list of functions and so on. List is created using list() function.

Creating a List:

```
#Create a list containing strings, numbers, vectors)
list_1 < - list(“Apple”, “Mango”, 25.25, 60.5, c(16,25,36))
print(list_1)
```

O/P:
[[1]]
[1] “Apple”
[[2]]
[1] “Mango”
[[3]]
[1] 25.25
[[4]]
[1] 60.5
[[5]]
[1] 16 25 36

Naming List Element: The list elements can be given and they can be accessed using these names.

```
list_1 < - list(c(“Mon”, “Tues”, “Wed”), matrix(c(1,2,3,4,5,6), nrow = 2)
#Give names to the elements in the list.
names(list_1) < - c(“Days of Week”, “Matrix”)
print(list_1)
```

O/P

```
$Days of  Week
 [1] “Mon”   “Tues”   “Wed”
$Matrix
          [,1]  [,2]  [,3]
[1,]       1     2     3
[2,]       4     5     6
```

Accessing List Elements:

Elements of the list can be accessed by the index of the element in the list. In case of named lists it can also be accessed using the names.

```
list_1 < - list(c(“Mon”, “Tues”, “Wed”), matrix(c(1,2,3,4,5,6), nrow = 2)
names(list_1) < - c(“Days of Week”, “Matrix”)
print(list_1[1])
        print(list_1$Matrix)
```

O/P:

```
$Days of  Week
        [1] “Mon”   “Tues”          “Wed”
              [,1]  [,2]  [,3]
        [1,]       1     2     3
        [2,]       4     5     6
```

Manipulating List Elements:

We can add, delete and update list elements as shown below. We can add and delete elements only at the end of a list. But we can update any element.

```
list_1 < - list(c("Mon", "Tues", "Wed"), matrix(c(1,2,3,4,5,6), nrow = 2)
names(list_1) < - c("Days of Week", "Matrix")
        #add element at the end of the list
        list_1[3] < - "Add Element"
        print(list_1[4])
        O/P
        [[1]]
        [1] "Add Element"
```

Merging Lists:

You can merge many lists into one list by placing all the lists inside one list() function.

```
        list_a < - list(1,2)
        list_b < - list("Ankit", "Pooja")
        #merge tow lists
        merged.list < - c(list_a, list_b)
        print(merged.list)
 [[1]]
 [1] 1
 [[2]]
 [1] 2
 [[3]]
 [1] Ankit
 [4]]
[1] Pooja
```

Converting List to vector:

A list can be converted to a vector so that the elements of the vector can be used for further manipulation. All the arithmetic operations on vectors can be applied after the list is converted into vectors.

```
        list_a < - list(10:13)
        print(list_a)
        list_b < - list(20:23)
        print(list_b)
        #Convert the lists to vectors
        x1 < - unlist(list_a)
        x2 < - unlist(list_b)
        print(x1)
        print(x2)
        add < -x1 + x2
```

```
print(add)
O/P
[[1]]
[1] 10  11      12      13
[[2]]
[1] 20  21      22      23
[1] 10  11      12      13
[1] 20  21      22      23
[1] 30  32      34      36
```

3.10 R ARRAYS

Arrays are the R data objects which can store data in more than two dimensions. In R, an array is created with the help of the array() function. This array() function takes a vector as an input and to create an array it uses vectors values in the **dim** parameter.

```
#create two vectors of different lengths
v1 < - c(1,2,3)
v2 < - c(4,5,6,7,8,9)
#Take these vectors as input to the array
array1 < - (c(v1,v2), dim = c(3,3,2))
print(array1)
O/P
, , 1
        [,1]    [,2]    [,3]
[1,]     1       4       7
 [2,]    2       5       8
[3,]     3       6       9
,  , 2
        [,1]    [,2]    [,3]
[1,]     1       4       7
[2,]     2       5       8
[3,]     3       6       9
```

Naming Columns and Rows: We can give names to the rows, columns and matrices in the array by using the dimnames parameter.

```
#create two vectors of different lengths
v1 < - c(1,2,3)
v2 < - c(4,5,6,7,8,9)
column.names <- c("Col1", "Col2", "Col3")
row.names <- c("Row1", "Row2", "Row3")
matrix.names <- c("Matrix1", "Matrix2")
array1 < - array(c(v1,v2),dim = c(3,3,2), dimnames = list(row.names, column.names, matrix.names))
print(array1)
O/P:
```

, , Matrx1

	Col1	Col2	Col3
Row1	1	4	7
Row2	2	5	8
Row3	3	6	9

, , Matrix2

	Col1	Col2	Col3
Row1	1	4	7
Row2	2	5	8
Row3	3	6	9

Accessing Array Elements:

```
#create two vectors of different lengths
v1 < - c(1,2,3)
v2 < - c(4,5,6,7,8,9)
column.names <- c("Col1", "Col2", "Col3")
row.names <- c("Row1", "Row2", "Row3")
matrix.names <- c("Matrix1", "Matrix2")
array1 < - array(c(v1,v2),dim = c(3,3,2), dimnames = list(row.names, column.names, matrix.names))
#Print the second row of the second matrix of the array.
print(array1[2,2])
#Print the element in the first row and 3rd column of the first matrix.
print(array1[1,3,1])
#Print the first Matrix
print(array1[, ,1])
```

O/P

Col1	Col2	Col3
2	5	8

[1] 7

	Col1	Col2	Col3
Row1	1	4	7
Row2	2	5	8
Row3	3	6	9

Manipulating Array Element: As array is made up matrices in multiple dimensions, the operations on elements of array are carried out by accessing elements of the matrices.

```
#create two vectors of different lengths
v1 < - c(5,9,3)
v2 < - c(10,11,12,13,14,15)
#Take these vectors as input to the array
array1 < - array(c(v1,v2),dim = c(3,3,2))
#create two vectors of different lengths
V3 < - c(9,1,0)
```

```
V4< - c(6,0,11,3,14,1,2,6,9)
array2 < - array(c(v3,v4),dim = c(3,3,2))
#create matrices from these arrays
matrix1 < - array1[, , 2]
matrix2 < - array2[, , 2]
#add the matrices
add1 < - matrix1+matrix2
print(add1)
```

O/P:

	[,1]	[,2]	[,3]
[1,]	10	20	26
[2,]	18	22	28
[3,]	6	24	30

3.11 SUMMARY

R is world's most widely used statistics programming language. It is the 1 choice of data scientists R is taught to solve critical business applications. In addition, R is a full-fledged programming language, with a rich complement of mathematical functions, matrix operations and control structures. It is very easy to write your own functions. In this chapter we covered basic programming to different types of data objects of R with suitable examples in simple and easy steps.

3.12 EXERCISE

1. Find the output of following code.

1)
```
b= "15"
a = switch( b,
"5"="Hello A",
            "10"="Hello B",
            "15"="Hello C",
        "20"="Hello D"  )
             print (a)
```

2)
```
a= 1
 b = 2
y = switch (a+b, "Hello, A",  "Hello B", "Hello C", "Hello D" )
print (y)
```

3)
```
# Create vegetable vector
vegetable <- c('Potato', 'Onion','Brinjal', 'Pumpkin')
for ( x in vegetable) {
print(x)
}
```

```
4) for ( i in c (5, 10, 15, 20, 0, 25)
{
        if (i == 0)
        {
                break
        }
print (i)
}
print(“outside loop”)

5) for ( i in c (5, 10, 15, 20, 0, 25)
{
        if (i == 0)
        {
                next
        }
print (i)
}
print(“outside loop”)

6) a < - 10
b<- 14
count=0
if(a<b)
{
   cat(a,"is a smaller number\n")
   count=1
}
if(count==1){
   cat("Block is successfully execute")
}

7) a <-1
b<-24
count=0
while(a<b){
                cat(a,"is a smaller number\n")
                 a=a+2
                  if(x==15)
          break
}
8) a < -24
if(a%%2==0){
                cat(a," is an even number")
}
if(a%%2!=0){
                        cat(a," is an odd number")
}
9) x <- c("Hardwork","is","the","key","of","success")
```

```
if("key" %in% x) {
                print("key is found")
} else {
                        print("key is not found")
}
```

```
10) Rectangle = function(l=6, w=5){
  area = l * w
  return(area)
}
print(Rectangle(3, 4))
 print(Rectangle(w = 9, l = 3))
 print(Rectangle())
```

3.13 REFERENCES

- The Art of R Programming: A Tour of Statistical Software Design by Norman Matloff
- Beginning R – The Statistical Programming Language by Mark Gardener
- https://www.javatpoint.com/
- https://www.ict.gnu.ac.in/content/r-programming
- https://www.geeksforgeeks.org/

*****

UNIT II

4

MOMENTS, SKEWNESS, AND KURTOSIS

Unit Structure

4.1 OBJECTIVE

After going through this unit, you will able to :

- Define Moments and calculate for ungroup and group data.
- Explain types of moments.
- Find relation between raw and central moment.
- Used Charlier's check method in computing moments by coding method.
- Define Sheppard's correction for moments.
- Define moments in dimensional form.
- Define Skewness and Kurtosis.
- Calculate moments, Skewness and Kurtosis using software.

4.2 INTRODUCTION

The measure of central tendency (location) and measure of dispersion (variation) both are useful to describe a data set but both of them fail to tell anything about the shape of the distribution. We need some other certain measure called the moments to identify the shape of the distribution

known as Skewness and Kurtosis.Moments are statistical measures that give certain characteristics of the distribution. Moments provide sufficient information to reconstruct a frequency distribution function. Moments are a set of statistical parameters to measure a distribution. Four moments are commonly used:1st moments for Average, 2nd for Variance, 3rd for Skewness and 4th moment for Kurtosis.

4.3 MOMENTS

The arithmetic mean of the rth power of deviations taken either from mean, zero or from any arbitrary origin is called moments. Assume there is sequence of random variables$x_1, x_2, x_3, \ldots\ldots\ldots x_n$. The first sample moment, usually called the average is defined by first moments. Three types of moments are defined as follow:

When the deviations are computed from the arithmetic mean, then such moments are called moments about mean (mean moments) or sometimes called central moments, denoted by μ_r and given as follows:Hence for ungroup data,

i) The first moment about A, as $\mu_1 = \frac{\sum(x-\bar{x})}{n}$.

ii) The second moment about A, as$\mu_2 = \frac{\sum(x-\bar{x})^2}{n}$.

iii) The third moment about A, as$\mu_3 = \frac{\sum(x-\bar{x})^3}{n}$.

iv) The fourth moments about A, as $\mu_4 = \frac{\sum(x-\bar{x})^4}{n}$

When the deviations of the values are computed from any arbitrary value (provisional mean) say *A*, then such moments are called moments about arbitrary or provisional mean denoted by $\mu_r(a)$.Hence for ungroup data,

i) The first moment about A, as $\mu_1(a) = \frac{\sum(x-A)}{n}$.

ii) The second moment about A, as$\mu_2(a) = \frac{\sum(x-A)^2}{n}$.

iii) The third moment about A, as$\mu_3(a) = \frac{\sum(x-A)^3}{n}$.

iv) The fourth moments about A, as $\mu_4(a) = \frac{\sum(x-A)^4}{n}$

When the deviations of the values are computed from the origin or zero, then such moments are called moments about origin or raw moments denoted by $\mu_r(a)$

i) The first moment about origin, as $\mu'_1 = \frac{\sum(x)}{n}$.

ii) The second moment about origin, as$\mu'_2 = \frac{\sum(x)^2}{n}$.

iii) The third moment about origin, as$\mu'_3 = \frac{\sum(x)^3}{n}$.

iv) The fourth moments about origin, as $\mu'_4 = \frac{\sum(x)^4}{n}$

Example 1: Find raw moments for the following data: 5, 8, 12, 4, 6.

Solution:

x	x^2	x^3	x^4
5	25	125	625
8	64	512	4096
2	4	8	16
4	16	64	256
6	36	216	1296
$\sum x = 25$	$\sum x^2 = 145$	$\sum x^3 = 925$	$\sum x^4$ =6289

i) The first moment about origin, as $\mu'_1 = \frac{\sum(x)}{n} = \frac{25}{5} = 5$.

ii) The second moment about origin, as$\mu'_2 = \frac{\sum(x)^2}{n} = \frac{145}{5} = 29.$

iii) The third moment about origin, as$\mu'_3 = \frac{\sum(x)^3}{n} = \frac{925}{5} = 185$.

iv) The fourth moments about origin, as $\mu'_4 = \frac{\sum(x)^4}{n} = \frac{6289}{5} = 1257.8$

4.3.1 Moments for Grouped Data:

1. Moments about arbitrary point: Let x represents a variable occurring with frequency f, in a given distribution, then the i^{yh} moment $\mu_i(a)$ about A is defined as

$\mu_i(a) = \frac{\sum f(x-A)^i}{N}$, where$N = \sum f$.

We generally find moments upto $i = 4$.

∴we can write :

i) The first moment about A, as $\mu_1(a) = \frac{\sum f(x-A)}{N}$.

ii) The second moment about A, as $\mu_2(a) = \frac{\sum f(x-A)^2}{N}$.

iii) The third moment about A, as $\mu_3(a) = \frac{\sum f(x-A)^3}{N}$.

iv) The fourth moments about A, as $\mu_4(a) = \frac{\sum f(x-A)^4}{N}$.

Example 2: For the following distribution find all four moments about 5.

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

Solution: let prepared table first,

x	f	$(x-5)$	$f(x-5)$	$f(x-5)^2$	$f(x-5)^3$	$f(x-5)^5$
2	4	-3	-12	36	-108	324
4	6	-1	-6	6	-6	6
6	12	1	12	12	12	12
8	5	3	15	45	135	405
10	3	5	15	75	375	1875
Total	30		24	174	408	2622

Moments about arbitrary A = 5 is given by

The first moment about A, as $\mu_1(a) = \frac{\sum f(x-A)}{N} = \frac{24}{30} = 0.8$.

The second moment about A, as $\mu_2(a) = \frac{\sum f(x-A)^2}{N} = \frac{174}{30} = 5.8$.

The third moment about A, as $\mu_3(a) = \frac{\sum f(x-A)^3}{N} = \frac{408}{30} = 13.6$.

The fourth moments about A, as $\mu_4(a) = \frac{\sum f(x-A)^4}{N} = \frac{2622}{30} = 87.4.$

2. Moments about mean (Central moments):

These are moments about the Arithmetic Mean $\bar{x}$. Hence when A is taken as $\bar{x}$, we obtain these moments. Thus it is given by

i) The first moment about $\bar{x}$, as

$\mu_1 = \frac{\sum f(x-\bar{x})}{N}$.

ii) The second moment about $\bar{x}$, as

$\mu_2 = \frac{\sum f(x-\bar{x})^2}{N}$.

iii) The third moment about $\bar{x}$, as

$\mu_3 = \frac{\sum f(x-\bar{x})^3}{N}$.

iv) The fourth moments about $\bar{x}$, as

$\mu_4 = \frac{\sum f(x-\bar{x})^4}{N}$.

From the definition of the mean $\bar{x}$ and the standard deviation σ, it immediately follows that $\mu_1 = 0$, $\mu_2 = \sigma^2$ and μ_3 measure the asymmetry of the curve. These moments are important study the nature of the distribution.

Example 3: Find the central moments for the following distribution:

X	1	2	3	4	5
F	2	5	6	5	2

Solution:

x	f	fx	$(x-\bar{x})$	$f(x-\bar{x})$	$f(x--x^2)$	$f(x-\bar{x})^3$	$f(x-\bar{x})^4$
1	2	2	-2	-4	8	-16	32
2	5	10	-1	-5	5	-5	5
3	6	18	0	0	0	0	0
4	5	20	1	5	5	5	5
5	2	10	2	4	8	16	32
Total	20	60		0	26	0	74

Here, $\bar{x} = \frac{\sum fx}{N} = \frac{60}{20} = 3$.

Therefore, the central moments are given by

i) The first moment about $\bar{x}$, as

$\mu_1 = \frac{\sum f(x-\bar{x})}{N} = \frac{0}{20} = 0$.

ii) The second moment about $\bar{x}$, as

$\mu_2 = \frac{\sum f(x-\bar{x})^2}{N} = \frac{26}{20} = 1.3$.

iii) The third moment about $\bar{x}$, as

$\mu_3 = \frac{\sum f(x-\bar{x})^3}{N} = \frac{0}{20} = 0$.

iv) The fourth moments about $\bar{x}$, as

$\mu_4 = \frac{\sum f(x-\bar{x})^4}{N} = \frac{74}{20} = 3.7$.

3. Moments about origin(Raw moments):

As the name suggests, taking A as the origin (A = 0), we get these moments. Thus it is given by

i) The first moment about Origin, as

$\mu_1' = \frac{\sum fx}{N}$.

ii) The second moment about Origin, as

$\mu_2' = \frac{\sum fx^2}{N}$.

iii) The third moment about Origin, as

$\mu_3' = \frac{\sum fx^3}{N}$.

iv) The fourth moments about Origin, as

$\mu_4' = \frac{\sum fx^4}{N}$.

Note that for first moment about origin is mean of the data.

Example 4: Find the raw moments for the following data:

X	-1	0	1	2	3	4
F	2	4	3	7	3	1

Solution : lets prepared table

x	f	fx	fx^2	fx^3	fx^4
-1	2	-2	2	-2	2
0	4	0	0	0	0
1	3	3	3	3	3
2	7	14	28	56	112
3	3	9	27	81	243
4	1	4	16	64	256
Total	**20**	**28**	**76**	**202**	**616**

Therefore, the raw moments are given by

i) The first moment about Origin, as

$\mu_1' = \frac{\sum fx}{N} = \frac{28}{20} = 1.4$.

ii) The second moment about Origin, as

$\mu_2' = \frac{\sum fx^2}{N} = \frac{76}{20} = 3.8$.

iii) The third moment about Origin, as

$\mu_3' = \frac{\sum fx^3}{N} = \frac{202}{20} = 10.1$.

iv) The fourth moments about Origin, as

$\mu_4' = \frac{\sum fx^4}{N} = \frac{616}{20} = 30.8.$

4.3.2 Relations between Moments:

We studied three different types of moments. Now it is very useful to simplifying relation between them. We will now give inter-relation between various moments and solve example using these relations.

Relation between moments about arbitrary point and the central moment:

i) $\mu_1 = \mu_1(a) - \mu_1(a) = 0$

ii) $\mu_2 = \mu_2(a) - \mu_1(a)^2$

iii) $\mu_3 = \mu_3(a) - 3\mu_1(a)\mu_2(a) + 2\mu_1(a)^3$

iv) $\mu_4 = \mu_4(a) - 4\mu_1(a)\mu_3(a) + 6\mu_1(a)^2\mu_2(a) - 3\mu_1(a)^4$

Conversely the moments $\mu_i(a)'s$ about A in term of μ_i'sare given as follows:

i) $\mu_1(a) = \bar{x} - A$

ii) $\mu_2(a) = \mu_2 + \mu_1(a)^2$

iii) $\mu_3(a) = \mu_3 + 3\mu_2\mu_1(a) + \mu_1(a)^3$

iv) $\mu_4(a) = \mu_4 + 4\mu_3\mu_1(a) + 6\mu_2\mu_1(a)^2 + \mu_1(a)^4$

Relation between Raw moments and central moments:

Recall that, the raw moments μ'_i are obtained from the general moments $\mu_i(a)$ when A is taken as '0'.

Hence taking A as '0'and replacing $\mu_i(a)$ by corresponding μ'_i in the formula, we get

i) $\mu_1 = \mu'_1 - \mu'_1 = 0$

ii) $\mu_2 = \mu_2' - \mu_1'^2$

iii) $\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3$

iv) $\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6\mu_1'^2\mu_2' - 3\mu_1'^4$

Conversely the moments μ'_i in term of μ_i are given as follows:

i) $\mu_1' = \bar{x}$

ii) $\mu_2' = \mu_2 + \mu_1'^2$

iii) $\mu_3' = \mu_3 + 3\mu_2\mu_1' + \mu_1'^3$

iv) $\mu_4' = \mu_4 + 4\mu_3\mu_1' + 6\mu_2\mu_1'^2 + \mu_1'^4$

Example 5: The first four central moments of a distribution are 0, 3, 5, 10. If the mean of the distribution is 2, find the moments about 3.

Solution: We have $A = 3, \bar{x} = 2, \mu_1 = 0, \mu_2 = 3, \mu_3 = 5, \mu_4 = 10.$
Using the relation between central moments and arbitrary moments,

i) $\mu_1(a) = \bar{x} - A = 2 - 3 = -1.$

ii) $\mu_2(a) = \mu_2 + \mu_1(a)^2 = 3 + (-1)^2 = 4$

iii) $\mu_3(a) = \mu_3 + 3\mu_2\mu_1(a) + \mu_1(a)^3 = 5 + 3(3)(-1) + (-1)^3 = -5.$

iv) $\mu_4(a) = \mu_4 + 4\mu_3\mu_1(a) + 6\mu_2\mu_1(a)^2 + \mu_1(a)^4$
$= 10 + 4(5) + 6(3)(-1)^2 + (-1)^4 = 49.$

Example 6: The first four raw moments about the origin are 2, 12, 74 and 384. Find the mean $\bar{x}$ and the first four central moments.

Solution: We already define the raw moments about the origin i.e. $\mu_i(a)'$s with $A = 0$. Given that $\mu_1(a) = 2, \mu_2(a) = 12, \mu_3(a) = 74, and\ \mu_4(a) = 384,\ with\ A = 0.$

Therefore, Mean= $\bar{x} = \mu_1(a) + A = 2 + 0 = 2.$

Using the relation between raw moments and central moments

i) $\mu_1 = \mu_1(a) - \mu_1(a) = 2 - 2 = 0$

ii) $\mu_2 = \mu_2(a) - \mu_1(a)^2 = 12 - 2^2 = 8$

iii) $\mu_3 = \mu_3(a) - 3\mu_1(a)\mu_2(a) + 2\mu_1(a)^3 = 74 - 3(2)(12) + 2(2^2) = 10$

iv) $\mu_4 = \mu_4(a) - 4\mu_1(a)\mu_3(a) + 6\mu_1(a)^2\mu_2(a) - 3\mu_1(a)^4$
$= 284 - 4(2)(74) + 6(2^2)(12) - 3(2^4)$
$= 384 - 592 + 288 - 48 = 128.$

Example 7: The first four central moments for a distribution are 0,3,0 and 7. If the mean $\bar{x}$ of the distribution is 4, find the first four raw moments.

Solution: The raw moments are the moments about origin. Given that $\mu_1 = 0, \mu_2 = 3, \mu_3 = 0\ and\ \mu_4 = 7\ with\ \bar{x} = 4.$
Using the relation between central moments and raw moments.

i) $\mu_1' = \bar{x} = 4$

ii) $\mu_2' = \mu_2 + \mu_1'^2 = 3 + 4^2 = 17.$

iii) $\mu_3' = \mu_3 + 3\mu_2\mu_1' + \mu_1'^3 = 0 + 3(3)(4) + 4^3 = 100.$

$\mu_4' = \mu_4 + 4\mu_3\mu_1' + 6\mu_2\mu_1'^2 + \mu_1'^4$

4.3.3 Computation of Moments for Grouped Data:

We have already found mean and standard deviation for the continuous data (grouped data). Now to calculate moments for the continuous data we used coding method (Short method).

When the values of x are not consecutive, but equally spaced at an interval of length$'c'$. We need to divide the expression by $'c'$. It is called change of scale by $'c'$.

Where we take $x = a + cu \; or \; u = \frac{x-a}{c}$.

We give below the effect of change of origin and scale on moments.

Let $x = a + cu$,

$\therefore \; \bar{x} = a + c\bar{u}$.

i)The moments of x about A are given by

$$\mu_i(a) = \frac{\sum f u^i}{N} \times c^i$$

ii) The central moments of x are given by

$$\mu_i = \frac{\sum f(u-\bar{u})^i}{N} \times c^i$$

Note: When A = 0, we get the raw moment.

Example 8: Find the central moments for the following data:

Class interval	0-20	20-40	40-60	60-80
Frequency	4	7	6	3

C.I	F	Class Marks (x)	$u = \frac{x-a}{c}$	Fu	$(u - 0.4)$	$f(u - 0.4)$	$f(u - 0.4)^2$	$f(u - 0.4)^3$	$f(u - 0.4)^4$
0-20	4	10	-1	-4	-1.4	-5.6	7.84	-10.976	15.3664
20-40	7	30	0	0	-0.4	-2.8	3.92	-0.448	0.1792
40-60	6	50	1	6	0.6	3.6	2.16	1.296	0.7776
60-80	3	70	2	6	1.6	4.8	7.68	12.288	19.6608
Total	20			8		0	21.6	2.16	35.984

Solution: first find mean by coding method, taking $a = 30$

Here, $\bar{u} = \frac{\sum fu}{N} = \frac{8}{20} = 0.4$.

The central moments of x are given by

$$\mu_i = \frac{\sum f(u-\bar{u})^i}{N} \times c^i$$

i) $\mu_1 = \frac{\sum f(u-\bar{u})^1}{N} \times c^1 = \frac{0}{20} \times 20 = 0.$

ii) $\mu_2 = \frac{\sum f(u-\bar{u})^2}{N} \times c^2 = \frac{21.6}{20} \times 20^2 = 432.$

iii) $\mu_3 = \frac{\sum f(u-\bar{u})^3}{N} \times c^3 = \frac{2.16}{20} \times 20^3 = 864.$

iv) $\mu_4 = \frac{\sum f(u-\bar{u})^4}{N} \times c^4 = \frac{35.984}{20} \times 20^4 = 2{,}87{,}872.$

4.4 CHARLIE'S CHECK AND SHEPPARD'S CORRECTIONS

A check which can be used to verify correct computations in a table of grouped classes. For example, consider the following table with specified class limits and frequencies f. The class marks x_i are then computed as well as the rescaled frequencies u_i, which are given by

$$\boldsymbol{u_i = \frac{f_i - x_0}{c}}$$

Where the class mark is taken as $x_0 = 44.5$ and the class interval is $c = 10$. The remaining quantities are then computed as follows.

Class interval	x_i	f_i	u_i	$f_i u_i$	$f_i {u_i}^2$	$f_i(u_i+1)^2$
0-9	4.5	2	-4	-8	32	18
10-19	14.5	3	-3	-9	27	12
20-29	24.5	11	-2	-22	44	11
30-39	34.5	20	-1	-20	20	0
40-49	44.5	32	0	0	0	32
50-59	54.5	25	1	25	25	100
60-69	64.5	7	2	14	28	63
Total		100		-20	176	236

In order to compute the variance, note that

$$V(u) = \frac{\sum \boldsymbol{f_i {u_i}^2}}{\sum \boldsymbol{f_i}} - \left(\frac{\sum \boldsymbol{f_i u_i}}{\sum \boldsymbol{f_i}}\right)^2$$

$$= \frac{176}{100} - \left(\frac{-20}{100}\right)^2 = 1.72$$

So the variance of the original data is

$$V(x) = c^2 V(u) = 100 \times 1.72 = 172.$$

Charlier's check makes use of the additional column $f_i(u_i + 1)^2$ added to the right side of the table. By noting that the identity

$$\sum \boldsymbol{f_i(u_i + 1)^2} = \sum \boldsymbol{f_i(u_i^2 + 2u_i + 1)}$$

$$= \sum \boldsymbol{f_i u_i^2 + 2\sum u_i + \sum f_i}$$

connects columns five through seven, it can be checked that the computations have been done correctly. In the example above,

236 = 176 +2 (- 20) +100 (8)

Hence, the computations pass Charlier's check.

Charlier's check in computing moments by the coding method uses the following identities:

$$\sum f(u + 1) = \sum fu + \sum f$$

$$\sum f(u + 1)^2 = \sum fu^2 + 2\sum fu + \sum f$$

$$\sum f(u + 1)^3 = \sum fu^3 + 3\sum fu^2 + 3\sum fu + \sum f$$

$$\sum f(u + 1)^4 = \sum fu^4 + 4\sum fu^3 + 6\sum fu^2 + 4\sum fu + \sum f$$

Sheppard's Corrections:

When the frequency distribution, consists of interval, we take x as the class mark of the interval and use this x in all the formulae.

While doing this, it is assumed that all the values in the interval concentrate at the class mark. But this assumption may not be always true and we are likely to get some errors in this calculation.

The well-known statistician Sheppard gave the corrected values of the moments as follows:

$$\mu_1(corrected) = \mu_1$$

$$\mu_2(corrected) = \mu_2 - \frac{c^2}{12}$$

$$\mu_3(corrected) = \mu_3$$

$$\mu_4 = \mu_4 - \frac{1}{2}c^2\mu_2 + \frac{7}{240}c^4$$

Where ‘c’ is the length of the class-interval, which is the same as the spacing between the mid values.

Note that even though this correction has great mathematical significance, we need not use these corrections in practice because the error is too small hence negligible and also in statistic, we look for estimates, which are approximate values.

Example 9: Apply Sheppard’s corrections to determine the moments about the mean for the data

Class Interval	0-10	10-20	20-30	30-40	40-50
Frequency	1	2	9	2	6

Solution: Lets prepared table, taking $A = 25$.

Class Interval	Class Mark (X)	f	$u = \frac{x-a}{c}$	fu	$(u - 0.5)$	$f(u - 0.5)$	$f(u - 0.5)^2$	$f(u - 0.5)^3$	$f(u - 0.5)^4$
0-10	5	1	-2	-2	-2.5	-2.5	12.5	-15.625	39.0625
10-20	15	2	-1	-2	-1.5	-3	4.5	-6.75	10.125
20-30	25	9	0	0	-0.5	-4.5	2.25	-1.125	0.5625
30-40	35	2	1	2	0.5	1	0.5	0.25	0.125
40-50	45	6	2	12	1.5	9	13.5	20.25	30.375
Total		20		10		0	33.25	-3	80.25

Here, $\bar{u} = \frac{\sum fu}{N} = \frac{10}{20} = 0.5$.

The central moments of x are given by

$$\mu_i = \frac{\sum f(u-\bar{u})^i}{N} \times c^i$$

i) $\mu_1 = \frac{\sum f(u-\bar{u})^1}{N} \times c^1 = \frac{0}{20} \times 20 = 0.$

ii) $\mu_2 = \frac{\sum f(u-\bar{u})^2}{N} \times c^2 = \frac{33.25}{20} \times 20^2 = 665.$

iii) $\mu_3 = \frac{\sum f(u-\bar{u})^3}{N} \times c^3 = \frac{-3}{20} \times 20^3 = -1200.$

iv) $\mu_4 = \frac{\sum f(u-\bar{u})^4}{N} \times c^4 = \frac{80.25}{20} \times 20^4 = 6{,}42{,}000.$

Sheppard gave the corrected values of the moments as follows:

$\mu_1(corrected) = \mu_1 = 0, \qquad \mu_2(corrected) = \mu_2 - \frac{c^2}{12} = 665 - \frac{20^2}{12} = 631.67$

$\mu_3(corrected) = \mu_3 = -1200$

$\mu_4 = \mu_4 - \frac{1}{2}c^2\mu_2 + \frac{7}{240}c^4 = 6{,}42{,}000 - \frac{400}{2} \times 665 + \frac{7}{240} \times 1{,}60{,}000 = 5{,}13{,}666.67.$

4.5 MOMENTS IN DIMENSIONLESS FORM

To avoid particular units, we can define the dimensionless central moments as

$$a_r = \frac{\mu_i}{\sigma^i}$$

Where σ is the standard deviation, so, as we have $\sigma = \sqrt{\mu_2}$,
We already know that for central moments, $\mu_1 = 0, \mu_2 = \sigma^2$.
So, we get $a_0 = 0 \; and \; a_1 = 1$.

4.6 SKEWNESS

Skewness is one more concept which deals with the symmetry or rather asymmetry of the values of distribution around its central value. When a frequency distribution is plotted on a chart, an ideal distribution by a nice, symmetric, bell-shaped curve around the central value. Such a distribution is called symmetric distribution or a normal distribution. However in practice every distribution that we across need not be normal. Their graph will be asymmetric or skew. Such distributions are called skewed distribution.

Definition: Skewness defined by famous statistician Garrett " A distribution is said to be skewed when the mean and median fall at different points of the distribution and balance is shifted to one side or the other to left or right."

Types of Skewness: In order to understand this concept we draw the following graphs, where$\bar{x}$ = Mean, M_e = Median and M_o = Mode of the distributions.

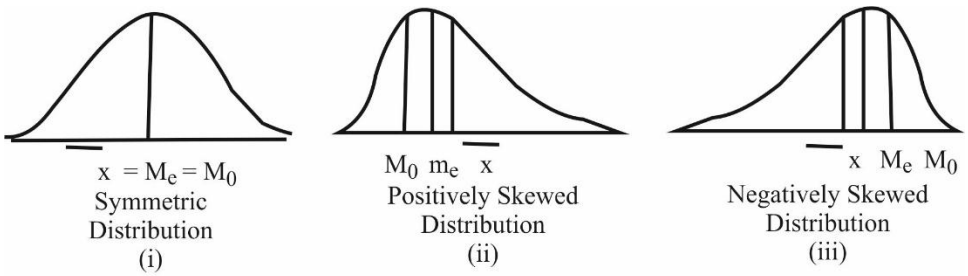


Figure 4.1

It is clear from the diagram that

i) Represents a symmetric distribution, for which Mean = Median = Mode.

ii) Represents a positive skewed distribution for which Mode < Median <Mode.

iii) Represents a negative skewed distribution for which Mean < Median <Mode.

Measure of Skewness:

Since mean, median and mode are different for a skewed distribution, the simplest measure would be the difference between two of these in pairs. Though such measures are simple to calculate, their main drawback is the following: these measures are expressed with respect to the corresponding units of the distribution. Therefore two distributions with different units cannot be compared. In order to overcome this difficulty, relative measures are defined. These are called Coefficients of Skewness.

Karl Pearson's Coefficient of Skewness: it is defined as

$$S_k = \frac{Mean - Mode}{standard\ deviation} = \frac{\bar{x} - M_o}{\sigma}$$

Using the relation between mean, median and mode:

Mean – Mode = 3 (Mean – Median), we can write

$$S_k = \frac{3(Mean - Median)}{Standard\ deviation} = \frac{3(\bar{x} - M_e)}{\sigma}$$

Interpretation on S_k

i) If S_k is positive then the distribution is positively skewed.

ii) If S_k is negative then the distribution is negatively skewed.

iii) If $S_k = 0$ then the distribution is symmetric.

iv) Theoretically the limits of S_k are from -3 to +3.

Example 10: For the following ungrouped data find the Karl Pearson's Coefficient of Skewness.

12,18,25,15, 16, 10, 8 15, 27,14

Solution: For the Karl Pearson's Coefficient of Skewness we need to find mean, mode and standard deviation of the data.

Mean = $\bar{x} = \frac{\sum x}{n} = \frac{12+18+25+15+16+10+8+15+27+14}{10} = \frac{160}{10} = 16.$

Mode = 15 (number which repeated maximum time)

$$\sum x^2 = 144 + 324 + 625 + 225 + 256 + 100 + 64 + 225 + 729 + 196 = 2{,}888$$

$$\sigma = \sqrt{\frac{\sum x^2}{n} - (\bar{x})^2} = \sqrt{\frac{2{,}888}{10} - (15)^2} = \sqrt{288.8 - 225} = \sqrt{63.8} = 7.99$$

Therefore, the Karl Pearson's Coefficient of Skewness is

$$S_k = \frac{Mean - Mode}{standard\ deviation} = \frac{\bar{x} - M_o}{\sigma} = \frac{16 - 15}{7.99} = \frac{1}{7.99} = 0.125$$

Example 11: For the following grouped data find the Karl Pearson's Coefficient of Skewness. Also interpret the type of distribution.

C.I	0-4	4-8	8-12	12-16	16-20
F	1	3	10	4	2

Solution: First we find mean, mode and standard deviation.

C.I	F	x	fx	fx^2
0-4	1	2	2	4
4-8	3	6	18	108
8-12	10	10	100	1000
12-16	4	14	56	784
16-20	2	18	36	648
Total	20		212	2,544

Mean $=\bar{x} = \frac{\sum fx}{N} = \frac{212}{20} = 10.6.$

Standard deviation =

$$\sigma = \sqrt{\frac{\sum fx^2}{N} - (\bar{x})^2} = \sqrt{\frac{2{,}544}{20} - (10.6)^2} = \sqrt{127.2 - 112.36} = \sqrt{14.84}$$

$$= 3.85.$$

$$Mode = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

$$Mode = 8 + \frac{10 - 3}{2(10) - 3 - 4} \times 4 = 10.15.$$

Therefore, the Karl Pearson's Coefficient of Skewness is

$$S_k = \frac{Mean - Mode}{standard\ deviation} = \frac{\bar{x} - M_o}{\sigma} = \frac{10.6 - 10.15}{3.85} = \frac{0.45}{3.85} = 0.12.$$

Bowley's Coefficient of Skewness:

This measure is based on Quartiles, hence it is also known as Quartile Coefficient of Skewness. It is given by

$$\boldsymbol{S_B = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}}$$

The limits of **Bowley's Coefficient of Skewness**are between -1 to +1.

Example 12: Find the Bowley'scoefficient of Skewness for the following information are given:$Q_1 = 12.5, Q_2 = 17.2, Q_3 = 24.7$

Solution: Given that

$Q_1 = 12.5, Q_2 = 17.2, Q_3 = 24.7$

The Bowley's coefficient of Skewness is given by

$$S_B = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} = \frac{24.7 + 12.5 - 2(17.2)}{24.7 - 12.5} = \frac{2.8}{12.2} = 0.23$$

Example 13: Find the Bowley's coefficient of Skewness for the following distribution:

X	1	3	5	7	9	11
F	3	8	14	20	18	7

Solution: Let find all three quartiles for the distribution:

X	F	cf(Cummulative frequency)
1	3	3
3	8	11
5	14	25
7	20	45
9	19	64
11	7	71
Total	N =71	

Therefore,

$$Q_1 = value\ of\ \left[1\left(\frac{N+1}{4}\right)\right]^{th}\ item$$

$$= value\ of\ \left[1\left(\frac{71+1}{4}\right)\right]^{th}\ ite = value\ of\ 18th\ item = 5$$

$$Q_2 = value\ of\ \left[2\left(\frac{N+1}{4}\right)\right]^{th}\ item$$

$$= value\ of\ \left[2\left(\frac{71+1}{4}\right)\right]^{th}\ ite = value\ of\ 36th\ item = 7$$

$$Q_3 = value\ of\ \left[3\left(\frac{N+1}{4}\right)\right]^{th}\ item$$

$$= value\ of\ \left[3\left(\frac{71+1}{4}\right)\right]^{th}\ ite = value\ of\ 54th\ item = 9$$

The Bowley's coefficient of Skewness is given by

$$S_B = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} = \frac{9 + 5 - 2(7)}{9 - 5} = \frac{0}{.4} = 0.$$

Therefore, the distribution is symmetric.

4.7 KURTOSIS

Kurtosis in Greek means 'Bulginess'. In statistics, Kurtosis refers to the degree of flatness or peakedness around the mode of a frequency curve. The measure of kurtosis is with respect to a normal curve, which is accepted as a yardstick to decide the nature of other curves.
In other words measures of kurtosis tell us to what extent the given distribution is flat or peaked with respect to the standard normal curve.

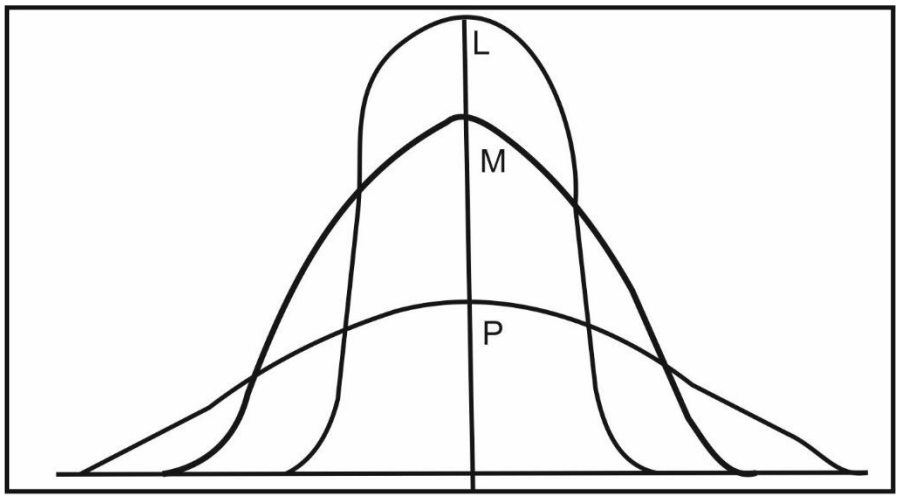


Figure 4.2

i) Normal curve is called Mesokurtic (M).

ii) Flat one is called Platykurtic (P).

iii) Peaked is called Leptokurtic (L).

Measures of Kurtosis:

The most prominent measure of kurtosis is the coefficientβ_2, given by

$$\beta_2 = \frac{\mu_4}{{\mu_2}^2}$$

Where $\mu_i's$are the moment about mean $\bar{x}$.

Bigger value of β_2 gives more peak to the distributions. For normal distribution β_2= 3.

Hence the given distribution is :

i) Leptokurtic, if β_2> 3.

ii) Mesokurtic, if β_2= 3.

iii) Platykurtic, if β_2< 3.

Example 14: for the following distribution find β_1 and β_2 and comment on the Skewness and Kurtosis of the distribution.

X	2	3	4	5
f	4	3	2	1

Solution: First calculate moments about mean for the given distribution.

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{2(4) + 3(3) + 4(2) + 5(1)}{4 + 3 + 2 + 1} = \frac{30}{10} = 3.$$

x	f	$(x-3)$	$f(x-3)$	$f(x-3)^2$	$f(x-3)^3$	$f(x-3)^4$
2	4	-1	-4	4	-4	4
3	3	0	0	0	0	0
4	2	1	2	2	2	2
5	1	2	2	4	8	16
Total	10		0	10	6	22

The central moments are given by

i) The first moment about $\bar{x}$, as

$\mu_1 = \frac{\sum f(x-\bar{x})}{N} = \frac{0}{10} = 0$.

ii) The second moment about $\bar{x}$, as

$\mu_2 = \frac{\sum f(x-\bar{x})^2}{N} = \frac{10}{10} = 1$.

iii) The third moment about $\bar{x}$, as

$\mu_3 = \frac{\sum f(x-\bar{x})^3}{N} = \frac{6}{10} = 0.6$.

iv) The fourth moments about $\bar{x}$, as

$\mu_4 = \frac{\sum f(x-\bar{x})^4}{N} = \frac{22}{10} = 2.2.$

$$\beta_1 = \frac{{\mu_3}^2}{{\mu_2}^3} = \frac{(0.6)^2}{(1)^3} = 0.36.$$

$$\beta_2 = \frac{\mu_4}{{\mu_2}^2} = \frac{2.2}{(1)^2} = 2.2$$

Since $\beta_1 \neq 0$, the curve is not symmetric Also $\mu_3 = 0.6 > 0$.

Therefore, the curve is positively skewed.

Since $\beta_2 = 2.2 < 3$.

Therefore, the curve is flat as compared to the normal curve.

Hence the distribution is platykurtic.

4.8 SOFTWARE COMPUTATION OF SKEWNESS AND KURTOSIS

To compute Skewness and Kurtosis by using different software are given below:

Sigma Magic: Using the Sigma Magic software, calculating the Skewness and Kurtosis is relatively straightforward. Just add a new Basic Statistics template to Excel by clicking on Stat > Basic Statistics. Copy and paste

the data for which you want to Skewness and Kurtosis into the input area and then click on Compute Outputs. The analysis results will include the Skewness and Kurtosis values.

Excel: You could also calculate these values in Excel by using the formula =SKEW(…) for the Skewness value, =KURT(…) for the Kurtosis value.

Minitab: If you use the Minitab software, you can copy and paste the data into Minitab and then click on Stat > Basic Statistics > Display Descriptive Statistics. Then select the data column and then click on OK. This will print out the quartiles for the sample values. If you want the Skewness and Kurtosis values, you have to go back to the menu and click on Statistics and select the checkbox next to Skewness and Kurtosis in the statistics options. Note that the values provided by Minitab may be slightly different from Excel and Sigma Magic software.

4.10 SUMMARY

In this unit, we have discussed:

- Moments and its types for ungroup and grouped data.
- The relation between raw, arbitrary and central moments.
- The effect of change of origin and scale on moments.
- Charlie's check, and Shephard's Correction for Moments.
- Skewness and about symmetry of distribution.
- Kurtosis.

4.11 EXERCISE

1.The first four moments of a distribution are 1, 4, 10 and 46 respectively. Compute the moment coefficients of skewness and kurtosis and comment upon the nature of the distribution.

2. Compute the first four central moments from the following data. Also find the two beta coefficients.

X	5	10	15	20	25	30	35
f	8	15	20	32	23	17	5

3. The first four central moments of a distribution are 0, 2.5, 0.7 and 18.75. Examine the skewness and kurtosis of the distribution.

4. Calculate first four central moments for the following distribution:

Class interval	0-4	4-8	8-12	12-16	16-20
Frequency	5	8	13	9	5

5. Find the first four arbitrary moments about A = 7 for the following:

10, 5, 8, 7, 2, 3, 12, 14

6. Find raw moment for the following data:

C.I	5-10	10-15	15-20	20-25	25-30
f	3	4	7	4	2

7. The first four central moments of a distribution are 0, 15, 36, 78. If the mean of the distribution is 8, find the moments about A = 5.

8. The first four raw moments about origin are 4, 16, 33, 89. Find mean and the first four central moments.

9. For the following data verify Charlie's check :

C.I	2-8	8-14	14-20	20-26
f	1	4	3	2

10. Find the first four central moments using coding method, also find Sheppard's correction for moments.

C.I	0-5	5-10	10-15	15-20	20-25	25-30
f	3	8	12	13	7	2

11. For the following data , find Karl Pearson's coefficient of Skewness and also find the type of distribution:

i) 12,15,17, 12,8,25,16,6,7,41

ii) 3,7,8, 12, 15

X	1	2	3	4	5	6	7
f	2	8	12	15	18	9	6

C.I	0-2	2-4	4-6	6-8	8-10
frequency	4	7	13	10	6

12. Find the Bowely's coefficient of Skewness for each of the following:

i) $Q_1 = 165.5,\ Q_2 = 184.3,\ Q_3 = 196.7.$

ii) 2,8,7,12,14,17,20.

X	1	2	3	4	5	6	7
F	2	8	12	15	18	9	6

4.12 LIST OF REFERENCES

- Statistics by Murry R. Spiegel, Larry J. Stephens. Publication McGRAWHILL INTERNATIONAL.
- Fundamental Statistic by S.C. Gupta

*****

5

ELEMENTARY PROBABILITY THEORY

Unit Structure

5.1 OBJECTIVE

After going through this unit, you will able to :

- Determine the probability of different experimental results.
- Explain the concept of probability.
- Calculate probability for simple, compound and complimentary events.
- Conditional probability and its examples.
- Independent events and multiplication theorem of probability.
- Probability distribution and its Expected value of probability distribution.
- Combination and Stirling's number approximation.
- Relations between probability and set theory with help of Venn diagram.

5.2 INTRODUCTION

Some times in daily life certain things come to mind like "I will be success today', I will complete this work in hour, I will be selected for job and so

on. There are many possible results for these things but we are happy when we get required result. Probability theory deals with experiments whose outcome is not predictable with certainty. Probability is very useful concept. These days many field in computer science such as machine learning, computational linguistics, cryptography, computer vision, robotics other also like science, engineering, medicine and management.

Probability is mathematical calculation to calculate the chance of occurrence of some happening, we need some basic concept on random experiment, sample space, and events.

Basic concept of probability:

Random experiment: When experiment can be repeated any number of times under the similar conditions but we get different results on same experiment, also result is not predictable such experiment is called random experiment. For.e.g. A coin is tossed, A die is rolled and so on.

Outcomes: The result which we get from random experiment is called outcomes of random experiment.

Sample space: The set of all possible outcomes of random experiment is called sample space. The set of sample space is denoted by S and number of elements of sample space can be written as $n(S)$. For e.g. A die is rolled, we get $= \{1,2,3,4,5,6\}$, $n(S) = 6$.

Events: Any subset of the sample space is called an event. Or a set of sample point which satisfies the required condition is called an events. Number of elements in event set is denoted by $n(E)$.For example in the experiment of throwing of a dia. The sample space is

S = {1, 2, 3, 4, 5, 6 } each of the following can be an event :

i) A: even number i.e. A = { 2, 4, 6} ii) B: multiple of 3 i.e. B = { 3, 6} iii) C: prime numbers i.e. C = { 2, 3, 5}.

Types of events:

Impossible event: An event which does not occurred in random experiment is called impossible event. It is denoted by $\emptyset$ set. i. e. $n(\emptyset) = 0$. For example getting number 7 when die is rolled. The probability measure assigned to impossible event is Zero.

Equally likely events: when all events get equal chance of occurrences is called equally likely events. For e.g. Events of occurrence of head or tail in tossing a coin are equally likely events.

Certain event: An event which contains all sample space elements is called certain events. i.e.$n(A) = n(S)$.

Mutually exclusive events: Two events A and B of sample space S, it does not have any common elements are called mutually exclusive events. In the experiment of throwing of a die A: number less than 2 , B: multiple of 3. There fore $n(A \cap B) = 0$

Exhaustive events: Two events A and B of sample space S, elements of event A and B occurred together are called exhaustive events. For e.g. In a

thrown of fair die occurrence of even number and occurrence of odd number are exhaustive events. There fore $n(A \cup B) = 1$.

Complement event: Let S be sample space and A be any event than complement of A is denoted by $\bar{A}$ is set of elements from sample space S, which does not belong to A. For e.g. if a die is thrown, S = {1, 2, 3, 4, 5, 6} and A: odd numbers, A = {1, 3, 5}, then $\bar{A} = \{2,4,6\}$.

5.3 DEFINITIONS OF PROBABILITY

Probability: For any random experiment, sample space S with required chance of happing event E than the probability of event E is define as

$$P(E) = \frac{n(E)}{n(S)}$$

Basic properties of probability:

1) The probability of an event E lies between 0 and 1. i.e. $0 \leq P(E) \leq 1$.
2) The probability of impossible event is zero. i.e. $P(\emptyset) = 0$.
3) The probability of certain event is unity. i.e. $P(E) = 1$.
4) If A and B are exhaustive events than probability of $P(A \cup B) = 1$.
5) If A and B are mutually exclusive events than probability of $P(A \cap B) = 0$.
6) If A be any event of sample space than probability of complement of A is given by $P(A) + P(\bar{A}) = 1 \Rightarrow \therefore P(\bar{A}) = 1 - P(A)$.

Probability Axioms:

Let S be a sample space. A probability function P from the set of all event in S to the set of real numbers satisfies the following three axioms for all events A and B in S.

i) $P(A) \geq 0$.

ii) $P(\emptyset) = 0$ and $P(S) = 1$.

iii) If A and B are two disjoint sets i.e. $A \cap B = \emptyset$) than the probability of the union of A and B is $P(A \cup B) = P(A) + P(B)$.

Theorem: Prove that for every event A of sample space S, $0 \leq P(A) \leq 1$.

Proof: $S = A \cup \bar{A}$, $\emptyset = A \cap \bar{A}$.

$$\therefore 1 = P(S) = P(A \cup \bar{A}) = P(A) + P(\bar{A})$$

$$\therefore 1 = P(A) + P(\bar{A})$$

$\Rightarrow P(A) = 1 - P(\bar{A})$ or $P(\bar{A}) = 1 - P(A)$.

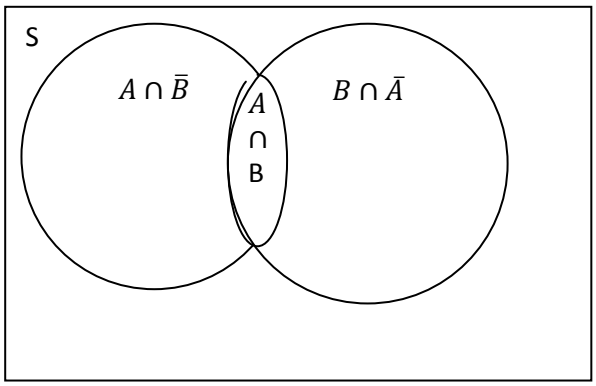
If $P(A) \geq 0$. than $P(\bar{A}) \leq 1$.

$\therefore$ for every event A; $0 \leq P(A) \leq 1$.

Addition theorem of probability:

Theorem: If A and B are two events of sample space S, then probability of union of A and B is given by $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof: A and B are two events of sample space S.



Now from diagram probability of union of two events A and B is given by,

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

But $P(A \cap \bar{B}) = P(A) - P(A \cap B)$ and $P(B \cap \bar{A}) = P(B) - P(A \cap B)$.

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

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

Note: The above theorem can be extended to three events A, B and C as shown below:

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

Example 1: A bag contains 4 black and 6 white balls; two balls are selected at random. Find the probability that balls are i) both are different colors. ii) both are of same colors.

Solution: Total number of balls in bag = 4 blacks + 6 white = 10 balls
To select two balls at random, we get
$n(S) = C(10,2) = 45.$

i) A be the event to select both are different colors.

$$\therefore n(A) = C(4,1) \times C(6,1) = 4 \times 6 = 24.$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{24}{45} = 0.53.$$

ii) To select both are same colors.

Let Abe the event to select both are black balls

$$n(A) = C(4,2) = 6$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{6}{45}$$

Let B be the event to select both are white balls.

$$n(B) = C(6,2) = 15$$

$P(B) = \frac{n(B)}{n(S)} = \frac{15}{45}$.

A and B are disjoint event.

∴ The required probability is

$$P(A \cup B) = P(A) + P(B) = \frac{6}{45} + \frac{15}{45} = \frac{21}{45} = 0.467.$$

Example 2: From 40 tickets marked from 1 to 40, one ticket is drawn at random. Find the probability that it is marked with a multiple of 3 or 4.

Solution: From 40 tickets marked with 1 to 40, one ticket is drawn at random

$$n(S) = C(40,1) = 40$$

it is marked with a multiple of 3 or 4, we need to select in two parts.

Let A be the event to select multiple of 3,

i.e. A = { 3,6,9,….,39}

$$n(A) = C(13,1) = 13$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{13}{40}$$

Let B be the event to select multiple of 4.

i.e. B = {4,8,12, …., 40}

$$n(B) = C(10,1) = 10$$

$$P(B) = \frac{n(B)}{n(s)} = \frac{10}{40}.$$

Here A and B are not disjoint.

$A \cap B$be the event to select multiple of 3 and 4.

i.e. $A \cap B$ = {12,24,36}

$$n(A \cap B) = C(3,1) = 3$$

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{3}{40}$$

∴ The required probability is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{13}{40} + \frac{10}{40} - \frac{3}{40} = \frac{20}{40} = 0.5.$$

Example 3: If the probability is 0.45 that a program development job; 0.8 that a networking job applicant has a graduate degree and 0.35 that applied for both. Find the probability that applied for atleast one of jobs. If number of graduate are 500 then how many are not applied for jobs?

Solution: Let Probability of program development job= $P(A) = 0.45$.

Probability of networking job= $P(B) = 0.8$.

Probability of both jobs = $P(A \cap B) = 0.35$.

Probability of atleast one i.e. to find $P(A \cup B)$.

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

$$P(A \cup B) = 0.45 + 0.8 - 0.35 = 0.9$$

Now there are 500 application, first to find probability that not applied for job.

$$P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - 0.9 = 0.1$$

Number of graduate not applied for job = $0.1 \times 500 = 50$.

Check your Progress:

1. A card is drawn from pack of 52 cards at random. Find the probability that it is a face card or a diamond card.
2. If $P(A) = \frac{3}{8}$and $(B) = \frac{5}{8}$, $P(A \cup B) = \frac{7}{8}$ than find i) $P(\overline{A \cup B})$ ii) $P(A \cap B)$.
3. In a class of 60 students, 50 passed in computers, 40 passed in mathematics and 35 passed in both. What is the probability that a student selected at random has i) Passed in atleast one subject, ii) failed in both the subjects, iii) passed in only one subject.

5.4 CONDITIONAL PROBABILITY

In many case we come across occurrence of an event A and for the same are required to find out the probability of occurrence an event B which depend on event A . This kind of problem is called conditional probability problems.

Definition: Let A and B be two events. The conditional probability of event B, if an event A has occurred is defined by the relation,

$P(B|A) = \frac{P(B \cap A)}{P(A)}$if and only if $P(A) > 0$.

In case when $P(A) = 0, P(B|A)$ is not define because $P(B \cap A) = 0$ and $P(B|A) = \frac{0}{0}$ which is an indeterminate quantity.

Similarly, Let A and B be two events. The conditional probability of event A, if an event B has occurred is defined by the relation,

$P(A|B) = \frac{P(A \cap B)}{P(B)}$ If and only if $P(B) > 0$.

Example 4: A pair of fair dice is rolled. What is the probability that the sum of upper most face is 6, given that both of the numbers are odd?

Solution: A pair of fair dice is rolled, therefore $n(S) = 36$.

A to select both are odd number, i.e. A = {(1,1), (1,3), (1,5), (3,1), (3,3), (3,5), (5,1),(5,3), (5,5)}.

$$P(A) = \frac{n(A)}{n(S)} = \frac{9}{36}$$

B is event that the sum is 6, i.e. B = { ((1,5),(2,4), (3,3),(4,2), (5,1)}.

$$P(B) = \frac{n(B)}{n(S)} = \frac{5}{36}$$

$A \cap B$ = { (1,5), (3,3), (5,1)}

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{3}{36}$$

By the definition of conditional probability,

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{3/36}{9/36} = \frac{1}{3}.$$

Example 5: If A and B are two events of sample space S, such that $P(A) = 0.85,\ P(B) = 0.7$and $P(A \cup B) = 0.95$. Find i) $P(A \cap B)$, ii) $P(A|B)$, iii) $P(B|A)$.

Solution: Given that $P(A) = 0.85,\ P(B) = 0.7$and $P(A \cup B) = 0.95$.

i) By Addition theorem,

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

$0.95 = 0.85 + 0.7 - P(A \cap B)$

$P(A \cap B) = 1.55 - 0.95 = 0.6.$

ii) By the definition of conditional probability ,

$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.6}{0.7} = 0.857.$

iii) $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.6}{0.85} = 0.706$

Example 6: An urn *A* contains 4 Red and 5 Green balls. Another urn *B* contains 5 Red and 6 Green balls. A ball is transferred from the urn *A* to the urn *B*, then a ball is drawn from urn *B*. find the probability that it is Red.

Solution: Here there are two cases of transferring a ball from urn A to B.

Case I: When Red ball is transferred from urn A to B.

There for probability of Red ball from urn A is $P(R_A) = \frac{4}{9}$

After transfer of red ball, urn B contains 6 Red and 6 Green balls.

Now probability of red ball from urn B = $P(R_B|R_A) \times P(R_A) = \frac{6}{12} \times \frac{4}{9} = \frac{24}{108}$.

Case II: When Green ball is transferred from urn A to B.

There for probability of Green ball from urn A is $P(G_A) = \frac{5}{9}$

After transfer of red ball, urn B contains 5 Red and 7 Green balls.

Now probability of red ball from urn B = $P(R_B|G_A) \times P(G_A) = \frac{5}{12} \times \frac{5}{9} = \frac{25}{108}$.

Therefore required probability =$\frac{24}{108} + \frac{25}{108} = \frac{49}{108} = 0.4537$.

Check your progress:

1. A family has two children. What is the probability that both are boys, given at least one is boy?
2. Two dice are rolled. What is the condition probability that the sum of the numbers on the dice exceeds 8, given that the first shows 4?
3. Consider a medical test that screens for a COVID-19 in 10 people in 1000. Suppose that the false positive rate is 4% and the false negative rate is 1%. Then 99% of the time a person who has the condition tests positive for it, and 96% of the time a person who does not have the condition tests negative for it. a) What is the probability that a randomly chosen person who tests positive for the COVID-19 actually has the disease? b) What is the probability that a randomly chosen person who tests negative for the COVID-19 does not indeed have the disease?

5.4.1 Independent and Dependent Events, Mutually Exclusive Events:

Independent events:

Two events are said to be independent if the occurrence of one of them does not affect and is not affected by the occurrence or non-occurrence of other.

i.e. $P(B/A) = P(B)$ or $P(A/B) = P(A)$.

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

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

Where $P(B/A)$ denotes the conditional probability of event B given that event A has already occurred.

OR

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

Where $P(A/B)$ denotes the conditional probability of event A given that event B has already occurred.

11.5.1 For Independent events multiplication theorem:

If A and B are independent events then multiplication theorem can be written as,

$$\boldsymbol{P(A \cap B) = P(A)P(B)}$$

Proof. Multiplication theorem can be given by,

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

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

By definition of independent events, $P\left(B/A\right) = P(B)$ or $P\left(A/B\right) = P(A)$.

$\therefore \boldsymbol{P(A \cap B) = P(A)P(B)}$.

Note:

1) If A and B are independent event then, $\bar{A}$ and $\bar{B}$ are independent event.

2) If A and B are independent event then, $\bar{A}$ and B are independent event.

3) If A and B are independent event then, A and $\bar{B}$ are independent event.

Example 7: Manish and Mandar are trying to make Software for company. Probability that Manish can be success is $\frac{1}{5}$ and Mandar can be success is $\frac{3}{5}$, both are doing independently. Find the probability that i) both are success. ii) Atleast one will get success. iii) None of them will success. iv) Only Mandar will success but Manish will not success.

Solution: Let probability that Manish will success is $P(A) = \frac{1}{5} = 0.2$.

Therefore probability that Manish will not success is $P(\bar{A}) = 1 - P(A) = 1 - 0.2 = 0.8$.

Probability that Mandar will success is $P(B) = \frac{3}{5} = 0.6$.

Therefore probability that Mandar will not success is $P(\bar{B}) = 1 - P(B) = 1 - 0.6 = 0.4$.

i) Both are success i.e. $P(A \cap B)$.

$P(A \cap B) = P(A) \times P(B) = 0.2 \times 0.6 = 0.12$ ∵A and B are independent events.

ii) Atleast one will get success. i.e. $P(A \cup B)$

By addition theorem,

$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.2 + 0.6 - 0.12 = 0.68$.

iii) None of them will success.$P(\overline{A \cup B})$ or $P(\bar{A} \cap \bar{B})$

[ByDeMorgan's law both are same]

$P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - 0.68 = 0.32$.

Or

If A and B are independent than $\bar{A}$ and $\bar{B}$ are also independent.

$$P(\bar{A} \cap \bar{B}) = P(\bar{A}) \times P(\bar{B}) = 0.8 \times 0.4 = 0.32.$$

iv) Only Mandar will success but Manish will not success. i.e. $P(\bar{A} \cap B)$.

$$P(\bar{A} \cap B) = P(\bar{A}) \times P(B) = 0.8 \times 0.6 = 0.48$$

Example 8: 50 coding done by two students A and B, both are trying independently. Number of correct coding by student A is 35 and student B is 40. Find the probability of only one of them will do correct coding.

Solution: Let probability of student A get correct coding is $P(A) = \frac{35}{50} = 0.7$

Probability of student A get wrong coding is $P(\bar{A}) = 1 - 0.7 = 0.3$

Probability of student B get correct coding is $P(B) = \frac{40}{50} = 0.8$

Probability of student B get wrong coding is $P(\bar{B}) = 1 - 0.8 = 0.2.$

The probability of only one of them will do correct coding.

i.e. A will correct than B will not or B will correct than A will not.

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

$$= 0.7 \times 0.2 + 0.8 \times 0.3 = 0.14 + 0.24$$

$$= 0.38$$

Example 9: Given that $P(A) = \frac{3}{7},\ P(B) = \frac{2}{7}$, if A and B are independent events than find i) $P(A \cap B)$, ii) $P(\bar{B})$, iii) $P(A \cup B)$, iv) $P(\bar{A} \cap \bar{B})$.

Solution: Given that $P(A) = \frac{3}{7},\ P(B) = \frac{2}{7}$.

i) A and B are independent events,

$$\therefore P(A \cap B) = P(A) \times P(B) = \frac{3}{7} \times \frac{2}{7} = \frac{6}{49} = 0.122$$

ii) $P(\bar{B}) = 1 - P(B) = 1 - \frac{2}{7} = \frac{5}{7} = 0.714.$

iii) By addition theorem,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{3}{7} + \frac{2}{7} - \frac{6}{49} = \frac{29}{49} = 0.592.$$

iv) $P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - 0.592 = 0.408.$

Check your progress:

1. If $P(A) = \frac{2}{5}$, $P(B) = \frac{1}{3}$ and if A and B are independent events, find $(i)P(A \cap B),\ (ii)P(A \cup B),\ (iii)P(\bar{A} \cap \bar{B})$.
2. The probability that A , B and C can solve the same problem independently are $\frac{1}{3}, \frac{2}{5} and \frac{3}{4}$ respectively. Find the probability that i) the problem remain unsolved, ii) the problem is solved, iii) only one of them solve the problem.
3. The probability that Ram can shoot a target is $\frac{2}{5}$ and probability of Laxman can shoot at the same target is $\frac{4}{5}$. A and B shot independently. Find the probability that (i) the target is not shot at all, (ii) the target is shot by at least one of them. (iii) the target shot by only one of them. iv) target shot by both.

5.5 PROBABILITY DISTRIBUTIONS

In order to understand the behavior of a random variable, we may want to look at its average value. For probability we need to find Average is called expected value of random variable X. for that first we have to learn some basic concept of random variable.

Random Variable: A probability measurable real valued functions, say X, defined over the sample space of a random experiment with respective probability is called a random variable.

Types of random variables: There are two type of random variable.

Discrete Random Variable: A random variable is said to be discrete random variable if it takes finite or countably infinite number of values. Thus discrete random variable takes only isolated values.

Continuous Random variable: A random variable is continuous if its set of possible values consists of an entire interval on the number line.

Probability Distribution of a random variable: All possible values of the random variable, along with its corresponding probabilities, so that$\sum_{\square=1}^{\square} \square_{\square} = 1$, is called a probability distribution of a random variable.

The probability function always follows the following properties:

i) $P(x_i) \geq 0$ for all value of i.

ii) $\sum_{i=1}^{n} P_i = 1$.

The set of values x_i with their probability P_i constitute a discrete probability distribution of the discrete variable X.

For e.g. Three coins are tossed, the probability distribution of the discrete variable X is getting head.

X= x_i	0	1	2	3
$P(x_i)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

5.6 MATHEMATICAL EXPECTATION

All the probability information of a random variable is contained in probability mass function for random variable, it is often useful to consider various numerical characteristics of that random variable. One such number is the expectation of a random variable.

If random variable X takes values $x_1, x_2, \ldots\ldots x_n$ with corresponding probabilities $P_1, P_2, \ldots\ldots P_n$ respectively, then expectation of random variable X is

$E(X) = \sum_{i=1}^{n} p_i x_i$ where $\sum_{i=1}^{n} P_i = 1$

Example 10: In Vijay sales every day sale of number of laptops with his past experience the probability per day are given below:

No. of laptop	0	1	2	3	4	5
Probability	0.05	0.15	0.25	0.2	0.15	0.2

Find his expected number of laptops can be sale?

Solution: Let X be the random variable that denote number of laptop sale per day.

To calculate expected value, $E(X) = \sum_{i=1}^{n} p_i x_i$

$$E(X) = (0 \times 0.05) + (1 \times 0.15) + (2 \times 0.25) + (3 \times 0.2) + (4 \times 0.15) + (5 \times 0.2)$$

$$E(X) = 2.85 \sim 3$$

Therefore expected number of laptops sale per day is 3.

Example 11: A random variable X has probability mass function as follow:

X=x_i	-1	0	1	2	3
P(x_i)	K	0.2	0.3	2k	2k

Find the value of k, and expected value.

Solution: A random variable X has probability mass function,

$\sum_{i=1}^{n} P_i = 1.$

$\Rightarrow$ k + 0.2 + 0.3 + 2k +2k = 1

$\Rightarrow$5k = 0.5

$\Rightarrow$ k = 0.1

Therefore the probability distribution of random variable X is

X=x_i	-1	0	1	2	3
P(x_i)	0.1	0.2	0.3	0.2	0.2

To calculate expected value, $E(X) = \sum_{i=1}^{n} p_i x_i$

$E(X) = (-1 \times 0.1) + (0 \times 0.2) + (1 \times 0.3) + (2 \times 0.2) + (3 \times 0.2) = 1.2$.

Example 12: A box contains 5 white and 7 black balls. A person draws 3 balls at random. He gets Rs. 50 for every white ball and losses Rs. 10 every black ball. Find the expectation of him.

Solution: Total number of balls in box = 5 white + 7 black = 12 balls.

To select 3 balls at random, $n(s) = C(12,3) = \frac{12 \times 11 \times 10}{3 \times 2 \times 1} = 220.$

Let A be the event getting white ball.

A takes value of 0, 1, 2 and 3 white ball.

Case I : no white ball. i.e. A = 0,

$$P(A=0)=\frac{C(7,3)}{220}=\frac{35}{220}$$

Case II: one white ball i.e. A = 1,

$$P(A=1)=\frac{C(5,1)\times C(7,2)}{220}=\frac{105}{220}$$

Case III: two white balls i.e. A = 2,

$$P(A=2)=\frac{C(5,2)\times C(7,1)}{220}=\frac{70}{220}$$

Case IV: three white balls i.e. A = 3,

$$P(A=3)=\frac{C(5,3)}{220}=\frac{10}{220}$$

Now let X be amount he get from the game.

Therefore the probability distribution of X is as follows:

X= x_i	-30	30	90	150
P(x_i)	$\frac{35}{220}$	$\frac{105}{220}$	$\frac{70}{220}$	$\frac{10}{220}$

To calculate expected value, $E(X)=\sum_{i=1}^{n} p_i x_i$

$E(X)=\left(-30\times\frac{35}{220}\right)+\left(30\times\frac{105}{220}\right)+\left(90\times\frac{70}{220}\right)+\left(150\times\frac{10}{220}\right)$ =Rs. 45.

5.7 COMBINATORIAL ANALYSIS

Multiplication Rule:

If the procedure can be broken into first and second stages, and if there are m possible outcomes for the first stage and for each of these outcomes, there are n possible outcomes for second stage, then the total procedure can be carried out in the designate order, in $m\times n$ ways.

This principle can be extended to a general form as follows:

Theorem : If a process consists of *n* steps, and

i) The first step can be performed by n_1 ways.

ii) The second step can be performed by n_2 ways.

iii) The i^{th} step can be performed by n_i ways.

Then the whole process can be completed by $n_1\times n_2\times \ldots\ldots\times n_i$ different ways.

Example 13: There are 8 men and 7 women in a drama company. How many way the director has to choose a couple to play lead roles in a stage show?

Solution: The director can choose a man (task 1) in 8 ways and then a woman (task 2) in 7 ways. Then by multiplication rule he can choose a couple from $8 \times 7 = 56$ ways.

Example 14: How many four digits numbers can be formed contains each of the digits 7, 8, and 9 exactly once?

Solution: To construct four digits number we have four places.

_____ _____ _____ _____

Thousand place Hundred place Ten place Unit place

First for '7' there are 4 places, for '8' there are 3 places and for '9' there are 2 places. For last digit, we can choose any of 0,1,2,3,4,5,6 so there will be 7 digits.

Thus these can be done by $4 \times 3 \times 2 \times 7 = 168$ ways.

Example 15: To generate typical personal identification number (PIN) is a sequence of any four symbols chosen from the letters in the alphabet and the digits , How many different PIN's are generated?

i) repetition is not allowed.

ii) repetition is allowed.

Solution: There are 26 letters of alphabets and 10 digits. Total different symbols are 36.

i) Repetition is not allowed:

There are four place to generate PIN with four symbols,

First place can be filled by 36 ways, second place can be filled by 35 ways, third place can be filled by 34 ways and last fourth place can be filled by 33 ways.

By the multiplication rule,

Therefore these can be done by $36 \times 35 \times 34 \times 33 = 1{,}413{,}720$ ways.

ii) Repetition is allowed:

Since repetition is allowed, so each place can be filled by 36 ways,
By multiplication rule,
These can be done by $36 \times 36 \times 36 \times 36 = 1{,}679{,}616$ ways.

Check your progress:

1. A license plate can be made by 2 letters followed by 3 digits. How many different license plates can be made if i) repetition is not allowed. ii) Repetition is allowed.
2. Mr. Modi buying a personal computer system is offered a choice of 4 models of basic units, 2 models keyboard, and 3 models of printer. How many distinct systems can be purchased?

Counting elements of disjoint sets with Addition Rule:

In above section we have discussed counting problem that can be solved using possibility tree. Here we discuss counting problem that can be solved using the operation sets like union , intersection and the difference between two sets.

The addition rule:

If a task can be performed in m ways and another task in n ways assuming that these two tasks cannot perform simultaneously, then the performing either task can be accomplished in any one of the $m + n$ ways.

In general from as follows:

If there are$n_1, n_2, n_3, \ldots\ldots, n_m$ different objects in m different sets respectively and the sets are disjoint, then the number of ways to select an object from one of the m sets is $n_1 + n_2 + n_3 + \ldots\ldots + n_m$

Example 16: How many different number of signals that can be sent by 5 flags of different colours taking one or more at a time ?

Solution: Let number of signal made by one colour flag $= 5$ ways.

Number of signal made by two colours flag $= 5 \times 4 = 20$ ways.

Number of signal made by three flag colours$= 5 \times 4 \times 3 = 60$ ways.

Number of signal made by four flag colours$= 5 \times 4 \times 3 \times 2 = 120$ ways.

Number of signal made by five flag colours$= 5 \times 4 \times 3 \times 2 \times 1 = 120$ ways.

Using Addition rule we get,

Therefore total number of signals $= 5 + 20 + 60 + 120 + 120 = 325$ ways.

Example 17: There are 4 different English books, 5 different Hindi books and 7 different Marathi books. How many ways are there to pick up an pair of two books not both with the same subjects?

Solution: One English and one Hindi book is chosen, that selection can be done by $= 4 \times 5 = 20$ ways.

One English and one Marathi book is chosen, that selection can be done by $= 4 \times 7 = 28$ ways.

One Hindi and one Marathi book is chosen, that selection can be done by $= 5 \times 7 = 35$ ways.

These three types of selection are disjoint, therefore by addition rule,

Total selection can be done by $= 20 + 28 + 35 = 83$ ways.

Additive Principle with Disjoint sets:

Given two sets A and B, both sets are disjoint i.e. if $A \cap B = \emptyset$, than $|A \cup B| = |A| + |B|$.

Example 18: In college 200 students visit to canteen every day of which 80 likes coffee and 70 likes tea. If no one student like both than find i) number of students like atleast one of them? ii) number of students like none of them?

Solution: Total number of students = 200

Total number of students who like coffee = $|A| = 80$.

Total number of students who like tea = $|B| = 70$.

Total number of students like atleast one $|A \cup B| = |A| + |B| = 80 + 70 = 150$.

Total number of students like none of them $= 200 - 150 = 50$.

Definition: An r-combination of n distinct objects is an unordered selection, or subset, of r out of the n objects. We use $C(n,r) or {}_r^nC$ to denote the number of r-combinations. This number is called as binomial number.

If $x_1, x_2, x_3, \ldots\ldots, x_n$ are n distinct objects, and r is any integer, with $1 \leq r \leq n$. Therefore selecting r-objects from n objects is given by

$$C(n,r) = \frac{n!}{r!\,(n-r)!}$$

Example 19: How many elements of set 3-bit string with weight 2?

Solution: there are 3-bit with weight 2, i.e. $n = 3,\ r = 2$.

These can be done by = $C(n,r) = C(3,2) = 3$.

Therefore the bit string is 011, 101, 110.

Example 20: A bag contains 4 red marbles and 5 green marbles. Find the number of ways that 4 marbles can be selected from the bag, if selection contain i) No restriction of colors. ii) all are of same colors.

Solution: Total number of marbles: 4 Red + 5 Green = 9 marbles

To select 4 marbles from the bag with condition,

i) No restriction of colors:

These can be done by :$C(9,4) = 126$ ways.

ii) All are of same colors:

First select the colors by : $C(2,1) = 2$.

If all is Red in colors than these can be done by $= C(4,4) = 1$.

If all is Green in colors then these can be done by $= C(5,4) = 5$.

Therefore total number of ways $= 2 \times 1 \times 5 = 10$ ways.

Example 21: There are 10 members in a society who are eligible to attend annual meeting. Find the number of ways a 4 members can be selected that

i) No restriction

ii) If 2 of them will not attend meeting together.

iii) If 2 of them will always attend meeting together.

Solution:

i) To select 4 members from 10 members, it can be done by $= C(10,4) = 210$ ways.

ii) If 2 of them will not attend meeting together,

Let A and B denote the 2 members who will not attend meeting together.

i.e. A or B but not both are together , these can be done by $= 2 \times C(8,3) = 112$ ways.

It possible that both will not attend meeting, i.e. Neither A nor B will attend meeting, these can be done by $= C(8,2) = 28$ ways.

Therefore total number of ways = 112 + 28 =140 ways.

iii) If 2 of them will attend meeting together,

Let A and B denote the 2 members who will attend meeting together.

i.e. A or B $= C(8,2) = 28$ ways.

It possible that both will not attend meeting, i.e. Neither A nor B will attend meeting, these can be done by $= C(8,4) = 70$ ways.

Therefore total number of ways = 28 +70 = 88 ways.

Example 21: How many diagonal has a regular polygon with n sides?

Solution: The regular polygon with n sides has n vertices. Any two vertices determine either a side or diagonal. Therefore these can be done by $= C(n,2) = \frac{n(n-1)}{2}$. But there are n sides which are not diagonal.

Therefore total number of diagonals are$= \frac{n(n-1)}{2} - n = \frac{n^2-n}{2} - \frac{2n}{2} = \frac{n^2-3n}{2} = \frac{n(n-3)}{2}$ diagonals.

r-combinations with Repetition Allowed:

Till now, we have seen the formula for the number of combinations when r objects are chosen from the collection of n distinct objects. The following results is very important to find the number of selection of n objects when not all n are distinct.

The number of selection with repetition of r objects chosen from n types of objects is

$$C(n+r-1,r)$$

Example 22: How many ways are there to fill a box with a dozen marbles chosen five different colors of marbles with the requirement that at least one fruit of each colors is picked?

Solution: One can pick one marble of each colors and then the remaining seven marbles in any way. There is no choice in picking one marble of

each type. The choice occurs in picking the remaining 7 marbles from 5 colors. By the result of r-combination with repetition allowed,

These can be done by $= C(5 + 7 - 1, 7) = C(11,7) = 330$ ways.

Example 23: How many solution does the following equation $x_1 + x_2 + x_3 + x_4 = 15$ have $x_1,\ x_2,\ x_3,$ and x_4 are non-negative integers?

Solution: Assume we have four types of unknown $x_1,\ x_2,\ x_3,$ and x_4 . There are 15 items or units (since we are looking for an integer solution). Every time an item is selected it adds one to the type it picked it up. Observe that a solution corresponds to a way of selecting 15 items from set of four elements. Therefore, it is equal to r-combinations with repetition allowed from set with four elements, we have

$$C(4 + 15 - 1, 15) = C(18, 15) = C(18, 3) = \frac{18 \times 17 \times 16}{3 \times 2 \times 1} = 816$$

Example 24: In how many ways can a teacher choose one or more students from 5 students?

Solution: Let set of student are 5, therefore total number of subsets are $2^5 = 32$.

To select one or more students, we must deleted empty set,.

Therefore total number of selection = 32 – 1= 31 ways.

5.8 COMBINATIONS, STIRLING'S APPROXIMATION TO N!

A helpful and commonly used approximate relationship for the evaluation of the factorials of large numbers is Stirling's approximation. It is a good approximation, leading to accurate results even for small values of*n*. it is given by

$$n! \approx \frac{n^n}{e^n}\sqrt{2\pi n}$$

Where *e* = 2:71828 is the natural base of logarithms.

5.9 RELATION OF PROBABILITY TO POINT SET THEORY, EULER OR VENN DIAGRAMS AND PROBABILITY

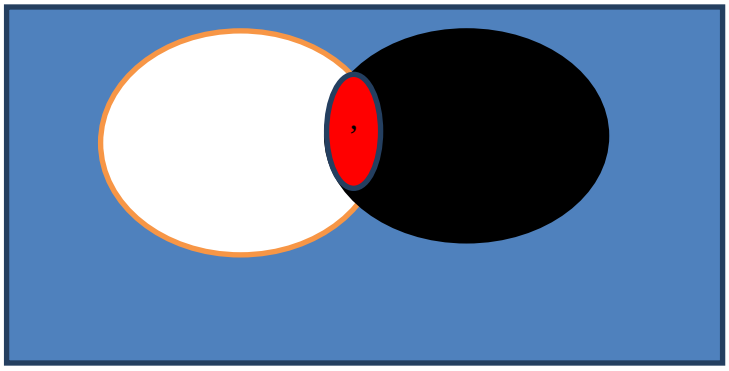
a) Relation of Probability to Point Set Theory:

- In discrete probability we assume well defined experiment such as flipping a coin or rolling a die. Each individual result which could occur is called an outcome. The set of all outcomes is called sample space, and any subset of the sample space is called an event.
- The union of two or more sets is the set that contains all the elements of the two or more sets. Union is denoted by the symbol ∪.The general

probability addition rule for the union events states that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Where $A \cap B$ is the intersection of the two sets.

Euler or Venn Diagrams and Probability:

In probability, a Venn diagram is a figure with one or more circles inside a rectangle that describes logical relation between events. The rectangle in a Venn diagram represents the sample space or the universal set, that is , the set of all possible outcomes. A circle inside the rectangle represents an event, that is, a subset of the sample space. We consider the following Venn diagram involving two events, A and B.



In the above diagram, we have two events A and B within the sample space S(or Universal set)

White and Red region represent event A, Black and Red region represent event B, only Red region represent $A \cap B$, and White, Black and Red region together represents $A \cup B$. Also Blue region represent $\overline{A \cup B}$.

- If the circles do not overlap than A and B are mutually exclusive events.

  i.e$A \cap B = \emptyset$.

- If the circles are overlap than A and B are intersecting each other. i.e$A \cap B \neq \emptyset$.

- The region outside both the circles but within the rectangle represents complement of union of both the events. i.e.$\overline{A \cup B}$.

Within each divided region of a Venn diagram, we can add data in one of the following ways:

- The outcomes of the event,
- The number of outcomes in the event,
- The probability of the event.

5.10 LET US SUM UP

In this unit we have learn:

- Basic concept and Definitions of Probability.
- Conditional Probability.
- Independent and Dependent Events, Mutually Exclusive Events.
- Probability Distributions for discrete distribution.
- Mathematical Expectation for probability distribution.
- Relation between Population, Sample Mean, and Variance.
- Combinations, Stirling's Approximation to n!.
- Relations of Probability to Point Set Theory with represent Euler or Venn Diagrams.

5.11 UNIT END EXERCISES

1. A card is drawn at random from well shuffled pack of card find the probability that it is red or king card.
2. There are 30 tickets bearing numbers from 1 to 15 in a bag. One ticket is drawn from the bag at random. Find the probability that the ticket bears a number, which is even, or a multiple of 3.
3. In a group of 200 persons, 100 like sweet food items, 120 like salty food items and 50 like both. A person is selected at random find the probability that the person (i). Like sweet food items but not salty food items (ii). Likes neither.
4. A bag contains 7 white balls & 5 red balls. One ball is drawn from bag and it is replaced after noting its color. In the second draw again one ball is drawn and its color is noted. The probability of the event that both the balls drawn are of different colors.
5. The probability of A winning a race is $\frac{1}{3}$& that B wins a race is $\frac{3}{5}$. Find the probability that (a). either of the two wins a race.b), no one wins the race.
6. Three machines A, B & C manufacture respectively 0.3, 0.5 & 0.2 of the total production. The percentage of defective items produced by A, B & C is 4 , 3 & 2 percent respectively. for an item chosen at random , what is the probability it is defective.
7. An urn A contains 3 white &5 black balls. Another urn B contains 5 white & 7 black balls. A ball is transferred from the urn A to the urn B, then a ball is drawn from urn B. find the probability that it is white.
8. A husband & wife appear in an interview for two vacancies in the same post. The probability of husband selection is $\frac{1}{7}$& that of wife's selection is $\frac{1}{5}$. What is the probability that,a). both of them will be

selected.b). only one of them will be selected.c). none of them will be selected?

9. A problem statistics is given to 3 students A,B & C whose chances of solving if are $\frac{1}{2}$, $\frac{3}{4}$&$\frac{1}{4}$ respectively. What is the probability that the problem will be solved?
10. A bag contains 8 white & 6 red balls. Find the probability of drawing 2 balls of the same color.
11. Find the probability of drawing an ace or a spade or both from a deck of cards?
12. A can hit a target 3 times in a 5 shots, B 2 times in 5 shots & C 3 times in a 4 shots. they fire a volley. What is the probability that a).2 shots hit? b). at least 2 shots hit?
13. A purse contains 2 silver & 4 cooper coins & a second purse contains 4 silver & 4 cooper coins. If a coin is selected at random from one of the two purses, what is the probability that it is a silver coin?
14. The contain of a three urns are : 1 white, 2 red, 3 green balls; 2 white, 1 red, 1 green balls & 4 white, 5 red, 3 green balls. Two balls are drawn from an urn chosen at random. This are found to be 1 white & 1 green. Find the probability that the balls so drawn come from the second urn.
15. Three machines A,B& C produced identical items. Of there respective output 2%, 4% & 5% of items are faulty. On a certain day A has produced 30% of the total output, B has produced 25% & C the remainder. An item selected at random is found to be faulty. What are the chances that it was produced by the machine with the highest output?
16. A person speaks truth 3 times out of 7. When a die is thrown, he says that the result is a 1. What is the probability that it is actually a 1?
17. There are three radio stations A, B and C which can be received in a city of 1000 families. The following information is available on the basis of a survey:

(a) 1200 families listen to radio station A

(b) 1100 families listen to radio station B.

(c) 800 families listen to radio station C.

(d) 865 families listen to radio station A & B.

(e) 450 families listen to radio station A & C.

(f) 400 families listen to radio station B & C.

(g) 100 families listen to radio station A,B & C.

The probability that a family selected at random listens at least to one radio station.

18. The probability distribution of a random variable x is as follows.

X	1	3	5	7	9
P(x)	K	2k	3k	3k	K

Find value of (i). K (ii). E(x)

19. A player tossed 3 coins. He wins Rs. 200 if all 3 coins show tail, Rs. 100 if 2 coins show tail, Rs. 50 if one tail appears and loses Rs. 40 if no tail appears. Find his mathematical expectation.
20. The probability distribution of daily demand of cell phones in a mobile gallery is given below. Find the expected mean .

Demand	5	10	15	20
Probability	0.4	0.22	0.28	0.10

21. If $P(A) = \frac{4}{15}$, $P(B) = \frac{7}{15}$ and if A and B are independent events, find $(i)P(A \cap B)$, $(ii)P(A \cup B)$, $(iii)P(\bar{A} \cap \bar{B})$.
22. If $P(A) = \frac{5}{9}$, $P(\bar{B}) = \frac{2}{9}$ and if A and B are independent events, find $(i)P(A \cap B)$, $(ii)P(A \cup B)$, $(iii)P(\bar{A} \cap \bar{B})$.
23. If $P(A) = 0.65$, $P(B) = 0.75$ and $P(A \cap B) = 0.45$, where A and B are events of sample space S, find $(i)P(A|B)$, $(ii)P(A \cup B)$, $(iii)P(\bar{A} \cap \bar{B})$.
24. A box containing 5 red and 3 black balls, 3 balls are drawn at random from box. Find the expected number of red balls drawn.
25. Two fair dice are rolled. X denotes the sum of the numbers appearing on the uppermost faces of the dice. Find the expected value.
26. A bag contains 5 black marbles and 6 white marbles. Find the number of ways that five marbles can be drawn from the bag such that it contains i) No restriction ii) no black marbles, iii) 3 black and 2 white, iv) at least 4 black, v) All are of same colors.
27. A student is to answer 8 out of 10 questions on an exam. Find the number of ways that the student can chose the 8 questions if i) No restriction, ii) student must answer the first 4 questions, iii) student must answer atleast 4 out of the five questions.
28. There are 12 points in a given plane, no three on the same line. i) How many triangle are determine by the points? ii) How many of these triangle contain a particular point as a vertex?
29. How many committees of two or more can be selected from 8 people?
30. Find the number of combinations if the letters of the letters of the word EXAMINATION taken out at a time.

5.12 LIST OF REFERENCES

- Statistics byMurry R. Spiegel, Larry J. Stephens. Publication McGRAWHILL INTERNATIONAL.
- Fundamental Mathematics and Statistics by S.C. Gupta and V.K kapoor Mathematical Statistics by J.N. Kapur and H.C. Saxena.

*****

6

ELEMENTARY SAMPLING THEORY

Unit Structure

6.1 OBJECTIVE

After going through this chapter you will able to know:

- Sampling and its requirements in statistics.
- Random sampling with and without replacement.
- Sampling distribution of Mean, Proportions, difference and sum.
- Standard errors in sampling distribution.
- Some software to use for sampling.

6.2 INTRODUCTION

In the previous chapters, we have discussed probability theory. In this chapter, we will introduce some basic concepts in statistics. The basic idea of statistical inference is to assume that the observed data is generated from some unknown probability distribution, which is often assumed to have a known functional form up to some unknown parameters. The purpose of statistical inference is to develop theory and methods to make inference on the unknown parameters based on observed data.

Sampling theory provides the tools and techniques for data collection keeping in mind the objectives to be fulfilled and nature of population. Sample surveys collect information on a fraction of total population whereas in census, the information is collected on the whole population.

The concept of sampling has a huge implementation and its application is seen in many vital fields. The importance of sampling theory is when it comes into play while making statistical analysis with different efficiency levels, there are three different methods of sampling. We have adequately thrown light on the process and methods of sampling.

6.3 SAMPLING THEORY

Often we are interested in drawing some valid inferences about a large group of individuals or objects called population in statistics. Instead of studying the entire population, which may be difficult or even impossible to study, we may study only a small portion of the population. Our objective is to draw valid inferences about certain facts for the population from results found in the sample; a process known as statistical inferences. The process of obtaining samples is called sampling and theory concerning the sampling is called sampling theory.

The sampling theory definition of the statistic is the creation of a sample set. This is recognized as one of the major processes. It retains the accuracy in bringing out the correct statistical information. The population tree is huge set and it turns out to be exhausting for the actual study and estimation process. Both money and time get exhausting in the process. The creation of the sample set saves time and effort and is a vital theory in the process of statistical data analysis.

Process of Sampling:
In this part of the chapter, we will discuss a few details regarding the process of sampling. So the steps are mentioned in the steps below:

- The first step is a wise choice of the population set.
- The second step is focusing on the sample set and the size of it.
- Then, one needs to choose an identifiable property based on which the samples will be created out of the population set.
- Then, the samples can be chosen using any of the types of sampling theory – Simple random, systematic, or stratified. Each of them is thoroughly discussed in the article ahead.
- Checking the inaccuracy, if there is any.
- Hence, the set is achieved in the result.

Sampling can be done in their different method and they are given below:

1. Simple random type.
2. Systematic Sampling.
3. Stratified sampling.

6.3.1 Random Samples and Random Numbers:

Definition: Simple random sampling is defined as a sampling technique where every item in the population has an even chance and likelihood of

being selected in the sample. Here the selection of items entirely depends on luck or probability, and therefore this sampling technique is also sometimes known as a method of chances. For e.g. Using the lottery method is one of the oldest ways and is a mechanical example of random sampling. In this method, the researcher gives each member of the population a number. Researchers draw numbers from the box randomly to choose samples. The use of random numbers is an alternative method that also involves numbering the population. The use of a number table similar to the one below can help with this sampling technique.

Simple random sampling (SRS) is a method of selection of a sample comprising of *n* a number of sampling units out of the population having *N* number of sampling units such that every sampling unit has an equal chance of being chosen.

Simple random sampling methods:

Researchers follow these methods to select a simple random sample:

1. They prepare a list of all the population members initially, and then each member is marked with a specific number (for example, there are nth members, then they will be numbered from 1 to N).
2. From this population, researchers choose random samples using two ways: random number tables and random number generator software. Researchers prefer a random number generator software, as no human interference is necessary to generate samples.

Advantages of simple random sampling:

1. It is a fair method of sampling, and if applied appropriately, it helps to reduce any bias involved compared to any other sampling method involved.
2. Since it involves a large sample frame, it is usually easy to pick a smaller sample size from the existing larger population.
3. The person conducting the research doesn't need to have prior knowledge of the data he/ she is collecting. Once can ask a question to gather the researcher need not be subject expert.
4. This sampling method is a fundamental method of collecting the data. You don't need any technical knowledge. You only require essential listening and recording skills.
5. Since the population size is vast in this type of sampling method, there is no restriction on the sample size that the researcher needs to create. From a larger population, you can get a small sample quite quickly.
6. The data collected through this sampling method is well informed; more the sample better is the quality of the data.

Disadvantages:

1. Sampling is not feasible where knowledge about each element or unit or a statistical population is needed.

2. The sampling procedures must be correctly designed and followed otherwise, what we call as wild sample would crop up with mis-leading results.
3. Each type of sampling has got its own limitations.
4. There are numerous situations in which units, to be measured, are highly variable. Here a very large sample is required in order to yield enough cases for achieving statistically reliable information.
5. To know certain population characteristics like population growth rate, population density etc. census of population at regular intervals is more appropriate than studying by sampling.

6.3.2 Sampling With and Without Replacement:

Selection with Replacement (SWR): In this case, a unit is selected from a population with a known probability and the unit is returned to the population before the next selection is made (after recording its characteristic). Thus, in this method at each selection, the population size remains constant and the probability at each selection or draw remains the same. Under this sampling plan, a unit has chances of being selected more than once. For example a card is randomly drawn from a pack of cards and placed back in the pack, after noting its face value before the next card is drawn. Such a sampling method is known as sampling with replacement. There are N^npossible samples of size *n* from a population of *N* units in case of sampling with replacement.

Sampling without replacement (SWOR): In this selection procedure, if a unit from a population of size *N* selected, it is not returned to the population. Thus, for any subsequent selection, the population size is reduced by one. Obviously, at the time of the first selection, the population size is *N* and the probability of a unit being selected randomly is $1/N$; for the second unit to be randomly selected, the population size is $(N-1)$ and the probability of selection of any one of the remaining sampling unit is $1/(N-1)$, similarly at the third draw, the probability of selection is $1/(N-2)$ and so on.

6.4 SAMPLING DISTRIBUTIONS

Sampling distribution is a statistic that determines the probability of an event based on data from a small group within a large population. Its primary purpose is to establish representative results of small samples of a comparatively larger population. Since the population is too large to analyze, the smaller group is selected and repeatedly sampled, or analyzed. The gathered data, or statistic, is used to calculate the likely occurrence, or probability, of an event. Using a sampling distribution simplifies the process of making inferences, or conclusions, about large amounts of data.

The idea behind a sampling distribution is that when you have a large amount of data (gathered from a large group, the value of a statistic from random samples of a small group will inform you of that statistic's value for the entire group. Once the data is plotted on a graph, the values of any given statistic in random samples will make a normal distribution from which you can draw inferences.

Each random sample selected will have a different value assigned to the statistic being studied. For example, if you randomly sample data three times and determine the mean, or the average, of each sample, all three means are likely to be different and fall somewhere along the graph. That's variability. You do that many times, and eventually the data you plot should look like a bell curve. That process is a sampling distribution.

Factors that influence sampling distribution:

The sampling distribution's variability can be measured either by standard deviation, also called "standard error of the mean," or population variance, depending on the context and inferences you are trying to draw. They both are mathematical formulas that measure the spread of data points in relation to the mean.

There are three primary factors that influence the variability of a sampling distribution. They are:

- **The number observed in a population:** This variable is represented by "*N*." It is the measure of observed activity in a given group of data.
- **The number observed in the sample:** This variable is represented by "*n*." It is the measure of observed activity in a random sample of data that is part of the larger grouping.
- **The method of choosing the sample:** How the samples were chosen can account for variability in some cases.

Types of distributions:

There are three standard types of sampling distributions in statistics.

1. Sampling Distribution of Means.
2. Sampling Distribution of Proportions.
3. Sampling Distributions of Differences and Sums.

6.4.1 Sampling Distribution of Means:

The most common type of sampling distribution is of the mean. It focuses on calculating the mean of every sample group chosen from the population and plotting the data points. The graph shows a normal distribution where the center is the mean of the sampling distribution, which represents the mean of the entire population.

The mean of the sampling distribution of the mean is the mean of the population from which the scores were sampled. Therefore, if a population has a mean μ, then the mean of the sampling distribution of the mean is also μ. The symbol $\mu_{\bar{X}}$ is used to refer to the mean of the sampling

distribution of the mean. Therefore, the formula for the mean of the sampling distribution of the mean can be written as:

$$\mu_{\bar{X}} = \mu$$

The standard deviation of the sampling distribution of the mean is computed as follows:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$$

That is, the standard deviation of the sampling distribution of the mean is the population Standard deviation divided by $\sqrt{N}$, the sample size (the number of scores used to compute a mean). Thus, the larger the sample size, the smaller the Standard deviation of the sampling distribution of the mean.

For sampling is drawn without replacement,

The mean of the sampling distribution of means given by

$$\mu_{\bar{X}} = \mu$$

The standard deviation of the sampling distribution of means is given by

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}\sqrt{\frac{N-n}{N-1}}.$$

Example 1: A population consists of the five numbers 5, 6, 7, 12, and 15. Consider all possible samples of size 2 that can be drawn with replacement from this population. Find

a) the mean of the population,
b) the standard deviation of the population,
c) the mean of the sampling distribution of means,
d) the standard deviation of the sampling distribution of means.

Solution: Here $population\ N = 5, sample\ size\ n = 2.$

a) The mean of the population is given by

$$\mu = \frac{5 + 6 + 7 + 12 + 15}{5} = \frac{45}{5} = 9$$

b) The standard deviation of the population is given by

$$\sigma = \sqrt{\frac{(x - \bar{x})^2}{N}}$$

$$\sigma = \sqrt{\frac{(5-9)^2 + (6-9)^2 + (7-9)^2 + (12-9)^2 + (15-9)^2}{5}}$$

$$\sigma = \sqrt{\frac{16 + 9 + 4 + 9 + 36}{5}}$$

$$\sigma = \sqrt{\frac{74}{5}} = 3.85$$

c) The mean of the sampling distribution of means:

Hare $5(5) = 25$ samples of size 2, that can be drawn with replacement, i.e. Samples are

$$(5,5), (5,6), (5,7), (5,12), (5,15), (6,5), (6,6), (6,7), (6,12), (6,15),$$
$$(7,5), (7,6), (7,7), (7,12), (7,15),$$
$$(12,5), (12,6), (12,7), (12,12), (12,15), (15,5), (15,6), (15,7),$$
$$(15,12), (15,15)$$

Therefore, the corresponding sample means are

5, 5.5, 6, 8.5, 10, 5.5, 6, 6.5, 9,
10.5, 6, 6.5, 7, 9.5, 11,
8.5, 9, 9.5, 12, 13.5, 10,
10.5, 11, 13.5, 15.

The mean of the sampling distribution of means is given by

$$\mu_{\bar{X}} = \frac{Sum\ of\ sample\ means}{25} = \frac{225}{25} = 9.$$

i.e. $\mu_{\bar{X}} = \mu$

d) The standard deviation of the sampling distribution of means.

$$\sigma_{\bar{X}} = \sqrt{\frac{(5-9)^2 + (5.5-9)^2 + (6-9)^2 + \cdots \ldots \ldots \ldots + (11-9)^2 + (13.5-9)^2 + (15-9)^2}{25}}$$

$$= \sqrt{\frac{185}{25}} = \sqrt{7.4} = 2.72$$

Therefore, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} = \frac{3.85}{\sqrt{2}} = 2.72.$

Example 2: A population consists of the five numbers 7, 9, 10, 14, and 20. Consider all possible samples of size 2 that can be drawn without replacement from this population. Find

a) the mean of the population,

b) the standard deviation of the population,

c) the mean of the sampling distribution of means,

d) the standard deviation of the sampling distribution of means.

Solution : Here $population\ N = 5, sample\ size\ n = 2.$

a) The mean of the population is given by

$$\mu = \frac{7 + 9 + 10 + 14 + 20}{5} = \frac{60}{5} = 12$$

b) The standard deviation of the population is given by

$$\sigma = \sqrt{\frac{(x - \bar{x})^2}{N}}$$

$$\sigma = \sqrt{\frac{(7 - 12)^2 + (9 - 12)^2 + (10 - 12)^2 + (14 - 12)^2 + (20 - 12)^2}{5}}$$

$$\sigma = \sqrt{\frac{25 + 9 + 4 + 4 + 64}{5}}$$

$$\sigma = \sqrt{\frac{106}{5}} = 4.60$$

c) The mean of the sampling distribution of means:

Here ${}^5C_2$ = 10 samples of size 2, that can be drawn without replacement, i.e.(all sample are distinct selection). Samples are

(7,9), (7,10), (7,14), (7,20), (9,10), (9,14), (9,20),

(10,14), (10,20), (14,20)

Therefore, the corresponding sample means are

8, 8.5, 10.5, 13.5, 9.5, 11.5, 14.5,
12, 15, 17.

The mean of the sampling distribution of means is given by

$$\mu_{\bar{X}} = \frac{Sum\ of\ sample\ means}{10} = \frac{120}{10} = 12.$$

i.e.$\mu_{\bar{X}} = \mu$

d) The standard deviation of the sampling distribution of means.

$$\sigma_{\bar{X}}$$
$$= \sqrt{\frac{(8-12)^2 + (8.5-12)^2 + (10.5-12)^2 + \cdots \ldots \ldots \ldots + (12-12)^2 \\ +(15-12)^2 + (17-12)^2}{10}}$$
$$= \sqrt{\frac{79.5}{10}} = \sqrt{7.95} = 2.81$$

Therefore, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}\sqrt{\frac{N-n}{N-1}} = \frac{4.6}{\sqrt{2}}\sqrt{\frac{5-2}{5-1}} = 2.81.$

6.4.2 Sampling Distribution of Proportions:

This sampling distribution focuses on proportions in a population. Samples are selected and their proportions are calculated. The mean of the sample proportions from each group represent the proportion of the entire population.

Suppose random samples of size n are drawn from a population in which the proportion with a characteristic of interest is p.

The Sampling Distribution of Proportion measures the proportion of success, i.e. a chance of occurrence of certain events, by dividing the number of successes i.e. chances by the sample size 'n'. Thus, the sample proportion is defined as

$$p = \frac{x}{n}$$

Therefore the mean μ_p and standard deviation σ_p are given by

$$\mu_p = p,$$
$$\sigma_p = \sqrt{\frac{pq}{n}}$$

Where q is probability of non-occurrence of event, which is given by$q = 1 - p$.

The following formula is used when population is finite, and the sampling is made without the replacement:

$$\sigma_p = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{pq}{n}}$$

If n is large, and p is not too close to 0 or 1, the binomial distribution can be approximated by the normal distribution.Practically, the Normal approximation can be used when both $np \geq 10$.

Once we have the mean and standard deviation of the survey data, we can find out the probability of a sample proportion. Here, the Z score conversion formula will be used to find out the required probability, i.e.

$$Z = \frac{X-\mu}{\sigma}.$$

Example 3:A random sample of 100 students is taken from the population of all part-time students in the Maharashtra, for which the overall proportion of females is 70%. Find sample mean and sample standard deviation.

Solution: Here $n = 100, p = 70\% = \frac{70}{100} = 0.7$

$$\therefore q = 1 - p = 1 - 0.7 = 0.3$$

the mean μ_p is given by

$$\mu_p = p = 0.7$$

and standard deviation σ_p is given by

$$\sigma_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.7 \times 0.3}{100}} = \sqrt{\frac{0.21}{100}} = 0.0458.$$

Example 4: Suppose it is known that 47% of Indian own smart phone. If a random sample of 50 Indians were surveyed, what is the probability that the proportion of the sample who owned smart phone is between 50% and 54%?

Solution: Here we have, $n = 50, p = 47\% = 0.47,$

$$\therefore q = 1 - p = 1 - 0.47 = 0.53.$$

Now, we should check our conditions for the sampling distribution of the sample proportion.

$np = 50 \times 0.47 = 23.5 \geq 10,\ nq = 50 \times 0.53 = 26.5 \geq 10.$

Since both the condition satisfy,

∴ The sampling distribution that is approximately normal with mean and standard deviation,

$$\mu_P = \mu = 0.47,$$

$$\sigma = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.47 \times 0.53}{50}} = 0.07$$

Now, The probability that the proportion of the sample who owned smart phone is between 50% and 54% is given by

$$P(0.50 < Z < 0.54) = P\left(\frac{0.50 - 0.47}{0.07} < Z < \frac{0.54 - 0.47}{0.07}\right)$$

$$= P(0.429 < Z < 1)$$

$$= P(Z < 1) - P(Z < 0.429)$$

$$= 0.8413 - 0.6627$$

$$= 0.1786$$

∴ If the true proportion of Indians who own smart phone is 47%, then there would be a 17.86% chance that we would see a sample proportion between 50% and 54% when the sample size is 50.

6.4.3 Sampling Distributions of Differences and Sums:

Statistical analyses are very often concerned with the difference between means. A typical example is an experiment designed to compare the mean of a control group with the mean of an experimental group. Inferential statistics used in the analysis of this type of experiment depend on the sampling distribution of the difference between means.

The sampling distribution of the difference between means can be thought of as the distribution that would result if we repeated the following three steps over and over again:

1. Sample n_1 scores from Population 1 and n_2 scores from Population 2.
2. compute the means of the two samples M_1 and M_2.
3. compute the difference between means, M_1 - M_2. The distribution of the differences between means is the sampling distribution of the difference between means.

As you might expect, the mean of the sampling distribution of the difference between means is:

$$\mu_{M_1-M_2} = \mu_{M_1} - \mu_{M_2}$$

Which says that the mean of the distribution of differences between sample means is equal to the difference between population means.

From the variance sum law, we know that:

$$\sigma^2{}_{M_1-M_2} = \sigma^2{}_{M_1} + \sigma^2{}_{M_2}$$

We can write the formula for the standard deviation of the sampling distribution of the difference between means as

$$\therefore \sigma_{M_1-M_2} = \sqrt{\sigma^2{}_{M_1} + \sigma^2{}_{M_2}}$$

Similarly we say about the sampling distribution of the sum between means is given by:

$$\mu_{M_1+M_2} = \mu_{M_1} + \mu_{M_2}$$

$$\sigma_{M_1+M_2} = \sqrt{\sigma^2{}_{M_1} + \sigma^2{}_{M_2}}$$

Example 5: Let S_1 be a variable that stands for any of the elements of the population 4, 6, 8 and S_2 be a variable that stands for any of the elements of the population 3, 5. Compute a) μ_{S_1}, b) μ_{S_2}, c) $\mu_{S_1-S_2}$, d)σ_{S_1}, e) σ_{S_2}, and f) $\sigma_{S_1-S_2}$.

Solution:

a) Here S_1 has sample of population 4, 6, 8.

$$\mu_{S_1} = \frac{4+6+8}{3} = \frac{18}{3} = 6.$$

b) Here S_2 has sample of population 3, 5.

$$\mu_{S_2} = \frac{3+5}{2} = 4$$

c) The population consisting of the differences of any member of S_1 and any member of S_2 mean is given by

$$\mu_{S_1-S_2} = \frac{(4-3)+(6-3)+(8-3)+(4-5)+(6-5)+(8-5)}{6}$$
$$= \frac{12}{6} = 2.$$

Therefore, we can verify that $\mu_{S_1-S_2} = \mu_{S_1} - \mu_{S_2} = 6 - 4 = 2.$

d) $\sigma_{S_1} = \sqrt{\frac{(4-6)^2+(6-6)^2+(8-6)^2}{3}} = \sqrt{\frac{8}{3}} = \sqrt{2.67} = 1.63$

e) $\sigma_{S_2} = \sqrt{\frac{(3-4)^2+(5-4)^2}{2}} = \sqrt{\frac{2}{2}} = 1.$

f) The population consisting of the differences of any member of S_1 and any member of S_2Standard deviation is given by

$$\sigma_{S_1-S_2} = \sqrt{\frac{(1-2)^2+(3-2)^2+(5-2)^2+(-1-2)^2+(1-2)^2+(3-2)^2}{6}}$$

$$= \sqrt{\frac{22}{6}} = \sqrt{\frac{11}{3}} = 1.91$$

Therefore, we can verify that,

$$\therefore \sigma_{S_1-S_2} = \sqrt{\sigma^2{}_{S_1} + \sigma^2{}_{S_2}} = \sqrt{2.67+1} = \sqrt{3.67} = 1.91$$

Example 6: The battery life of smart phone manufacturer A have a mean lifetime of 1050 days with a standard deviation of 150 days, while those of manufacturer B have a mean lifetime of 800 days with a standard deviation of 120 days. If random samples of 100 batteries of each brand are tested, what is the probability that the brand A batteries will have a mean lifetime that is at least (a) 200 days and (b) 280 days more than the brand B batteries?

Solution: Let $\overline{X_A}$ and $\overline{X_B}$ denote the mean lifetimes of samples A and B, respectively. Then

$$\mu_{\overline{X_A}-\overline{X_B}} = \mu_{\overline{X_A}} - \mu_{\overline{X_B}} = 1050 - 800 = 250\ days.$$

$$\sigma_{\overline{X_A}-\overline{X_B}} = \sqrt{\frac{\sigma_{\overline{X_A}}{}^2}{N_A} + \frac{\sigma_{\overline{X_B}}{}^2}{N_B}} = \sqrt{\frac{(150)^2}{100} + \frac{(120)^2}{100}} = \sqrt{\frac{36900}{100}}$$
$$= \sqrt{369} = 19.21$$

Therefore, the standardized variable for the difference in means is

$$\boldsymbol{Z} = \frac{(\overline{X_A} - \overline{X_B}) - (\mu_{\overline{X_A}-\overline{X_B}})}{\sigma_{\overline{X_A}-\overline{X_B}}}$$

and is very closely normally distributed.

a) the probability that the brand A batteries will have a mean lifetime that is at least 1200 days,

$$P(Z > 200) = 0.5 + P\left(Z = \frac{200 - 250}{19.21}\right) = 0.5 + P(Z > -2.6)$$
$$= 0.5 + 0.4953 = 0.9953.$$

b) the probability batteries will have a mean lifetime that is at least 960 days more than the brand B batteries,

$$P(Z > 280) = 0.5 - P\left(Z = \frac{280 - 250}{19.21}\right) = 0.5 - P(Z = 1.56)$$
$$= 0.5 - 0.4406 = 0.0594.$$

6.5 STANDARD ERRORS

Another measure is standard error, which is the standard deviation of the sampling distribution of an estimator. The idea is that if we draw a number of repeated samples of fixed size n from a population having a mean μ and variance σ^2 , each simple mean, say $\bar{x}$, will have a different value. Here $\bar{x}$it isa random variable and hence it has a distribution. The standard deviation of $\bar{x}$is called standard error. It has been proved that the standard error $'\sigma_{\bar{x}}^{'}$of the mean $\bar{x}$based on a sample of size n is,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

From above formula, it is obvious that the larger the sample size, the smaller the standard error and vice-versa. The advantage of considering standard error instead of a standard deviation is that this measure is not influenced by the extreme values present in a population under consideration.

In reality neither we use σ to calculate the standard error of $\bar{x}$nor we take more than one sample. As a matter of fact, what we do is, that we select only one sample, find its standard deviation s and use the following formula to find out the standard error of $\bar{x}$ .i.e.

$$S.E.(\bar{x}) = \frac{s}{\sqrt{n}}$$

Standard error is commonly used in testing of hypothesis and interval estimation. Many distributions, which are originally not normally distributed, have been taken as normal by considering the distribution of mean $\bar{x}$ for a large n.

6.6 SUMMARY

In this unit we have learn:

- In sampling theory we have Random Samples, Sampling With and Without Replacement of sample.
- Sampling Distributions and its types.
- Standard Errors of sampling distribution.

6.7 EXERCISE

1. A population consists of the four numbers 8, 11, 12, and 19. Consider all possible samples of size 2 that can be drawn with replacement from this population. Find
   a) the mean of the population,
   b) the standard deviation of the population,
   c) the mean of the sampling distribution of means,
   d) the standard deviation of the sampling distribution of means.
2. A population consists of the seven numbers 3, 5, 7, 9, 11, 13, and 15. Consider all possible samples of size 2 that can be drawn without replacement from this population. Find
   a) the mean of the population,
   b) the standard deviation of the population,
   c) the mean of the sampling distribution of means,
   d) the standard deviation of the sampling distribution of means.
3. Let S_1 be a variable that stands for any of the elements of the population 5, 8, 12 and S_2 be a variable that stands for any of the elements of the population 2, 6. Compute a) μ_{S_1}, b) μ_{S_2}, c) $\mu_{S_1-S_2}$, d)σ_{S_1}, e) σ_{S_2}, and f) $\sigma_{S_1-S_2}$.
4. A certain type of electric light bulb has a mean lifetime of 1500 h and a standard deviation of 150 h. Three bulbs are connected so that when one burns out, another will go on. Assuming that the lifetimes are normally distributed, what is the probability that lighting will take place for (a) at least 5000 h and (b) at most 4200 h?
5. Two distances are measured as 27.3 centimeters (cm) and 15.6 cm with standard deviations (standard errors) of 0.16 cm and 0.08 cm,

respectively. Determine the mean and standard deviation of (a) the sum and (b) the difference of the distances.

6. A and B play a game of ‘‘heads and tails,’’ each tossing 50 coins. A will win the game if she tosses 5 or more heads than B; otherwise, B wins. Determine the odds against A winning any particular game.
7. The average income of men in a city is Rs. 20,000 with standard deviation Rs. 10,500 and the average income of women is Rs. 16,000 and standard deviation Rs. 8,000. There are 100 sample selected from population. Find the probability of income between Rs. 15,000 to Rs. 18000.
8. A sample of 300 items selected at random had 32 defective items. Find mean and standard deviation of sampling distribution of proportion.
9. Ball bearings of a given brand weigh 0.50 g with a standard deviation of 0.02 g. What is the probability that two lots of 1000 ball bearings each will differ in weight by more than 2 g?
10. Find the probability that in 120 tosses of a fair coin (a) less than 40% or more than 60% will be heads and (b) 5/8 or more will be heads.

6.8 LIST OF REFERENCES

- Statistics by Murry R. Spiegel, Larry J. Stephens. Publication McGRAWHILL INTERNATIONAL.
- Fundamental Mathematics and Statistics by S.C. Gupta and V.K kapoor
- Mathematical Statistics by J.N. Kapur and H.C. Saxena.

*****

UNIT III

7

STATISTICAL ESTIMATION THEORY

Unit Structure

7.0 OBJECTIVES

- To understand basic definitions related to point estimation.
- To find the best point estimators to represent population characteristics.
- In this chapter, students can learn the requirements of a good estimator.

- To find an appropriate confidence interval for the population parameters.

7.1 BASIC DEFINITIONS

7.1.1 Population:

- A collection of all well-defined objects under study is called population.
- *Example:* Suppose we want to study the economic conditions of primary teachers in Maharashtra, then the group of all primary teachers in the state of Maharashtra is a population.

7.1.2 Sample:

- A well defined finite subset of the population is called a sample.
- *Example*: Suppose we want to study the economic conditions of primary teachers in the state of Maharashtra, then the few primary teachers (set of few teachers) in the state of Maharashtra forms a sample.

7.1.3 Parameter:

- An unknown constant of a population that summarises or describes an aspect of the population (such as a mean or a standard deviation) is called parameter. Let $f(x, \theta)$ be the pdf of a random variable 'X' having an unknown constant θ.

7.1.4 Statistic:

- Any function of a sample value (observed value) is called a statistic. The sample statistic is constants but it differ from sample to sample.

7.1.5 Sampling distribution:

- The probability distribution of the sample statistic is called a sampling distribution.

7.1.6 Parameter space:

- The set of all admissible values of a parameter of the distribution is called parameter space. It is denoted by Θ.
- *Example*: $X \sim \text{Normal}(\mu, \sigma^2)$
  $\Theta = \{(\mu, \sigma^2) / -\infty<\mu<\infty, \sigma>0\}$

7.1.7 Estimator:

- Let $x_1, x_2, \ldots, x_n$ be a sample of size n taken from a distribution having pdf $f(x, \theta)$ where $\theta \in \Theta$ is an unknown parameter. A function $T = T(x_1, x_2, \ldots, x_n)$ which maps sample space (S) to parameter space Θ is called an estimator. In other words, If a statistic $T = T(x_1, x_2, \ldots, x_n)$ is used to estimate θ, and its value belongs to parameter space then it is said to be an estimator of θ.

7.1.8 Estimate:

- A particular value of an estimator corresponding to the given sample values is called an estimate of the population parameter.

In the theory of estimation, there are two parts, 1) Point Estimation, 2) Interval Estimation.

7.2 POINT ESTIMATION

- Let $x_1, x_2, \ldots, x_n$ be a sample of size n taken from a distribution having pdf (probability density function) *f (x, θ)* where $\theta \in \Theta$ is an unknown parameter. The method of using sample statistic 'T' to estimate the value of parameter *θ*, which is a point on real line **R**, is called "Point Estimation".
- Requirements of good and reliable estimators:
  1. Unbiasedness
  2. Consistency
  3. Efficiency
  4. Sufficiency

7.2.1 Unbiasedness:

An estimator T = T $(x_1, x_2, \ldots, x_n)$ is said to be an unbiased estimator of θ iff;

$$E(T) = \theta, \quad \forall\, \theta \in \Theta,$$

a parameteric function $\phi(\theta)$ is said to be estimable if there exists a statistic h (T) such that, $E\big(h(T)\big) = \phi(\theta) \quad \forall\, \theta \in \Theta$.

7.2.1.1 The bias of an estimator:

An estimator T = T $(x_1, x_2, \ldots, x_n)$ is said to be a biased estimator with a bias $\frac{b(\theta)}{n}$ if;

$$E(T) = \theta + \frac{b(\theta)}{n} \quad \forall\, \theta \in \Theta$$

Or

$$E(T - \theta) = \frac{b(\theta)}{n} \quad \forall\, \theta \in \Theta$$

- If $b(\theta) > 0$, then estimator T is called biased estimator with an upward (positive) bias $\frac{b(\theta)}{n}$.
- If $b(\theta) < 0$, then estimator T is called biased estimator with a downward (negative) bias $\frac{b(\theta)}{n}$.
- If $b(\theta) = 0$, then estimator T is called an unbiased estimator.

- *Example*: Let x_1, x_2, x_3 be independent observation from Poisson (λ), then show that, $T = \frac{x_1 + x_2 + x_3}{3}$ is an unbiased estimator of λ.

Since x_1, x_2, x_3 are i.i.d. Poisson (λ).

$$E[X_1] = E[X_2] = E[X_3] = \lambda$$

Consider;

$$E[T] = E\left[\frac{x_1 + x_2 + x_3}{3}\right] = \frac{1}{3} E[x_1 + x_2 + x_3] = \lambda$$

Hence, T is an unbiased estimator of λ.

7.1.1.1 MSE of an estimator:

- Mean square error (MSE) of an estimator T = T (x_1, x_2, …, x_n) is,

$$MSE = E(T - \theta)^2$$

7.1.1 Consistency:

- A sequence of the estimator $T_n = T(x_1, x_2, \dots, x_n)$ is said to be a consistent estimator for parameter θ if any given $\epsilon > 0$,

$$P[|T_n - \theta| < \epsilon] \to 1 \quad as \quad n \to \infty \quad \forall \epsilon > 0$$

- Difference between T_n & θ becomes smaller and smaller as n goes to a large number (infinity).
- A sequence of an estimator $T_n = T(x_1, x_2, \dots, x_n)$ is said to be a consistent estimator for parameter θ if;

$$E(T_n) = \theta \quad \forall \theta \in \Theta$$

$$V[T_n] \to 0 \quad as\ n \to \infty$$

- ***Example*:** Show that the sample mean is a consistent estimator of the population mean.

Let $x_1, x_2, \dots, x_n$ be a random sample of size n from a population with mean μ and variance σ^2.

First, we have to define the sample mean,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$E[x_i] = \mu \quad \& V[x_i] = \sigma^2 \qquad \forall i = 1, 2, \dots, n$$

$$E[\bar{x}] = E\left[\frac{\sum_{i=1}^{n} x_i}{n}\right]$$

$$= \frac{\sum E[x_i]}{n}$$

$$= \frac{n\mu}{n}$$

$$E\,[\bar{x}] = \mu$$

Consider,

$$V\,[\bar{x}] = V\left[\frac{\sum_{i=1}^{n} x_i}{n}\right]$$

$$= \frac{1}{n^2}\sum V\,(x_i\,)$$

$$= \frac{1}{n^2}\,n\sigma^2$$

$$V\,[\bar{x}] = \frac{\sigma^2}{n}$$

$\text{Lim}_{n\to\infty}\, E\,[\bar{x}] = \mu$

& $V\,[\bar{x}] = \frac{\sigma^2}{n} \to 0.$

Therefore, the sample mean is a consistent estimator of the population mean.

- ***Remark***: If T_n is a consistent estimator for θ, then $\phi\,(T_n)$ is a consistent estimator for $\phi\,(\theta)$, where ϕ is a continuous function.

7.1.2 Efficiency:

- An estimator T_1 is said to be more efficient than estimator T_2 of parameter θ if $V\,(T_1) < V\,(T_2)$. The relative efficiency of T_2 with respect to T_1 is defined as,

$$e = \frac{V\,[T_1]}{V\,[T_2]}.$$

- *Example*: Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from a Normal $(\mu,\ \sigma^2)$. Then select the most efficient estimator between $\bar{x}$ (Mean) & x_m (Median).

We know that,

$$E\,[\bar{x}] = \mu$$

$$V\,[\bar{x}] = \frac{\sigma^2}{n} \to 0 \quad as \quad n \to \infty$$

& $$E\,[x_m] = \mu$$

$$V\,[x_m] = \left(\frac{\pi}{2}\right).\left(\frac{\sigma^2}{n}\right).$$

Now we have to calculate efficiency as,

$$e = \frac{V\,[\bar{x}]}{V\,[x_m]} = \frac{2}{\pi} < 1.$$

Therefore, the sample mean is a more efficient estimator than the sample median.

7.1.3 Minimum Variance Unbiased Estimator (MVUE):

- An estimator T = T $(x_1, x_2, \dots, x_n)$ is said to be minimum variance unbiased estimator of parameter θ if,

1) $E\,[T] = \theta \quad \forall\, \theta \in \Theta$ and,

2) $V\,[T] < V\,[T']$; where, $T'$is any other unbiased estimator of θ.

7.1.4 Uniformly minimum variance unbiased estimator (UMVUE):

Let θ be the unknown parameter & Θ be the parameter space of θ. Let $U\,(\theta)$ be the set of a class of unbiased estimator of $T\,(\theta)$ such that,

$$E\,[T^2] < \infty \quad \forall\, \theta \in \Theta$$

i.e.,

$$U\,(\theta) = \{T:\ E\,[T] = \theta, \qquad E\,[T^2] < \infty \ \ \forall\, \theta \in \Theta\}$$

Then, $T_0 \in U\,(\theta)$ is UMVUE of θ if,

$$E\,[T_0 - \theta]^2 \le E\,[T - \theta]^2 \qquad \forall\, \theta \in \Theta \ \&\, T \in U\,(\theta).$$

7.1.5 Likelihood function:

- Let $x_1, x_2, \dots, x_n$ be a sample of size n taken from a distribution having pdf *f (x, θ)* where $\theta \in \Theta$ is an unknown parameter. Then the likelihood function of $x_1, x_2, \dots, x_n$ is defined as

$$L\,(X_1, X_2, \dots, X_n, \theta) = f(X_1, \theta)\,.\,f(X_2, \theta)\,.\,f(X_3, \theta) \dots f(X_n, \theta)$$

$$L\,(\underline{X}, \theta) = \prod_{i=1}^{n} f(x_i, \theta).$$

- T is said to be the maximum likelihood estimator of θ, which maximizes the likelihood function $L\,(\underline{X}, \theta)$.

7.1.6 Sufficiency:

- **Definition I** A statistic $T = T(x_1, x_2, \dots, x_n)$ based on a sample of size n having pmf/pdf $f(X, \theta)$ $\theta \in \Theta$ is said to sufficient statistic for θ if and only if the information contains in T about θ is same as the information contains in $x_1, x_2, \dots, x_n$ about θ.
- **Definition II** A statistic $T = T(x_1, x_2, \dots, x_n)$ based on a sample of size n having pmf/pdf $f(X, \theta)$ $\theta \in \Theta$ is said sufficient statistic for θ, if and only if the conditional distribution of $x_1, x_2, \dots, x_n$ given T is independent of θ.
- **Definition III (Neyman factorization criterion)** If $x_1, x_2, \dots, x_n$ is a sample of size n having pmf/pdf $f(X, \theta)$ $\theta \in \Theta$ and $T = T(x_1, x_2, \dots, x_n)$ be a statistic which is said to be sufficient for θ if and only if the joint probability distribution function of $x_1, x_2, \dots, x_n$ can be expressed as a product of a function of T and θ, and function of $x_1, x_2, \dots, x_n$ only.

i.e., $L\,(\underline{X}, \theta) = g(T, \theta)\,.\,h(\underline{X})$,

then T is said to be sufficient statistic for θ, where $g(T,\theta)$ is a function of T & θ only and $h(\underline{X})$ is a function of $x_1, x_2, \dots, x_n$ only.

- ***Example*:** Let $x_1, x_2, \dots, x_n$ be a random sample from a population having pdf,

$$f(x) = \begin{cases} \theta x^{\theta-1}, & 0 \le x \le 1 \\ 0, & otherwise \end{cases}$$

The likelihood function of $x_1, x_2, \dots, x_n$ is,

$$L(\underline{X},\theta) = \prod_{i-1}^{n} \theta\, x_i^{\theta-1}$$

$$= \theta^n \prod_{i-1}^{n} x_i^{\theta-1}$$

$$= \left[\theta^n \prod x_i^{\theta}\right]\left[\frac{1}{\prod x_i}\right]$$

$$L(\underline{X},\theta) = g(T,\theta)\,.\,h(\underline{X})$$

Where, $(T,\theta) = \theta^n \prod x_i^{\theta}$, $h(\underline{X}) = \frac{1}{\prod x_i}$

Therefore, from the Neyman factorization criterion $\prod x_i$ is sufficient statistic for θ.

7.2 INTERVAL ESTIMATION

- A confidence interval for an unknown parameter is an interval of possible values for the parameter. It is constructed so that, with a chosen degree of confidence, the actual value of the parameter lies within the lower and upper bounds of the interval.

Let T_1 and T_2 be two statistics such that,

$$P(T_1 > \theta) = \alpha_1 \quad \dots (1)$$

and $$P(T_2 < \theta) = \alpha_2 \quad \dots (2)$$

where α_1 and α_2 are constants independent of θ. Equations (1) & (2) can be combined to give

$$P(T_1 < \theta < T_2) = 1 - \alpha \quad \dots (3)$$

where $\alpha = \alpha_1 + \alpha_2$

- ***Example*:** If we take a large sample from a normal population with mean μ and standard deviation σ, then $100(1-\alpha)\%$ confidence interval for a population mean at $\alpha = 0.05\ is$

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

In general, confidence interval for μ can be constructed by using the following normal probability approach,

$$P\left(Z_{\alpha/2} \leq Z \leq Z_{1-\alpha/2}\right) = 1-\alpha,$$

Where $P(Z < Z_{\alpha/2}) = \alpha/2$.

In particular case, $\alpha = 0.05$ then $Z_{\alpha/2} = -Z_{1-\frac{\alpha}{2}} = -1.96$ implies,

$P\ (-1.96 \leq Z \leq 1.96) = 0.95$ … (From Normal Probability Tables)

$$\Longrightarrow P\left(-1.96 \leq \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

$$\Longrightarrow P\left(\bar{x} - 1.96\,\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\,\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Thus, $\bar{x} \pm 1.96\,\frac{\sigma}{\sqrt{n}}$ are 95% confidence limits for the unknown parameter μ, the population mean and interval $\left(\bar{x} - 1.96\,\frac{\sigma}{\sqrt{n}},\ \bar{x} + 1.96\,\frac{\sigma}{\sqrt{n}}\right)$ is called the 95% confidence interval.

- **Remark** For a small sample case with unknown variance, to construct a confidence interval for a population mean, use students-t distribution.

7.2.1 Probable error:

- Probable error defines the half-range of an interval about a mean for the distribution, such that half of the values from the distribution will lie within the interval and half outside.
- Thus for a symmetric distribution it is equivalent to half the interquartile range, or the median absolute deviation.
- For a normal distribution probable error is 0.6745σ

7.3 EXERCISE

1. What do you understand by point estimation?
2. Explain the terms statistic and its sampling distribution.
3. Define estimator and estimate.
4. Write a note on unbiasedness, consistency, the efficiency of an estimator. Also, define sufficient statistics.
5. Define MVUE and UMVUE.
6. Define Neyman factorization criteria.
7. Construct a confidence interval for a population mean where population variance is known.

7.4 SUMMARY

- In this chapter we studied basic definitions related to point estimation of population parameters.
- The good estimator must satisfy conditions of Unbiasedness, Consistency, Efficiency and Sufficiency.
- Uniformly Minimum variance unbiased estimators popularly used techniques to estimate population parameters.
- In this chapter we studied confidence interval estimation of the population parameter.

7.5 REFERENCES

1. Gupta S. C. and Kapoor V. K., 2011, Fundamentals of Mathematical Statistics, 11th Ed, Sultan and Chand.
2. Rohatgi V. K, 1939, Introduction to Probability and Statistics, Wiley
3. Murray R. Spiegel, Larry J. Stephens, STATISTICS, 4th Ed, McGRAW – HILL INTERNATIONAL.
4. J.N. KAPUR and H.C. SAXENA, 2005, MATHEMATICAL STATISTICS, 12th rev, S.Chand.
5. Agrawal B. L, 2003, Programmed Statistics, 2nd Ed, New Age International.

*****

8

STATISTICAL DECISION THEORY

Unit Structure

8.6 Exercise

8.7 References

8.0 OBJECTIVES

In this chapter, students can learn,

- Basic definitions related to the testing of hypothesis/decision making
- Decision making through Testing of hypothesis
- Large and small sample tests
- Decision Making through statistical control chart

8.1 BASIC DEFINITIONS

8.1.1 Statistical decision:

- The decisions are made based on observations of a phenomenon that carry out probabilistic laws that are not completely known.

8.1.2 Hypothesis:

- A definite statement about the population parameter is called a hypothesis. A hypothesis is a claim to be tested.
- *Example*: A particular scooter gives an average of 50 km per litre.

8.1.3 Null hypothesis:

- A hypothesis having no difference is called the null hypothesis.
- ***Example*:** The population mean is μ_0 the hypothesis will be $H_0: \mu = \mu_0$.

8.1.4 Alternative hypothesis:

- A hypothesis that is accepted in the case H_0 is rejected is called the alternative hypothesis and usually denoted by H_1. It is exactly opposite to H_0.
- *Example*: If $H_0: \mu = \mu_0$ i.e., the population has a specified mean μ_0, then the alternative hypothesis could be;

i. $H_1: \mu \neq \mu_0 \quad (\mu > \mu_0 \; or \; \mu < \mu_0)$ … Two-tailed alternative

ii. $H_1: \mu > \mu_0$ … Right tailed alternative

iii. $H_1: \mu < \mu_0$ … Left tailed alternative

8.1.5 Simple and composite hypothesis:

- A statistical hypothesis that completely specifies the population parameter is called a simple hypothesis, and the hypothesis that does not specify the population parameter is called a composite hypothesis.
- *Example*: If $x_1, x_2, \ldots, x_n$ is a random sample from normal with mean μ and variance σ^2 then $H_0: \mu = \mu_0$ and $\sigma^2 = \sigma_1^2$ is a simple

hypothesis. The following fully not specified hypotheses is called a composite hypothesis.

1) $H_1: \mu \neq \mu_0$ 2) $H_0: \sigma^2 \neq \sigma_0^2$ 3) $H_0: \mu = \mu_0 \; and \; \sigma^2 > \sigma_0^2$ etc.

8.1.6 Errors in the test of significance:

- The main objective in the sampling theory is to draw a valid inference about the population parameters based on sample results. In practice, we decide to accept or reject the lot after examining a sample drawn from it. In sampling theory, we are liable to commit two types of errors: a rejection of a good lot and acceptance of a bad lot.

i. ***Type-I error***: Rejecting H_0 when H_0 is true.

ii. ***Type-II error***: Accepting H_0 when it is false (Accepting H_0 when H_1 is true).

iii. ***Size of Type-I and Type-II errors***

$$P\,[Reject\ H_0\ when\ it\ is\ true] = P\,[Reject\ H_0 \mid H_0] =$$
$$P\,[Reject\ a\ lot\ when\ it\ is\ good] = \alpha$$
$$P\,[Accept\ H_0\ when\ it\ is\ wrong] = P\,[Accept\ H_0 \mid H_1] =$$
$$P\,[Accept\ a\ lot\ when\ it\ is\ bad] = \beta$$

In the above probabilities, α & β are called the Type-I & Type-II errors, respectively.

- The four types of decisions are shown in the table as follows.

Actual Situation	Decision	
	Reject H_0	Accept H_0
H_0 is true	Type-I Error	Correct Decision
H_0 is false	Correct Decision	Type-II Error

8.1.7 Critical region:

- A region in sample space **S** which amounts to a rejection of H_0 is called a critical region or rejection region of H_0.

8.1.8 Level of significance:

- The probability 'α' is that the value of the test statistic belongs to the critical region, known as 'level of significance'. That is the probability of the occurrence of the type I error is the level of significance. Usually, we use the level of significance of 5% or 1%. The level of significance is always fixed in advance before collecting the sample information.

8.1.9 Test statistic:

- A function of sample observations is used to test H_0 is called test statistic.

8.1.10 One-tailed and two-tailed tests:

- A function of any statistical hypothesis where the alternative hypothesis is one-tailed (right-tailed or left-tailed) is called a one-tailed test.
- *Example*: A test for testing the mean of a population

  $H_0: \mu = \mu_0$ Versus $H_1: \mu > \mu_0$ (Right Tailed) or $H_1: \mu < \mu_0$ (Left Tailed), is a one-tailed test. In the right-tailed test, the critical region lies entirely in the right tail of the sampling distribution of $\bar{X}$, while for the left tailed test, the critical region lies entirely in the left tail of the sampling distribution of $\bar{X}$.
- A test of any statistical hypothesis where the alternative hypothesis is two-tailed such as; $H_0: \mu = \mu_0$ Versus $H_1: \mu \neq \mu_0$ $(\mu > \mu_0 \; or \; \mu < \mu_0)$ is known as a two-tailed test, and in such a case, the critical region is given by the portion of the area lying in both the tails (sides) of the probability curve of the test statistic $\bar{X}$.

8.1.11 Central limit theorem:

- In many cases, the exact probability distribution of the test statistics *T* cannot be obtained. The difficulty is overcome using the normal approximation. The probability distribution of standardized *T* is assumed to be *N (0, 1)* as the sample size $n \to \infty$ (i.e., *n* is sufficiently large). The corresponding theorem in support of the normal approximation is known as the central limit theorem.

Case-I: Parent population is Normal:

Let the random sample drawn from $N(\mu, \sigma^2)$. By the definition of a random sample, variates values $x_1, x_2, \ldots, x_n$ of the sample are independent and identically distributed as $N(\mu, \sigma^2)$ then the sample mean $(\bar{X})$ is distributed normally with μ and variance σ^2/n i.e., $\bar{X} \sim N(\mu, \sigma^2/n)$. The result shows how the precision of a sample mean increases as the sample size increases and $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$, standard normal variate.

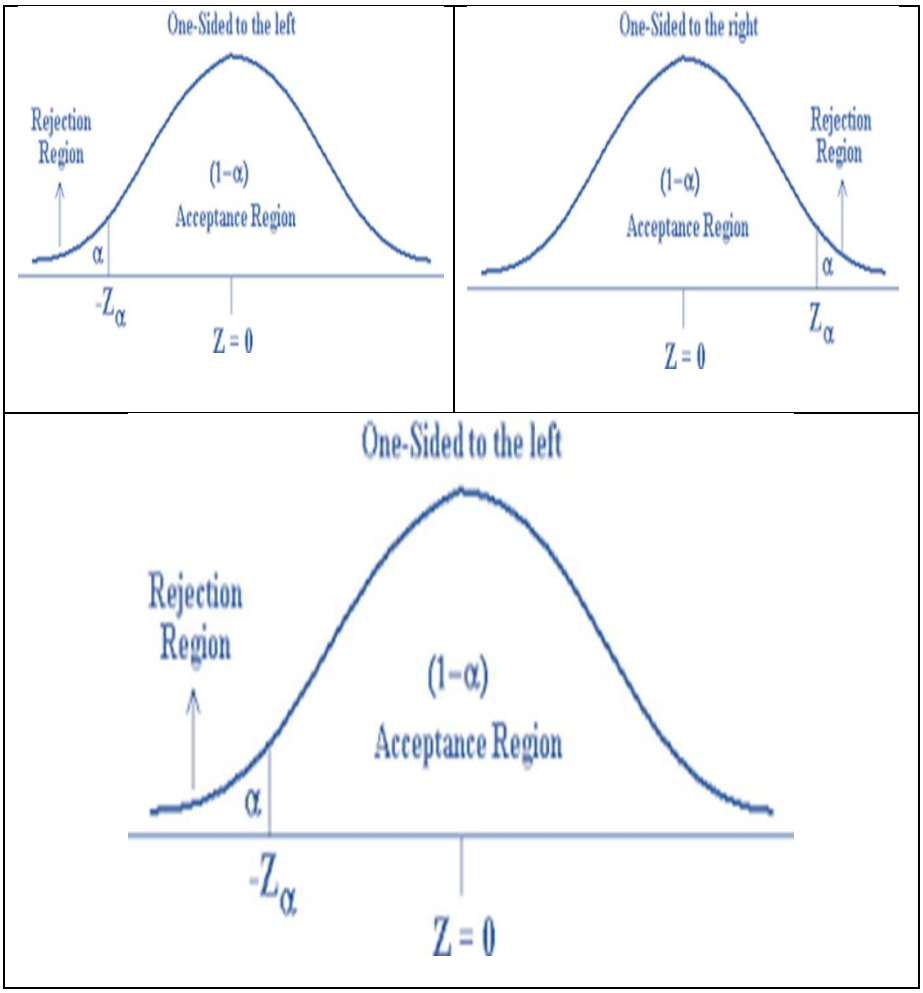
Case-II: Parent population is Non-Normal:

If the population from which the random sample is drawn has a non-normal distribution with finite mean μ and finite standard deviation σ then the variate, by owing to central limit theorem,

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \quad as \; n \to \infty$$

8.1.12 Critical value:

- The value of the test statistic, which separates the critical (rejection) region and acceptance region, is called the *'critical value'*. It depends upon (1) The level of significance α used and (2) The alternative hypothesis, whether it is two-tailed or one-tailed.



8.1.13 P-value:

- Another approach for testing is to find out the 'p' value at which H_0 is significant. That is, to find the smallest level of significance, α at which H_0 is rejected. About the acceptance or rejection of H_0, the experimenter can himself decide the level α by comparing it with the p-value. The criterion for this is that if the p-value is less than or equal to α, reject H_0 otherwise, accept H_0.
- ***Example*:**

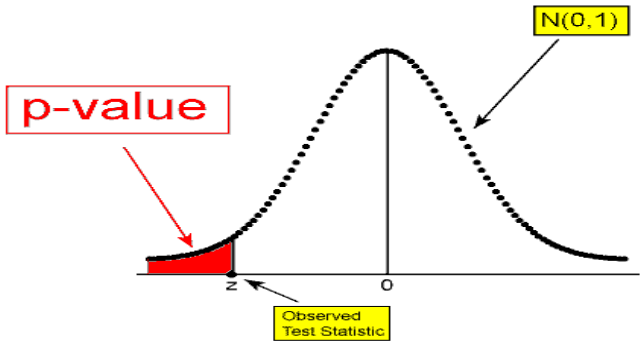


Fig.4: P-value for One-tailed Z-Test

8.1.14 Power of test:

- $$Power\ of\ Test = P\ [Rejecting\ H_0\ given\ that\ H_1\ is\ true]$$
$$= P\ [Rejecting\ H_0 \mid H_1\ is\ true]$$
$$= 1 - P\ [Accepting\ H_0 \mid H_1\ is\ true]$$
$$= 1 - P\ [Type\ II\ Error]$$
$$= 1 - \beta$$

8.1.15 The procedure of testing of hypothesis:

In any testing of a hypothesis, it is a stepwise procedure that leads to rejection or acceptance of the null hypothesis based on samples drawn from the population. The various steps in testing a statistical hypothesis are as follows;

i. Set up the null hypothesis H_0.

ii. Set up the alternative hypothesis H_1, this will enable us to decide whether we have to use a one-tailed (right or left) test or a two-tailed test.

iii. Choose an appropriate level of significance (α), i.e., α is fixed in advance.

iv. Choose the appropriate test statistic *Z* or *T* and find its value under the null hypothesis H_0.

v. Determine the critical values and critical region corresponding to the level of significance and the alternative hypothesis (Z_α for one-tailed H_1or $Z_{\alpha/2}$ for two-tailed H_1).

vi. Decision rule: We compare the calculated value of *Z*, with the tabulated value of Z_α if H_1 is one-tailed & compare |*Z*| with $Z_{\alpha/2}$ (For symmetrical distribution). If the calculated value is greater than the tabulated value, then we reject H_0 at α % level of significance and conclude that there is a significant difference at α % level of significance otherwise accept H_0 at α % significance level and conclude that there is no significant difference at α % level of significance.

8.2 LARGE SAMPLE TESTS

8.2.1 Test for testing of a population mean:

- Let consider $x = \{x_1, x_2, \dots, x_n\}$ be a random sample of size n taken from a normally distributed population having population mean μ and population variances σ^2 respectively.

- We have to test the hypothesis,

$$H_0: \mu = \mu_0$$

Against

$$H_1: \mu \neq \mu_0 \text{ OR}$$
$$H_1: \mu > \mu_0 \text{ OR}$$

$$H_1: \mu < \mu_0$$

- Under H_0, the test statistic is,

$$Z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1),$$

where n is the sample size, $\bar{x}$ is the sample mean, σ population standard deviation.

Let Z_α be the critical value at α level of significance. We compare the calculated value of *Z*, with the tabulated value of Z_α. Where $P(Z \leq Z_\alpha) = \alpha$.

- H_1 is two-tailed $(\mu \neq \mu_0)$

$|Z| \geq Z_{\alpha/2}$, then reject H_0.

- H_1 is one-tailed $(\mu > \mu_0)$

$Z \geq Z_{1-\alpha}$, then reject H_0.

- H_1 is one-tailed $(\mu < \mu_0)$

$Z \leq Z_\alpha$, then reject H_0.

8.2.2 Test for the testing difference of the two population mean:

- Let consider $x_1 = \{x_{11}, x_{12}, \ldots, x_{1n_1}\}$ and $x_2 = \{x_{21}, x_{22}, \ldots, x_{2n_2}\}$ be two independent normally distributed samples having population means μ_1 and μ_2 and population variances σ_1^2 and σ_2^2 respectively.
- $\bar{x}_1$ and $\bar{x}_2$ are the sample means.
- We have to test the hypothesis,

$H_0: \mu_1 - \mu_2 = \mu_0$

Against

$$H_1: \mu_1 \neq \mu_2 \text{ OR}$$
$$H_1: \mu_1 - \mu_2 > \mu_0 \text{ OR}$$
$$H_1: \mu_1 - \mu_2 < \mu_0$$

Under H_0, the test statistic is,

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)-\mu_0}{\sqrt{\left[\frac{\sigma_1^2}{n_1}\right]+\left[\frac{\sigma_2^2}{n_2}\right]}} \sim N(0,1).$$

Let Z_α be the critical value at α level of significance. We compare the calculated value of *Z*, with the tabulated value of Z_α. Where $P(Z \leq Z_\alpha) = \alpha$.

- H_1 is two-tailed $(\mu_1 \neq \mu_2)$

$|Z| \geq Z_{\alpha/2}$, then reject H_0.

- H_1 is one-tailed $(\mu_1 - \mu_2 > \mu_0)$

$Z \geq Z_{1-\alpha}$, then reject H_0.

- H_1 is one-tailed $(\mu_1 - \mu_2 < \mu_0)$

$Z \leq Z_{\alpha}$, then reject H_0.

8.2.4 Test for testing of the population proportion:

- We have to test the hypothesis,

$$H_0: P = P_0$$

Against

$$H_1: P \neq P_0 \text{ OR}$$
$$H_1: P > P_0 \text{ OR}$$
$$H_1: P < P_0$$

- Under H_0, the test statistic is,

$$Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}} \sim N(0,1),$$

where n is the sample size, *p* is the sample proportion.

Let Z_α be the critical value at α level of significance. We compare the calculated value of *Z*, with the tabulated value of Z_α. Where $P(Z \leq Z_\alpha) = \alpha$.

- H_1 is two-tailed $(P \neq P_0)$

$|Z| \geq Z_{\alpha/2}$, then reject H_0.

- H_1 is one-tailed $(P > P_0)$

$Z \geq Z_{1-\alpha}$, then reject H_0.

- H_1 is one-tailed $(P < P_0)$

$Z \leq Z_{\alpha}$, then reject H_0.

8.2.4 Test for testing equality of two population Proportion:

- We have to test the hypothesis,

$$H_0: P_1 = P_2$$

Against

$$H_1: P_1 \neq P_2 \text{ OR}$$
$$H_1: P_1 > P_2 \text{ OR}$$
$$H_1: P_1 < P_2$$

- Under H_0, the test statistic is,

$$Z = \frac{(p_1 - p_2)}{\sqrt{\hat{P}\hat{Q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \qquad \hat{Q} = 1 - \hat{P}$$

Let Z_α be the critical value at α level of significance. We compare the calculated value of *Z*, with the tabulated value of Z_α. Where $P(Z \leq Z_\alpha) = \alpha$.

- H_1 is two-tailed $(P_1 \neq P_2)$

$|Z| \geq Z_{\alpha/2}$, then reject H_0.

- H_1 is one-tailed $(P_1 > P_2)$

$Z \geq Z_{1-\alpha}$, then reject H_0.

- H_1 is one-tailed $(P_1 < P_2)$

$Z \leq Z_\alpha$, then reject H_0.

8.3 SMALL SAMPLE TESTS

8.3.1 Test for the population mean:

- Let consider $x = \{x_1, x_2, \ldots, x_n\}$ be a random sample of size n (n small) taken from a normally distributed population having population mean μ and population variances σ^2 (mean and variances are unknown) respectively.

- We have to test the hypothesis,

$$H_0: \mu = \mu_0$$

Against

$$H_1: \mu \neq \mu_0 \text{ OR}$$

$$H_1: \mu > \mu_0 \text{ OR}$$

$$H_1: \mu < \mu_0$$

- Under H_0, the test statistic is,

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1} d.f.$$

Where n is the sample size, $\bar{x}$ is the sample mean, s sample standard deviation.

Let $t_{(n-1,\alpha)}$ be the critical value based on students t-distribution at n-1 degrees of freedom and α level of significance. We compare the

calculated value of *t*, with the tabulated value of $t_{(n-1,\alpha)}$. Where $P(T \leq t_{(n-1,\alpha)}) = \alpha$.

- H_1 is two-tailed $(\mu \neq \mu_0)$

$|t| \geq t_{(n-1,\alpha/2)}$, then reject H_0.

- H_1 is one-tailed $(\mu > \mu_0)$

$t \geq t_{(n-1,1-\alpha)}$, then reject H_0.

- H_1 is one-tailed $(\mu < \mu_0)$

$t \leq t_{(n-1,\alpha)}$, then reject H_0.

8.3.2 Test for the difference of two population means (two samples):

- Let consider $x = \{x_1, x_2, \dots, x_{n_1}\}$ and $y = \{y_1, y_2, \dots, y_{n_2}\}$ be two independent normally distributed samples having unknown population means μ_1 and μ_2 and unknown population variances σ_1^2 and σ_2^2 respectively.
- $\bar{x}$ and $\bar{y}$ are the sample arithmetic means.
- We have to test the hypothesis,

$H_0: \mu_1 - \mu_2 = \mu_0$

Against

$H_1: \mu_1 \neq \mu_2$ OR

$H_1: \mu_1 - \mu_2 > \mu_0$ OR

$H_1: \mu_1 - \mu_2 < \mu_0$

- Under H_0, the test statistic is,

$$t = \frac{\bar{X} - \bar{Y}}{s/\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} d.f. \qquad \text{where } s^2 = \frac{(n_1-1)s_x^2 + (n_2-1)s_y^2}{n_1 + n_2 - 2}.$$

Let $t_{(n_1+n_2-2,\alpha)}$ be the critical value based on students t-distribution at $n_1 + n_2 - 2$ degrees of freedom and α level of significance. We compare the calculated value of *t*, with the tabulated value of $t_{(n_1+n_2-2,\alpha)}$. Where $P(T \leq t_{(n_1+n_2-2,\alpha)}) = \alpha$.

- H_1 is two-tailed $(\mu \neq \mu_0)$

$|t| \geq t_{(n_1+n_2-2,\alpha/2)}$, then reject H_0.

- H_1 is one-tailed $(\mu > \mu_0)$

$t \geq t_{(n_1+n_2-2,1-\alpha)}$, then reject H_0.

- H_1 is one-tailed $(\mu < \mu_0)$

$t \leq t_{(n_1+n_2-2,\alpha)}$, then reject H_0.

8.3.3 Paired t-test for difference of mean:

- Let consider $\{x_i, y_i\}; i = 1,2, \dots, n$ be n-pair dependent normally distributed samples having unknown population means μ_1 and μ_2 and unknown population variances σ_1^2 and σ_2^2 respectively.

- We have to test the hypothesis,

$$H_0: \mu_1 - \mu_2 = 0$$

Against

$$H_1: \mu_1 - \mu_2 \neq 0 \text{ OR}$$
$$H_1: \mu_1 - \mu_2 > 0 \text{ OR}$$
$$H_1: \mu_1 - \mu_2 < 0$$

- Under H_0, the test statistic is,

$t = \frac{\bar{X} - \bar{Y}}{s/\sqrt{n}} \sim t_{n-1} d.f.$ where $d = \bar{X} - \bar{Y}$ & $s^2 = \frac{1}{n-1}\sum(d_i - \bar{d})^2$

Let $t_{(n-1,\alpha)}$ be the critical value based on students t-distribution at n-1 degrees of freedom and α level of significance. We compare the calculated value of *t*, with the tabulated value of $t_{(n-1,\alpha)}$. Where $P(T \leq t_{(n-1,\alpha)}) = \alpha$.

- H_1 is two-tailed $(\mu \neq \mu_0)$

$|t| \geq t_{(n-1,\alpha/2)}$, then reject H_0.

- H_1 is one-tailed $(\mu > \mu_0)$

$t \geq t_{(n-1,1-\alpha)}$, then reject H_0.

- H_1 is one-tailed $(\mu < \mu_0)$

$t \leq t_{(n-1,\alpha)}$, then reject H_0.

8.4 STATISTICAL CONTROL CARTS

8.4.1 A lot acceptance sampling plan (LASP):

- A lot acceptance sampling plan (LASP) is a sampling scheme and a set of rules for making decisions. The decision, based on counting the number of defectives in a sample, can be to accept the lot, reject the lot, or even, for multiple or sequential sampling schemes, to take another sample and then repeat the decision process.

8.4.2 OC curve:

- Operating Characteristic (OC) Curve plots the probability of accepting the lot (Y-axis) versus the lot fraction or percent defectives (X-axis). The OC curve is the primary tool for displaying and investigating the properties of a LASP.

8.4.3 Control charts:

- Control charts are a statistical process control tool used to determine if a manufacturing or business process is in a state of control. It is more appropriate to say that the control charts are the graphical device for Statistical Process Monitoring (SPM). Traditional control charts are mostly designed to monitor process parameters when an underlying form of the process distributions are known.

8.4.4 Variable control chart ($\bar{X}$ Chart):

- Dr Walter A. Shewhart proposed a general model for control charts in 1920. Let w be a sample statistic that measures some continuously varying quality characteristic of interest (e.g., thickness), and suppose that the mean of w is μ_w, with a standard deviation of σ_w. Then the centre line, the Upper Control Limit (UCL), and the Lower Control Limit (LCL) are:

$$UCL = \mu_w + k\sigma_w$$

$$\text{Center Line } = \mu_w$$

$$LCL = \mu_w + k\sigma_w$$

where k is the distance of the control limits from the centre line, expressed in terms of standard deviation units. When k is set to 3, we speak of $3 - sigma$ control charts. Historically, $k = \mathbf{3}$ has become an accepted standard in the industry. The centerline is the process mean, which in general is unknown. We replace it with a target or the average of all the data. The quantity that we plot is the sample average, $\bar{X}$ The chart is called the $\bar{X}$ chart.

We also have to deal with the fact that σ is, in general, unknown. Here we replace σw with a given standard value, or we estimate it by a function of the average standard deviation.

8.4.5 Attributes control charts:

The Shewhart control chart plots quality characteristics that can be measured and expressed numerically. We measure weight, height, position, thickness, etc. If we cannot represent a particular quality characteristic numerically, or if it is impractical to do so, we then often resort to using a quality characteristic to sort or classify an item that is inspected into one of two "buckets".

An example of a common quality characteristic classification would be designating units as "conforming units" or "nonconforming units". Another quality characteristic criteria would be sorting units into "non-defective" and "defective" categories. Quality characteristics of that type are called **attributes.**

- Control charts dealing with the number of defects or nonconformities are called **c charts** (for the count).

- Control charts dealing with the proportion or fraction of defective products are called **p charts** (for proportion).
- There is another chart that handles defects per unit, called the **u chart** (for the unit). This applies when we wish to work with the average number of nonconformities per unit of product.

8.6 SUMMARY

Students can get an idea about the testing of a hypothesis and make decisions about parameters of interest. In this chapter, we briefly studied various large/small One-sample and two-sample tests for mean and proportion along with variable and attribute control charts are also discussed.

8.5 EXERCISE

1. Explain the term hypothesis and its types.
2. Define Type-I and Type-II errors.
3. Explain the terms level of significance, p-value and power of a test.
4. Write down the stepwise procedure of testing of hypothesis.
5. Write down the procedure for testing the equality of two population proportions.
6. Write down the procedure to test the specified population mean, in the case of a small sample.
7. Explain in detail paired t-Test for difference Mean.

8.7 REFERENCES

- Gupta S. C. and Kapoor V. K., 2011, Fundamentals of Mathematical Statistics, 11th Ed, Sultan and Chand.
- Rohatgi V. K, 1939, Introduction to Probability and Statistics, Wiley
- Murray R. Spiegel, Larry J. Stephens, STATISTICS, 4th Ed, McGRAW – HILL INTERNATIONAL.
- J.N. KAPUR and H.C. SAXENA, 2005, MATHEMATICAL STATISTICS, 12th rev, S.Chand.
- Agrawal B. L, 2003, Programmed Statistics, 2nd Ed, New Age International.
- Kanji G. K., 2006, 100 Statistical tests, 3rd Ed, SAGE Publication.

*****

9

STATISTICS IN R

Unit Structure

9.0 OBJECTIVES

- Use of R-software to find basic statistical measures
- Use of R-software to simulate Normal Distributions
- Use of R-software to simulate Binomial Distributions
- Data import and Export in R.

9.1 DESCRIPTIVE STATISTICS IN R-SOFTWARE

Sr. No.	Descriptive Statistics	Syntax in R	Output
1.	Observation Vector (c-function)	x <- c(1,4,7,12,19,15,21,20)	x [1] 1 4 7 12 19 15 21 20
2.	Arithmetic mean	AM<-mean(x)	AM [1] 12.375

3.	Mode	# Create the function.
		getmode <- function(x) {
		uniqv <- unique(x)
		
		uniqv[which.max(tabulate(match(x, uniqv)))]
		}
		v <- c(2,1,2,3,1,2,3,4,1,5,5,3,2,3);
		mode<-getmode(v)
			Mode [1] 2

4.	median	Med=median(x)	Med [1] 13.5

9.2 NORMAL DISTRIBUTION

R has four inbuilt functions to generate normal distribution. They are described below.

Sr. no.	Function Name	Syntax	Description
1.	Density function	dnorm(x, mean, sd)	This function gives the height of the probability distribution at each point for a given mean and standard deviation.
2.	Cumulative Probability	pnorm(x, mean, sd)	This function gives the probability of a normally distributed random number less than the value of a given number. It is also called the "Cumulative Distribution Function".
3.	Inverse function	qnorm(p, mean, sd)	This function takes the probability value and gives a number whose cumulative value matches the probability value.
4.	Random Number	rnorm(n, mean, sd)	This function is used to generate random numbers whose distribution is normal. It takes the sample size as input and generates that many random numbers.

Following is the description of the parameters used in the above functions:

x: is a vector of numbers.

p: is a vector of probabilities.

n: is a number of observations(sample size).

mean: is the mean value of the sample data. Its default value is zero.

sd: is the standard deviation. Its default value is 1.

9.3 BINOMIAL DISTRIBUTION

R has four in-built functions to generate binomial distribution. They are described below.

Sr. no.	Function Name	Syntax	Description
1.	Probability	dbinom(x, size, prob)	This function gives the

			probability density distribution at each point.
2.	Cumulative Probability	pbinom(x, size, prob)	This function gives the cumulative probability of an event. It is a single value representing the probability.
3.	Inverse function	qbinom(p, size, prob)	This function takes the probability value and gives a number whose cumulative value matches the probability value.
4.	Random Number	rbinom(n, size, prob)	This function generates required number of random values of given probability from a given sample.

Following is the description of the parameters used −

x: is a vector of numbers.

p: is a vector of probabilities.

n: is a number of observations.

Size: is the number of trials.

prob: is the probability of success of each trial.

9.4 FREQUENCY DISTRIBUTION IN R

- Table function in R -*table()*, performs categorical tabulation of data with the variable and its frequency. *table()* function is also helpful in creating Frequency tables with the condition and cross-tabulation
- Syntax

```
x <- c(1,2,3,2,4,2,5,4,6,7,8,9)

freq <- data.frame(table(x))
```

- Output

> freq	x	Freq
1	1	1
2	2	3
3	3	1
4	4	2
5	5	1
6	6	1
7	7	1
8	8	1
9	9	1

9.5 DATA IMPORT AND EXPORT

i. Data Importing:

The sample data is frequently observed in Excel format and needs to be imported into R before use. For this, we can use the function read.xls from the gdata package. It reads from an Excel spreadsheet and returns a data frame.

```
library(gdata)              # load gdata package
> help(read.xls)             # documentation
> mydata = read.xls("mydata.xls")  # read from the first sheet
```

ii. Data export:

There are numerous methods for exporting R objects into other formats.

- *A tab delimited tex file*

```
write.table(mydata, "c:/mydata.txt", sep="\t")
```

- *MS-Excel Spread sheet*

```
library(xlsx)
write.xlsx(mydata, "c:/mydata.xlsx")
```

9.7 SUMMARY

R is a programming language and free software environment for statistical computing and graphics supported by the R Core Team and the R Foundation for Statistical Computing. It is widely used for data analysis purposes in analytical industries. In this chapter, we studied various descriptive measures in R, probabilities, quantiles and random number generation syntax of the normal and binomial distribution.

9.6 EXERCISE

1. Write down the R-command for the arithmetic mean and compute for the given data.

   1, 4, 7, 12, 19, 15, 21 and 20.

2. Generate 10 random numbers using R-command from a normal distribution with mean zero and standard deviation one.
3. Calculate cumulative distribution function at point zero for normal distribution with mean zero and standard deviation one.
4. Compute probability density function at 1, 2, 3, 4, 5 for normal distribution with mean zero and standard deviation one.
5. Compute quantiles at 0.1, 0.2, 0.3, 0.4, 0.5 for binomial distribution with n=5 and p=0.7.
6. Generate 10 random numbers using R-command from a binomial distribution with n=5 and p=0.5.

7. Explain frequency distribution in R.

9.8 REFERENCES

- Gupta S. C. and Kapoor V. K., 2011, Fundamentals of Mathematical Statistics, 11th Ed, Sultan and Chand.
- R.B. Patil, H.J. Dand and R. Bhavsar, 2017, A Practical Approach using R, 1st ed, SPD.
- https://www.tutorialspoint.com/r/r_normal_distribution.htm
- https://cran.r-project.org/bin/windows/base/

*****

UNIT IV

10

SMALL SAMPLING THEORY

Unit Structure

10.0 OBJECTIVES

In this chapter we will study about the test suitable for small samples i.e sample size less than or equal to 30 for which the tests studied in previous chapters are not applicable. We will also study the test for equality of variance.

10.1 INTRODUCTION

The entire large sample theory was based on the application of "Normal Test". However if the sample size n is small, the distribution of the various statistics, e.g.$Z = \frac{\bar{x}-\mu}{(\frac{\sigma}{\sqrt{n}})} or Z = \frac{X-nP}{\sqrt{nPQ}}$etc., are far from normality and as such 'normal test' cannot be applied if 'n' is small. In such cases exact sample tests, pioneered by W.S. Gosset (1908) who wrote under the pen name of Student, and later on developed and extended by Prof. R. A. Fisher (1926),

are used. In the following sections we shall discuss i) t – test and ii) F-test.

The exact sample tests can, however, be applied to large samples though the converse is not true. In all exact sample tests, the basic assumption is that "the population(s) from which the sample(s) is(are) drawn is(are) normal, i.e., the parent population(s) is(are) normally distributed."

10.2 STUDENT'S T DISTRIBUTION

Let x_i(i = 1,2,….,n) be a random sample of size n from a normal population with mean μ and variance σ^2. Then Student's t is defined by the statistic:

$$t = t_{cal} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}, s^2 = \frac{1}{n-1}\sum_{i=1}^{i=n}(x_i - \bar{x})^2, \bar{x} = \frac{\sum x_i}{n}$$

where $\bar{x}$ is the sample mean and $s^2$is an unbiased estimate of the population variance σ^2, and it follows Student's distribution with $v = (n-1)$ degree of freedom with probability density function:

$$f(t) = \frac{1}{\sqrt{v}B\left(\frac{1}{2},\frac{v}{2}\right)} \cdot \frac{1}{\left(1+\frac{t^2}{v}\right)^{\frac{v+1}{2}}}, -\infty < t\infty$$

10.3 GRAPH OF *t-DISTRIBUTION*

The probability density function of t-distribution with n degrees of freedom is:

$$f(t) = C.\left(1+\frac{t^2}{n}\right)^{-\frac{(n+1)}{2}}, \quad -\infty < t < \infty$$

Since f(t) is an even function, the probability curve is symmetric about the line t =0. As t increases, f(t) decreases rapidly and tends to zero as $t \to \infty$, so that t-axis is an asymptote to the curve. We know that

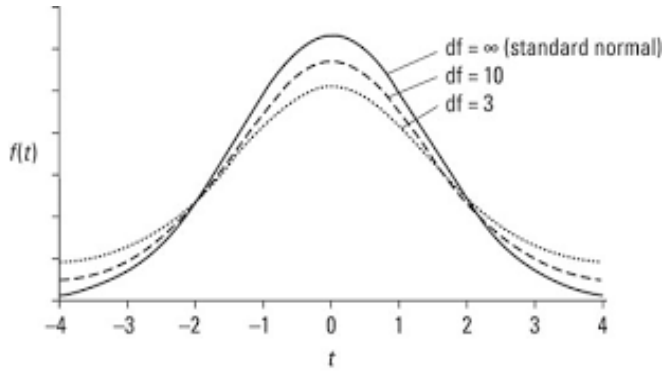
$$\mu_2 = \frac{n}{n-2}, n > 2; \ \beta_2 = \frac{3(n-2)}{n-4}, n > 4$$

Hence for n >2, $\mu_2 > 1$ i.e., the variance of t-distribution is greater than that of standard normal distribution and for n >4, $\beta_2 > 3$ and thus t-distribution is more flat on the top than the normal curve. In fact, for small n, we have

$$P(|t| \geq t_0) > -P(|Z| \geq t_0), \quad Z \sim N(0,1)$$

i.e, the tails of the t-distribution have a greater probability (area) than the tails of standard normal distribution. Moreover we can check that for large

n, t-distribution tends to standard normal distribution. Graph of t-distribution is given by the following diagram

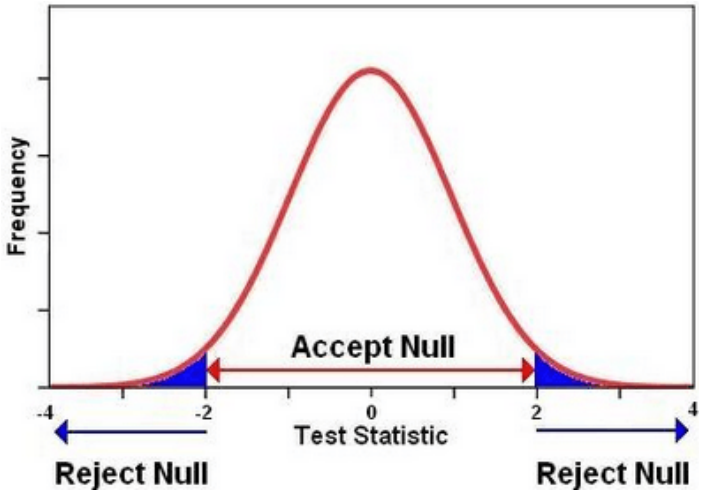


10.4 CRITICAL VALUES OF *t*

The critical values of t at level of significance α and degree of freedom v for two tailed test are given by the equation:

$$p\{|t| > t_v(\alpha)\} = \alpha$$

$$p\{|t| \leq t_v(\alpha)\} = 1 - \alpha$$



The values $t_v(\alpha)$ have been tabulated in table, for different values of $\alpha and v$ are given at the end of the chapter.

Since t-distribution is symmetric about t= 0, we get

$$P\big(t < t_v(\alpha)\big) + P\big(t < -t_v(\alpha)\big) = \alpha \Rightarrow 2P(\big(t > t_v(\alpha)\big) = \alpha$$

$$\Rightarrow P\big(t > t_v(\alpha)\big) = \frac{\alpha}{2} \therefore P\big(t > t_v(2\alpha)\big) = \alpha$$

$t_v(2\alpha)$ (from the tables at the end of the chapter) gives the significant value of t for a single tail test(Right tail or Left tail since the distribution is symmetrical), at level of significance $\alpha \& v$ degree of freedom.

Hence the significant values of t at level of significance $'\alpha'$ for a single tailed test can be obtained from those of two tailed test by looking the values at level of significance2α.For example

$t_8(0.05)\ for\ single\ tail\ test = t_8(0.1)\ for\ two\ tail\ test = 1.86$

$t_{15}(0.01)\ for\ single\ tail\ test = t_{15}(0.02)\ for\ two\ tail\ test = 2.602$

10.5 APPLICATION OF t-DISTRIBUTION

The t-distribution has a whole number of applications in Statistics, some of which are given here below:

i) to test if the sample mean $(\bar{x})$ differs significantly from the hypothetical value μ of the population mean.

ii) to test the significance of the difference between two sample means.

iii) to test the significance of an observed sample correlation coefficient and sample regression co efficient.

iv) to test the significance of observed partial correlation coefficient.

10.6 TEST OF HYPOTHESIS AND SIGNIFICANCE

Suppose we want to test:

i) if a random sample x_i(i=1,2,….,n) of size n has been drawn from a normal population with a specified mean say μ_0 orif the sample mean differs significantly from the hypothetical value of μ_0 of the population mean.
i.e $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0 or \mu > \mu_0 or \mu < \mu_0$where$H_0$ is the null hypothesis and H_1 is the alternative hypothesis

ii) Calculate

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}, S^2 = \frac{1}{n-1}\sum_{i=1}^{i=n}(x_i - \bar{x})^2$$

iii) degree of freedom df = v = n-1

iv) from the table calculate, $t_v(\alpha)$

v) Conclusion: Reject H_0if calculated |t| > tabulated t and Do not reject H_0 if calculated |t| ≤ tabulated t.

Remark: We know, the sample variance:

$$s^2 = \frac{1}{n}\sum(x_i - \bar{x})^2 \Rightarrow s^2 = \frac{1}{n}(n-1)S^2 \Rightarrow \frac{s^2}{n-1} = \frac{S^2}{n}$$

Hence for numerical problems, the test statistic t stated above will become

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}}$$

Eg: 1) A machinist is making engine parts with axle diameter of 0.7 inch. A random sample of 10 parts show a mean diameter of 0.742 inch with a standard deviation of 0.04 inch. Compute the statistic you would use to test whether the work is meeting the specifications and state the conclusion.

Solution:

i) $H_0 : \mu = 0.7, H_1 : \mu \neq 0.7$

ii) $\bar{x} = 0.742, s = 0.04, n = 10,$

$$t_{cal} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}} = \frac{0.742 - 0.7}{\frac{0.04}{\sqrt{9}}} = \frac{0.042 \times \sqrt{9}}{0.04} = 3.15$$

iii) degree of freedom i.e v = n-1, v = 9

iv) $t_v(\alpha) = t_9(0.05) = t_9(0.025) = 2.262$

v) conclusion$t_{cal} = 3.15 > t_9(0.05) = 2.262$
Since calculated 't' is greater than the tabulated 't' we reject the null hypothesis H_0 i.e the product is not conforming to specifications.

Eg: 2) The mean weakly sales of soap bars in departmental stores was 146.3 bars per store. After an advertising campaign the mean weekly sales in 22 stores for a typical week increased to 153.7 and showed a standard deviation of 17.2. Was the advertising campaign successful?

Solution:

i) $H_0 : \mu = 146.3, H_1 : \mu > 146.3$

ii) $\bar{x} = 153.7, s = 17.2, n = 22$

$$t_{cal} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}} = \frac{153.7 - 146.3}{\frac{17.2}{\sqrt{21}}} = \frac{7.4 \times \sqrt{21}}{17.2} = 1.97$$

iii) degree of freedom f i.e v = n-1, v = 21

iv) iv) $t_v(\alpha) = t_{21}(0.05) = t_{21}(0.025) = 1.72$

v) conclusion$t_{cal} = 1.97 > t_{21}(0.05) = 1.72$
Since calculated 't' is greater than the tabulated 't' we reject the null hypothesis H_0 i.e the advertising campaign was successful in promoting the sales.

Eg: 3) A random sample of 10 boys had the following I.Q.'s is 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Do these data support the assumption of a population mean I.Q of 100?

Solution: First we calculate find the value of $\bar{x}$&S^2.

Here n = 10, $\sum x = 972 \therefore \bar{x} = \frac{\sum x}{n} = \frac{972}{10} = 97.2$

x	70	120	110	101	88	83	95	98	107	100	Total
$x - \bar{x}$	-27.2	22.8	12.8	3.8	-9.2	-14.2	-2.2	0.8	9.8	2.8	
$(x - \bar{x})^2$	739.84	519.84	163.84	14.44	84.64	201.64	4.84	0.64	96.04	7.84	1833.6

$$S^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2 = \frac{1}{9}(1833.6) = 203.73 \Rightarrow S = \sqrt{203.73} = 14.273$$

i) $H_0 : \mu = 100, H_1 : \mu \neq 100$

ii) $\bar{x} = 97.2, S = 203.73, n = 10$

$$t_{cal} = \frac{|\bar{x} - \mu|}{\frac{S}{\sqrt{n}}} = \frac{|97.2 - 100|}{\frac{14.273}{\sqrt{10}}} = \frac{2.8 \times \sqrt{10}}{14.273} = 0.6203$$

iii) degree of freedom d f i.e v = n-1, v = 9

iv) $t_v(\alpha) = t_9(0.05) = t_9(0.025) = 2.262$

v) conclusion$t_{cal} = 0.6203 < t_9(0.05) = 2.262$

Since calculated 't' is less than the tabulated 't' we accept the null hypothesis H_0 i.e the data are consistent with the assumption of mean I.Q of 100 in the population.

10.7 CONFIDENCE INTERVAL

As done with normal distribution in earlier chapter, we can define 95%, 99 % or other confidence intervals by using the table given at the end of this chapter. We can estimate within specified limits of confidence the population mean μ. In general confidence limits are given by the formula

$$\bar{x} \pm \frac{t_v(\alpha)S}{\sqrt{n}}$$

In specific the 95 % confidence interval is given by

$$\left(\bar{x} - \frac{t_v(0.05)S}{\sqrt{n}}, \bar{x} + \frac{t_v(0.05)S}{\sqrt{n}}\right)$$

and 99% confidence interval is given by

$$\left(\bar{x} - \frac{t_v(0.01)S}{\sqrt{n}}, \bar{x} + \frac{t_v(0.01)S}{\sqrt{n}}\right)$$

Eg: 1) A random sample of 16 values from a normal population showed a mean of 41.5 inches and the sum of squares of deviations from this mean = 135 sq. inches. Obtain the 95 % and 99 % confidence limits for population mean.

Solution: n = 16, $\bar{x} = 41.5,\ \sum(x - \bar{x})^2 = 135$

$\therefore S^2 = \frac{1}{n-1}\sum(x - \bar{x})^2 = \frac{1}{15}(135) = 9 \Rightarrow S = 3$

from the table of t-distribution, we get $t_{15}(0.05) = 2.131 \& t_v(0.01) = 2.94795\%$ confidence limits for population mean are given by

$$\bar{x} \pm \frac{t_{15}(0.05)S}{\sqrt{n}} = 41.5 \pm 2.131 \times \frac{3}{\sqrt{16}} = 41.5 \pm 2.131 \times 0.75$$

$\Rightarrow 39.902 < \mu < 43.098$

99% confidence limits for population mean are given by

$$\bar{x} \pm \frac{t_{15}(0.01)S}{\sqrt{n}} = 41.5 \pm 2.947 \times \frac{3}{\sqrt{16}} = 41.5 \pm 2.947 \times 0.75$$

$\Rightarrow 39.29 < \mu < 43.71$

10.8 t-TEST FOR DIFFERENCE OF MEANS

Suppose we want to test if two independent samples $x_i(i = 1,2,3,\cdots\cdots,n_1)$ and $y_j(j = 1,2,3,\cdots\cdots,n_2)$ of size n_1 and n_2 have been drawn from two normal populations with means $\mu_x(\mu_1)$ & $\mu_y(\mu_2)$ respectively.

Under the null hypothesis (H_0) that the samples have been drawn from the normal populations with mean $\mu_x(\mu_1) \& \mu_y(\mu_2)$ and the under the assumption that the population variance are equal
i.e $\sigma_x^2 = \sigma_y^2 = \sigma^2$, the statistic

$$t = \frac{\left(\bar{x} - \bar{y} - (\mu_x - \mu_y)\right)}{S\left(\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)} = \frac{\bar{x} - \bar{y} - d_0}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},\ d_0 = \mu_x - \mu_y$$

Where $\bar{x} = \frac{\sum x_i}{n_1}, \bar{y} = \frac{\sum y_j}{n_2}$ & $S^2 = \frac{1}{n_1+n_2-2}\left[\sum(x_i - \bar{x})^2 + \sum(y_j - \bar{y})^2\right]$ is an unbiased estimate of the common population variance σ^2, follows

Students t-distribution with $v = (n_1 - 1) + (n_2 - 1) = (n_1 + n_2 - 2)$ degrees of freedom.

Paired t-test for Difference of means:

Let us now consider the case when (i) the sample sizes are equal i.e. $n_1 = n_2 = n$ and (ii) the two samples are not independent but the sample observations are paired together, i.e., the pair of observations $(x_i, y_i), 1 \leq i \leq n$ corresponds to the same i^{th} sample unit. The problem is to test if the sample means differ significantly or not.

For example, suppose we want to test the efficacy of a particular drug, say, for inducting sleep. Let x_i and y_i (i =1,2,……,n) be the readings, in hours of sleep, on the i^{th} individual, before and after the drug is given respectively. Here instead of applying the difference of the means test discussed above in the same section we apply the paired t-test given below:

Here we consider the increments, $d_i = x_i - y_i\ , 1 \leq i \leq n$

Under the null Hypothesis, H_0 that increments are due to fluctuations of sampling, i.e., the drug is not responsible for these increments, the statistic is

$$t = \frac{\bar{d}}{\frac{S}{\sqrt{n}}} = \frac{\bar{d} \times \sqrt{n}}{S}$$

where

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \;\&S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} d_i^2 - \frac{(\sum_{i=1}^{n} d)^2}{n}\right\}$$

follows Student's t-distribution with (n - 1) degree of freedom.

Eg: 1) For a random sample of 10 pigs fed on diet A, the increases in weight 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 the same period were: 7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 17.Test whether diets A and B differ significantly as regards to their effect on increase in weight.

Solution:

i) Null Hypothesis**,** $H_0: \mu_x = \mu_y$, i.e., there is no significant difference between the mean increase in weight due to diets A and B.

Alternative hypothesis, $H_1: \mu_x \neq \mu_y$(two tailed)

ii)

	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_j	$y_j - \bar{y}$	$(y_j - \bar{y})^2$
	10	-2	4	7	-8	64
	6	-6	36	13	-2	4
	16	4	16	22	7	49

	17	5	25	15	0	0
	13	1	1	12	-3	9
	12	0	0	14	-1	1
	8	-4	16	18	3	9
	14	2	4	8	-7	49
	15	3	9	21	6	36
	9	-3	9	23	8	64
				10	-5	25
				17	2	4
Total	120	0	120	180	0	314

$$\bar{x} = \frac{\sum x}{n_1} = \frac{120}{10} = 12 \ , \bar{y} = \frac{\sum y}{n_2} = \frac{180}{12} = 15$$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum (x_i - \bar{x})^2 + \left(y_j - \bar{y}\right)^2\right] = \frac{1}{20}[120 + 314]$$

$$S^2 = 21.7 \Rightarrow S = \sqrt{21.7} = 4.6583$$

$$t = \frac{\bar{x} - \bar{y} - d_0}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{12 - 15 - 0}{4.6583\sqrt{\frac{1}{10} + \frac{1}{12}}} = \frac{-3 \times \sqrt{120}}{4.6583 \times \sqrt{22}} = -\frac{32.8634}{21.8494}$$

$$\Rightarrow t = -1.5041$$

iii) df = v = 10 + 12 - 2 = 20

iv) $t_{20}(0.05) = 2.086$

v) Conclusion: $|t_{cal}| = 1.5041 < t_{20}(0.05) = 2.086$
Since calculated value of t is less than the tabulated value, the null hypothesis H_0 is accepted at 5% level of significance and we may conclude that the two diets do not differ significantly as regards their effect on increase in weight.

Eg: 2) The yields if two types 'Type A' and 'Type B' of grains in pounds per acre in 6 replications are given below. What comments would you make on the difference in the mean yield.

Replication	1	2	3	4	5	6
Yield of Type A	20.5	24.6	2306	29.98	30.37	23.83
Yield of Type B	24.86	26.39	28.19	30.75	29.98	22.04

Solution: i) $H_0: \mu_x = \mu_y, H_1: \mu_x \neq \mu_y$

ii) $t = \frac{\bar{d}}{\left(\frac{S}{\sqrt{n}}\right)}$, $\bar{d} = \sum_{i=1}^{n} d_i \, , d_i = x_i - y_i, v = n - 1$

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

replication	type A	type B	d	d^2
1	20.5	24.86	-4.36	
2	24.6	26.39	-1.79	
3	23.06	28.19	-5.13	
4	29.98	30.75	-0.77	
5	30.37	29.97	0.4	
6	23.83	22.04	1.79	
total			-9.86	52.4876

$$(iii) S^2 = \frac{1}{5}\left[52.4876 - \frac{(-9.86)^2}{6}\right] = \frac{1}{5}\left[52.4876 - \frac{97.2196}{6}\right]$$

$$= \frac{36.2843}{5} = 7.2569$$

$$\therefore S = \sqrt{7.2569} = 2.6939$$

$$\bar{d} = -\frac{9.86}{6} = -1.6433, t = \frac{\bar{d}}{\frac{S}{\sqrt{n}}} = \frac{-1.6433\times\sqrt{6}}{2.6939} = -1.4942$$

iv) $st_5(0.05) = 2.571$

v) Conclusion: $|t_{cal}| = 1.4942 < t_5(0.05) = 2.571$

Since the calculated value of t is less than the tabulated value of t we accept H_0 at 5% level of significance i.e., there is no major difference between the mean yield of two types diets.

10.9 DEGREES OF FREEDOM

In order to discuss the statistic t as discussed above, it is necessary to use observations obtained from a sample as well as certain population parameters. If these parameters are unknown, they must be estimated from the sample.

The number of degrees of freedom of a statistic, generally denoted by v, is defined as the number N of independent observation in the sample (i.e. the sample size) minus the number k of population parameters, which must be estimated from sample observations. In symbols,
$v = N - k.$

In case of the statistic t, the number of independent observations in the sample is N, from which we compute$\bar{x}\&S$. However, since we must estimate $\mu, k = 1$and $v = N - 1$.

10.10 THE *f*-DISTRIBUTION

If X and Y are two independent chi-square variates with v_1 and v_2 degree of freedom respectively, then F-statistic is defined by

$$F = \frac{X/v_1}{Y/v_2}$$

In other words, F is defined as the ration of two independent chi-square variates divided by their respective degrees of freedom and it follows Snedecor's F-distribution with (v_1, v_2) degree of freedom with probability function given by:

$$f(F) = \frac{\left(\frac{v_1}{v_2}\right)^{\frac{v_1}{2}}}{B\left(\frac{v_1}{2}, \frac{v_2}{2}\right)} \cdot \frac{F^{\frac{v_1}{2}-1}}{\left(1 + \frac{v_1}{v_2}F\right)^{\frac{(v_1+v_2)}{2}}}$$

Remark: The sampling distribution of F-statistic does not involve any population parameters and depends only on the degrees of freedom$v_1 \& v_2$.

F-test for Equality of Two population Variances.

Suppose we want to test (i) whether two independent samples x_i, (i=1,2,…n_1) and y_j, (j =1,2,…n_2) have been drawn from the normal population with the same variance σ^2 or (ii) whether the two independent estimates of the population variance are homogenous or not.

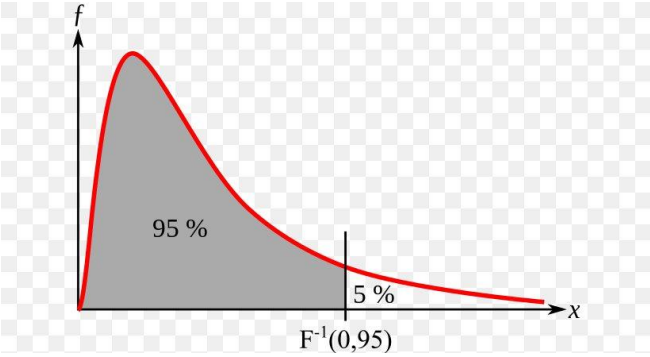
Under the Null hypothesis H_0: $\sigma_x^2 = \sigma_y^2 = \sigma^2$ i.e., the population variances are equal, or two independent estimates of the population variance are homogenous, the statistics F is given by

$$F = \frac{S_X^2}{S_Y^2}$$

where

$$S_X^2 = \frac{1}{n_1 - 1}\sum(x_i - \bar{x})^2, \quad S_Y^2 = \frac{1}{n_2 - 1}\sum(y_i - \bar{y})^2$$

are unbiased estimates of the common population variance σ^2 obtained from two independent samples and it follows Snedecor's F-distribution with $(v_1, v_2) = (n_1 - 1, n_2 - 1)$ degree of freedom.



The shaded in the diagram indicates the acceptance region $(1 - \alpha)$ and the unshaded region indicates the rejection region α.

Eg: 1) In one sample of 8 observations, the sum of the squares of deviations of the sample values from the sample mean was 84.4 and in the other sample of 10 observations it was 102.6. Test whether this difference is significant at 5% LOS given that 5% point of F for v_1 =7 and v_2 = 9 degree of freedom is 3.29.

Solution: $H_0: \sigma_x^2 = \sigma_y^2$ i.e., the estimate of variance given by the samples are homogenous,

$H_1 : \sigma_x^2 \neq \sigma_y^2$

$n_1 = 8, n_2 = 10, \sum(x_i - \bar{x})^2 = 84.4, \sum(y_i - \bar{y})^2 = 102.6$

$$S_x^2 = \frac{1}{n_1 - 1}\sum(x_i - \bar{x})^2 = \frac{1}{7} \times 84.4 = 12.0571$$

$$S_y^2 = \frac{1}{n_2 - 1}\sum(y_i - \bar{y})^2 = \frac{1}{9} \times 102.6 = 11.4$$

$$F = \frac{S_x^2}{S_y^2} = \frac{12.0571}{11.4} = 1.0576$$

tabulated $F_{7,9}(0.05) = 3.29$

$F_{cal} = F = 1.0576 < F_{0.05}(7,9) = 3.29$

$\therefore Accept\ H_0\ at\ 5\%\ LOS$

Eg : 2) Two random samples gave the following results:

Sample no	Size	sum of squares of deviation from the mean
1	10	90
2	12	108

Test at 5% LOS whether there is a difference in variance.
{Given $F_{0.05}(9, 11) = 2.9$, $F_{0.05}(11,9) = 3.1$}

Solution: $H_0: \sigma_x^2 = \sigma_y^2$, $H_1: \sigma_x^2 \neq \sigma_y^2$

$n_1 = 10, n_2 = 12, \sum(x_i - \bar{x})^2 = 90, \quad \sum(y_i - \bar{y})^2 = 108$

$$S_x^2 = \frac{1}{n_1 - 1}\sum(x_i - \bar{x})^2 = \frac{1}{9} \times 90 = 10$$

$$S_y^2 = \frac{1}{n_2 - 1}\sum(y_i - \bar{y})^2 = \frac{1}{11} \times 108 = 9.8182$$

$$F_{cal} = \frac{S_x^2}{S_y^2} = \frac{10}{9.8182} = 1.0185$$

$F_{cal} = 1.0185 < F_{0.05}(9,11) = 2.9$

$\therefore accept\ H_0$

10.11 SUMMARY

In this chapter we had learnt about to apply t-test for the sample size less than or equal to 30. We had also learnt to test if the sample mean ($\bar{x}$) differs significantly from the hypothetical value μ of the population mean and to test the significance of the difference between two sample means. WE had also learnt about the confidence interval and the F-test to whether the two independent estimates of the population variance are homogenous or not.

10.13 EXERCISES

1) A researcher is interested in determing whether or not review sessions affect exam performance. The independent variable, a review session, is administered to a sample of students (n=9) in an attempt to determine if this has an effect on the dependent variable, exam performance. Based on the information gathered in previous semesters, the researcher knows that the population mean for a given exam is 24. The sample mean is 25. with a S.D of 4, LOS = 5%.

2) You conduct a survey of a sample of 25 members of this year's graduating marketing students and find that average GPA is 3.2. The standard deviation of the sample is 4. Over the last year the average GPA has been 3.0. Is the GPA of this year's students significantly different from long run average?

3) The heights of 10 males of given locality are found to be 70, 67, 62, 68, 61, 68, 70, 64, 64, 66 inches. Is it reasonable to believe that the average height is greater than 64 inches?

4) A random sample of 16 values from a normal population showed a mean of 41.5 inches and the sum of squares of deviations from this mean = 135 sq. inches. Show that the assumptions of mean 43.5 inches for the population is not reasonable. Obtain the 95 % and 99 % confidence limits.

5) Below are given the gain in weights (in kgs) of pigs fed on two Diets A and B

Diet A	25	32	30	34	24	14	32	24	30	31	35	25			
Diet B	44	34	22	10	47	31	40	30	32	35	18	21	35	29	22

Test if the two diets differ significantly as regards to their effect on increase in weight.

6) Samples of two types of electric light bulbs were tested for length of life and following data were obtained:

	Type I	Type II

Sample Size	8	7
Sample mean	1234	1036
Sample S.D.	36	40

Is the difference in the means sufficient to warrant that Type I is superior to Type II regarding length of life?

7) Two laboratories carry out independent estimates of a particular chemicals in a medicine produced by a certain firm. A sample is taken from each batch, halved and the separate halves sent to the two laboratories. The following data is obtained.

no. of samples	10
mean value of the diff. of estimates	0.6
sum of the squares of their diff. from their mean	20

Is the difference significant at 5% LOS.

8) A certain stimulus administered to each of the following 12 patients resulted in the following increase of blood pressure: 5, 2, 8, -1, 0,-2, 1, 5, 0,4 and 6. Can it be obtained that the stimulus will, in general, be accompanied by an increase in blood pressure?

9) Two independent samples of 8 and 7 items respectively had the following values of the variables:

Sample I	9	11	13	11	15	9	12	14
Sample II	10	12	10	14	9	8	10	

Do the estimates of population variance differ significantly?

10.14 SOLUTION TO EXERCISES

Q. No.	Solution	Q. No.	Solution
1	$H_0: \mu = 24,\ H_1: \mu > 24,$ t= 0.75	2	$H_0: \mu = 3, H_1: \mu \neq 3, t = 0.2445$
3	t = 2, table t = 1.833, Reject H_0	4	$H_0: \mu = 43.5\ , H_1: \mu \neq 43.5,\ \lvert t_{cal}\rvert = 2.667, tablet = 2.131$
5	t = -0.609 , accept H_0	6	t = 9.3924, Reject H_0
7	t = 2.2222, Accept H_0	8	t= 2.89, Reject H_0
9	Cal. F = 0.8315		

10.15 TABLES OF *t*-DISTRIBUTION AND *f*-DISTRIBUTION *t* DISTRIBUTION : CRITICAL VALUES OF *t*

Significance level

Degrees of freedom	*Two-tailed test*: *One-tailed test*:	10% 5%	5% 2.5%	2% 1%	1% 0.5%	0.2% 0.1%	0.1% 0.05%
1		6.314	12.706	31.821	63.657	318.309	636.619
2		2.920	4.303	6.965	9.925	22.327	31.599
3		2.353	3.182	4.541	5.841	10.215	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
5		2.015	2.571	3.365	4.032	5.893	6.869
6		1.943	2.447	3.143	3.707	5.208	5.959
7		1.894	2.365	2.998	3.499	4.785	5.408
8		1.860	2.306	2.896	3.355	4.501	5.041
9		1.833	2.262	2.821	3.250	4.297	4.781
10		1.812	2.228	2.764	3.169	4.144	4.587
11		1.796	2.201	2.718	3.106	4.025	4.437
12		1.782	2.179	2.681	3.055	3.930	4.318
13		1.771	2.160	2.650	3.012	3.852	4.221
14		1.761	2.145	2.624	2.977	3.787	4.140
15		1.753	2.131	2.602	2.947	3.733	4.073
16		1.746	2.120	2.583	2.921	3.686	4.015
17		1.740	2.110	2.567	2.898	3.646	3.965
18		1.734	2.101	2.552	2.878	3.610	3.922
19		1.729	2.093	2.539	2.861	3.579	3.883
20		1.725	2.086	2.528	2.845	3.552	3.850
21		1.721	2.080	2.518	2.831	3.527	3.819
22		1.717	2.074	2.508	2.819	3.505	3.792
23		1.714	2.069	2.500	2.807	3.485	3.768
24		1.711	2.064	2.492	2.797	3.467	3.745
25		1.708	2.060	2.485	2.787	3.450	3.725
26		1.706	2.056	2.479	2.779	3.435	3.707
27		1.703	2.052	2.473	2.771	3.421	3.690
28		1.701	2.048	2.467	2.763	3.408	3.674
29		1.699	2.045	2.462	2.756	3.396	3.659
30		1.697	2.042	2.457	2.750	3.385	3.646
32		1.694	2.037	2.449	2.738	3.365	3.622
34		1.691	2.032	2.441	2.728	3.348	3.601
36		1.688	2.028	2.434	2.719	3.333	3.582
38		1.686	2.024	2.429	2.712	3.319	3.566
40		1.684	2.021	2.423	2.704	3.307	3.551
42		1.682	2.018	2.418	2.698	3.296	3.538
44		1.680	2.015	2.414	2.692	3.286	3.526
46		1.679	2.013	2.410	2.687	3.277	3.515
48		1.677	2.011	2.407	2.682	3.269	3.505
50		1.676	2.009	2.403	2.678	3.261	3.496
60		1.671	2.000	2.390	2.660	3.232	3.460
70		1.667	1.994	2.381	2.648	3.211	3.435
80		1.664	1.990	2.374	2.639	3.195	3.416

90		1.662	1.987	2.368	2.632	3.183	3.402
100		1.660	1.984	2.364	2.626	3.174	3.390
120		1.658	1.980	2.358	2.617	3.160	3.373
150		1.655	1.976	2.351	2.609	3.145	3.357
200		1.653	1.972	2.345	2.601	3.131	3.340
300		1.650	1.968	2.339	2.592	3.118	3.323
400		1.649	1.966	2.336	2.588	3.111	3.315
500		1.648	1.965	2.334	2.586	3.107	3.310
600		1.647	1.964	2.333	2.584	3.104	3.307
∞		1.645	1.960	2.326	2.576	3.090	3.291

F Distribution : Critical Values of F (5%significancelevel)

v_1 1 2 3 4 5 6 7 8 9
10 12 14 16 18 20

v_2

1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.36	246.46	247.32	248.01
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.42	19.43	19.44	19.45
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.71	8.69	8.67	8.66
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.87	5.84	5.82	5.80
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.64	4.60	4.58	4.56
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.96	3.92	3.90	3.87
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.53	3.49	3.47	3.44
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.24	3.20	3.17	3.15
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.03	2.99	2.96	2.94
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.86	2.83	2.80	2.77
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.74	2.70	2.67	2.65
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.64	2.60	2.57	2.54
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.55	2.51	2.48	2.46
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.48	2.44	2.41	2.39
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.42	2.38	2.35	2.33
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.37	2.33	2.30	2.28
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.33	2.29	2.26	2.23
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.29	2.25	2.22	2.19
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.26	2.21	2.18	2.16
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.22	2.18	2.15	2.12
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.20	2.16	2.12	2.10
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.17	2.13	2.10	2.07
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.15	2.11	2.08	2.05
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.13	2.09	2.05	2.03
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.11	2.07	2.04	2.01
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.09	2.05	2.02	1.99
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.08	2.04	2.00	1.97
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.06	2.02	1.99	1.96
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.05	2.01	1.97	1.94
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.04	1.99	1.96	1.93
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.04	1.99	1.94	1.91	1.88
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.95	1.90	1.87	1.84
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95	1.89	1.85	1.81	1.78
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.86	1.82	1.78	1.75
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.89	1.84	1.79	1.75	1.72
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.88	1.82	1.77	1.73	1.70
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.86	1.80	1.76	1.72	1.69
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.85	1.79	1.75	1.71	1.68
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.78	1.73	1.69	1.66
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.82	1.76	1.71	1.67	1.64
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.80	1.74	1.69	1.66	1.62

250	3.88	3.03	2.64	2.41	2.25	2.13	2.05	1.98	1.92	1.87	1.79	1.73	1.68	1.65	1.61
300	3.87	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.78	1.72	1.68	1.64	1.61
400	3.86	3.02	2.63	2.39	2.24	2.12	2.03	1.96	1.90	1.85	1.78	1.72	1.67	1.63	1.60
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.77	1.71	1.66	1.62	1.59
600	3.86	3.01	2.62	2.39	2.23	2.11	2.02	1.95	1.90	1.85	1.77	1.71	1.66	1.62	1.59
750	3.85	3.01	2.62	2.38	2.23	2.11	2.02	1.95	1.89	1.84	1.77	1.70	1.66	1.62	1.58
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.76	1.70	1.65	1.61	1.58

(continued)

***f* Distribution : Critical Values of *f* (5%significancelevel)**

v_1 25 30 35 40 50 60 75 100 150 200

v_2

1	249.26	250.10	250.69	251.14	251.77	252.20	252.62	253.04	253.46	253.68
2	19.46	19.46	19.47	19.47	19.48	19.48	19.48	19.49	19.49	19.49
3	8.63	8.62	8.60	8.59	8.58	8.57	8.56	8.55	8.54	8.54
4	5.77	5.75	5.73	5.72	5.70	5.69	5.68	5.66	5.65	5.65
5	4.52	4.50	4.48	4.46	4.44	4.43	4.42	4.41	4.39	4.39
6	3.83	3.81	3.79	3.77	3.75	3.74	3.73	3.71	3.70	3.69
7	3.40	3.38	3.36	3.34	3.32	3.30	3.29	3.27	3.26	3.25
8	3.11	3.08	3.06	3.04	3.02	3.01	2.99	2.97	2.96	2.95
9	2.89	2.86	2.84	2.83	2.80	2.79	2.77	2.76	2.74	2.73
10	2.73	2.70	2.68	2.66	2.64	2.62	2.60	2.59	2.57	2.56
11	2.60	2.57	2.55	2.53	2.51	2.49	2.47	2.46	2.44	2.43
12	2.50	2.47	2.44	2.43	2.40	2.38	2.37	2.35	2.33	2.32
13	2.41	2.38	2.36	2.34	2.31	2.30	2.28	2.26	2.24	2.23
14	2.34	2.31	2.28	2.27	2.24	2.22	2.21	2.19	2.17	2.16
15	2.28	2.25	2.22	2.20	2.18	2.16	2.14	2.12	2.10	2.10
16	2.23	2.19	2.17	2.15	2.12	2.11	2.09	2.07	2.05	2.04
17	2.18	2.15	2.12	2.10	2.08	2.06	2.04	2.02	2.00	1.99
18	2.14	2.11	2.08	2.06	2.04	2.02	2.00	1.98	1.96	1.95
19	2.11	2.07	2.05	2.03	2.00	1.98	1.96	1.94	1.92	1.91
20	2.07	2.04	2.01	1.99	1.97	1.95	1.93	1.91	1.89	1.88
21	2.05	2.01	1.98	1.96	1.94	1.92	1.90	1.88	1.86	1.84
22	2.02	1.98	1.96	1.94	1.91	1.89	1.87	1.85	1.83	1.82
23	2.00	1.96	1.93	1.91	1.88	1.86	1.84	1.82	1.80	1.79
24	1.97	1.94	1.91	1.89	1.86	1.84	1.82	1.80	1.78	1.77
25	1.96	1.92	1.89	1.87	1.84	1.82	1.80	1.78	1.76	1.75
26	1.94	1.90	1.87	1.85	1.82	1.80	1.78	1.76	1.74	1.73

27	1.92	1.88	1.86	1.84	1.81	1.79	1.76	1.74	1.72	1.71
28	1.91	1.87	1.84	1.82	1.79	1.77	1.75	1.73	1.70	1.69
29	1.89	1.85	1.83	1.81	1.77	1.75	1.73	1.71	1.69	1.67
30	1.88	1.84	1.81	1.79	1.76	1.74	1.72	1.70	1.67	1.66
35	1.82	1.79	1.76	1.74	1.70	1.68	1.66	1.63	1.61	1.60
40	1.78	1.74	1.72	1.69	1.66	1.64	1.61	1.59	1.56	1.55
50	1.73	1.69	1.66	1.63	1.60	1.58	1.55	1.52	1.50	1.48
60	1.69	1.65	1.62	1.59	1.56	1.53	1.51	1.48	1.45	1.44
70	1.66	1.62	1.59	1.57	1.53	1.50	1.48	1.45	1.42	1.40
80	1.64	1.60	1.57	1.54	1.51	1.48	1.45	1.43	1.39	1.38
90	1.63	1.59	1.55	1.53	1.49	1.46	1.44	1.41	1.38	1.36
100	1.62	1.57	1.54	1.52	1.48	1.45	1.42	1.39	1.36	1.34
120	1.60	1.55	1.52	1.50	1.46	1.43	1.40	1.37	1.33	1.32
150	1.58	1.54	1.50	1.48	1.44	1.41	1.38	1.34	1.31	1.29
200	1.56	1.52	1.48	1.46	1.41	1.39	1.35	1.32	1.28	1.26
250	1.55	1.50	1.47	1.44	1.40	1.37	1.34	1.31	1.27	1.25
300	1.54	1.50	1.46	1.43	1.39	1.36	1.33	1.30	1.26	1.23
400	1.53	1.49	1.45	1.42	1.38	1.35	1.32	1.28	1.24	1.22
500	1.53	1.48	1.45	1.42	1.38	1.35	1.31	1.28	1.23	1.21
600	1.52	1.48	1.44	1.41	1.37	1.34	1.31	1.27	1.23	1.20
750	1.52	1.47	1.44	1.41	1.37	1.34	1.30	1.26	1.22	1.20
1000	1.52	1.47	1.43	1.41	1.36	1.33	1.30	1.26	1.22	1.19

f Distribution : Critical Values of *f*(1%significancelevel)

$v_1$1 2 3 4 5 6 7 8 9 10
12 14 16 18 20

v_2

1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6106.32	6142.67	6170.10	6191.53	6208.73
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.44	99.44	99.45
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.92	26.83	26.75	26.69
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.25	14.15	14.08	14.02
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.77	9.68	9.61	9.55
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.60	7.52	7.45	7.40
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.36	6.28	6.21	6.16
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.56	5.48	5.41	5.36
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	5.01	4.92	4.86	4.81
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.60	4.52	4.46	4.41
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.29	4.21	4.15	4.10
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.05	3.97	3.91	3.86
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.86	3.78	3.72	3.66
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.70	3.62	3.56	3.51
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.56	3.49	3.42	3.37
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.45	3.37	3.31	3.26
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.35	3.27	3.21	3.16
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.27	3.19	3.13	3.08
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.19	3.12	3.05	3.00
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.13	3.05	2.99	2.94
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.07	2.99	2.93	2.88
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	3.02	2.94	2.88	2.83

23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.97	2.89	2.83	2.78
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.93	2.85	2.79	2.74
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.89	2.81	2.75	2.70
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.86	2.78	2.72	2.66
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.82	2.75	2.68	2.63
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.79	2.72	2.65	2.60
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.77	2.69	2.63	2.57
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.74	2.66	2.60	2.55
35	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88	2.74	2.64	2.56	2.50	2.44
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.56	2.48	2.42	2.37
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.56	2.46	2.38	2.32	2.27
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.39	2.31	2.25	2.20
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.59	2.45	2.35	2.27	2.20	2.15
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.42	2.31	2.23	2.17	2.12
90	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52	2.39	2.29	2.21	2.14	2.09
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.37	2.27	2.19	2.12	2.07
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.23	2.15	2.09	2.03
150	6.81	4.75	3.91	3.45	3.14	2.92	2.76	2.63	2.53	2.44	2.31	2.20	2.12	2.06	2.00
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.27	2.17	2.09	2.03	1.97
250	6.74	4.69	3.86	3.40	3.09	2.87	2.71	2.58	2.48	2.39	2.26	2.15	2.07	2.01	1.95
300	6.72	4.68	3.85	3.38	3.08	2.86	2.70	2.57	2.47	2.38	2.24	2.14	2.06	1.99	1.94
400	6.70	4.66	3.83	3.37	3.06	2.85	2.68	2.56	2.45	2.37	2.23	2.13	2.05	1.98	1.92
500	6.69	4.65	3.82	3.36	3.05	2.84	2.68	2.55	2.44	2.36	2.22	2.12	2.04	1.97	1.92
600	6.68	4.64	3.81	3.35	3.05	2.83	2.67	2.54	2.44	2.35	2.21	2.11	2.03	1.96	1.91
750	6.67	4.63	3.81	3.34	3.04	2.83	2.66	2.53	2.43	2.34	2.21	2.11	2.02	1.96	1.90
1000	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.20	2.10	2.02	1.95	1.90

continued)

***f* Distribution : Critical Values of *f* (1%significancelevel)**

v_1	25	30	35	40	50	60	75	100	150	200
1	6239.83	6260.65	6275.57	6286.78	6302.52	6313.03	6323.56	6334.11	6344.68	6349.97
2	99.46	99.47	99.47	99.47	99.48	99.48	99.49	99.49	99.49	99.49
3	26.58	26.50	26.45	26.41	26.35	26.32	26.28	26.24	26.20	26.18
4	13.91	13.84	13.79	13.75	13.69	13.65	13.61	13.58	13.54	13.52
5	9.45	9.38	9.33	9.29	9.24	9.20	9.17	9.13	9.09	9.08
6	7.30	7.23	7.18	7.14	7.09	7.06	7.02	6.99	6.95	6.93
7	6.06	5.99	5.94	5.91	5.86	5.82	5.79	5.75	5.72	5.70
8	5.26	5.20	5.15	5.12	5.07	5.03	5.00	4.96	4.93	4.91
9	4.71	4.65	4.60	4.57	4.52	4.48	4.45	4.41	4.38	4.36
10	4.31	4.25	4.20	4.17	4.12	4.08	4.05	4.01	3.98	3.96
11	4.01	3.94	3.89	3.86	3.81	3.78	3.74	3.71	3.67	3.66
12	3.76	3.70	3.65	3.62	3.57	3.54	3.50	3.47	3.43	3.41
13	3.57	3.51	3.46	3.43	3.38	3.34	3.31	3.27	3.24	3.22
14	3.41	3.35	3.30	3.27	3.22	3.18	3.15	3.11	3.08	3.06
15	3.28	3.21	3.17	3.13	3.08	3.05	3.01	2.98	2.94	2.92
16	3.16	3.10	3.05	3.02	2.97	2.93	2.90	2.86	2.83	2.81
17	3.07	3.00	2.96	2.92	2.87	2.83	2.80	2.76	2.73	2.71
18	2.98	2.92	2.87	2.84	2.78	2.75	2.71	2.68	2.64	2.62
19	2.91	2.84	2.80	2.76	2.71	2.67	2.64	2.60	2.57	2.55
20	2.84	2.78	2.73	2.69	2.64	2.61	2.57	2.54	2.50	2.48
21	2.79	2.72	2.67	2.64	2.58	2.55	2.51	2.48	2.44	2.42

22	2.73	2.67	2.62	2.58	2.53	2.50	2.46	2.42	2.38	2.36
23	2.69	2.62	2.57	2.54	2.48	2.45	2.41	2.37	2.34	2.32
24	2.64	2.58	2.53	2.49	2.44	2.40	2.37	2.33	2.29	2.27
25	2.60	2.54	2.49	2.45	2.40	2.36	2.33	2.29	2.25	2.23
26	2.57	2.50	2.45	2.42	2.36	2.33	2.29	2.25	2.21	2.19
27	2.54	2.47	2.42	2.38	2.33	2.29	2.26	2.22	2.18	2.16
28	2.51	2.44	2.39	2.35	2.30	2.26	2.23	2.19	2.15	2.13
29	2.48	2.41	2.36	2.33	2.27	2.23	2.20	2.16	2.12	2.10
30	2.45	2.39	2.34	2.30	2.25	2.21	2.17	2.13	2.09	2.07
35	2.35	2.28	2.23	2.19	2.14	2.10	2.06	2.02	1.98	1.96
40	2.27	2.20	2.15	2.11	2.06	2.02	1.98	1.94	1.90	1.87
50	2.17	2.10	2.05	2.01	1.95	1.91	1.87	1.82	1.78	1.76
60	2.10	2.03	1.98	1.94	1.88	1.84	1.79	1.75	1.70	1.68
70	2.05	1.98	1.93	1.89	1.83	1.78	1.74	1.70	1.65	1.62
80	2.01	1.94	1.89	1.85	1.79	1.75	1.70	1.65	1.61	1.58
90	1.99	1.92	1.86	1.82	1.76	1.72	1.67	1.62	1.57	1.55
100	1.97	1.89	1.84	1.80	1.74	1.69	1.65	1.60	1.55	1.52
120	1.93	1.86	1.81	1.76	1.70	1.66	1.61	1.56	1.51	1.48
150	1.90	1.83	1.77	1.73	1.66	1.62	1.57	1.52	1.46	1.43
200	1.87	1.79	1.74	1.69	1.63	1.58	1.53	1.48	1.42	1.39
250	1.85	1.77	1.72	1.67	1.61	1.56	1.51	1.46	1.40	1.36
300	1.84	1.76	1.70	1.66	1.59	1.55	1.50	1.44	1.38	1.35
400	1.82	1.75	1.69	1.64	1.58	1.53	1.48	1.42	1.36	1.32
500	1.81	1.74	1.68	1.63	1.57	1.52	1.47	1.41	1.34	1.31
600	1.80	1.73	1.67	1.63	1.56	1.51	1.46	1.40	1.34	1.30
750	1.80	1.72	1.66	1.62	1.55	1.50	1.45	1.39	1.33	1.29
1000	1.79	1.72	1.66	1.61	1.54	1.50	1.44	1.38	1.32	1.28

10.12 REFERENCE FOR FURTHER READING

Following books are recommended for further reading:

- Statistics by Murray R, Spiegel, Larry J. Stephens, Mcgraw Hill International Publisher, 4th edition

- Fundamental of Mathematical Statistics by S. C. Gupta and V. K. Kapoor, Sultan Chand and Sons publisher, 11th edition

*****

11

THE CHI-SQUARE TEST

Unit Structure

11.0 OBJECTIVES

The chi-square test is a non-parametric test that compares two or more variables from randomly selected data. It helps find the relationship between two or more variables. The chi square distribution is a theoretical or mathematical distribution which has wide applicability in statistical work. The term 'chi square' (pronounced with a hard 'ch') is used because the Greek letter χ is used to define this distribution. It will be seen that the elements on which this distribution is based are squared, so that the symbol χ 2 is used to denote the distribution.

11.1 INTRODUCTION

We know that if the probability distribution of the discrete random variable X is known we can find the probability distribution of the random variable$Y = X^2$. One may be interested in knowing whether we can find the probability distribution of random variable $Y = X^2$ if the probability distribution of a continuous random variable X is known. The answer is affirmative and in particular, if X is a standard normal variate (a continuous random variable) so that its probability density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\left(-\frac{1}{2}\right)x^2}, -\infty < x < \infty \qquad = 0, \qquad otherwise$$

then $Y = X^2$is also a continuous random variable whose probability density function is given by

$$g(y) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-\frac{y}{2}}, \quad y \geq 0$$
$$= 0, \qquad otherwise$$

Here, the distribution of Y is known as Chi-square distribution with one degree of freedom.

More generally, if $X_1, X_2, \cdots\cdots X_n$ are n independent standard normal variates, then the distribution of random variable $U = X_1^2 + X_2^2 + \cdots\cdots + X_n^2$is given by the probability density function

$$h(u) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} e^{-\frac{u}{2}} u^{\frac{n}{2}-1}, \qquad u \geq 0, n \in \mathbb{N}$$
$$= 0, \qquad otherwise$$

where $\Gamma\left(\frac{n}{2}\right)$ is called gamma $\frac{n}{2}$and is given by $\Gamma\left(\frac{n}{2}\right) = \int_0^\infty e^{-t}\, t^{\frac{n}{2}-1}\, dt$

Here, the distribution of random variable U is called Chi-sqaure distribution with n degrees of freedom n degrees of freedom and U is called a Chi-square variate. Generally, a chi square variate is denoted by the square of greek letter chi, i.e χ^2. Thus if χ^2 denotes a Chi-square variate with n degrees of freedom, then its probability density function is given by

$$f(\chi^2) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n}{2}-1}, \qquad \chi^2 \geq 0, n \in \mathbb{N}$$
$$= 0, \qquad otherwise$$

11.2 PROPERTIES OF CHI-SQUARE VARIATE WITH *n* DEGREES OF FREEDOM

Let χ^2 denote a Chi-square variate with n degrees of freedom. Then we have the following properties:

1) The probability density function of χ^2 is given by

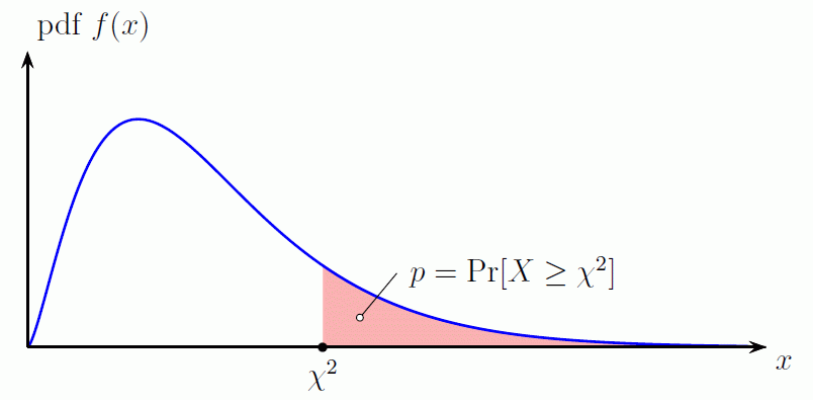
$$f(\chi^2) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n}{2}-1}, \qquad \chi^2 \geq 0, n \in \mathbb{N}$$
$$= 0, \qquad otherwise$$

2) The mean of χ^2 is E(χ^2) = n

3) The variance of χ^2 is $V(\chi^2)$ = 2n

4) The mode of χ^2 = n – 2

5) The frequency curve is given by $y = f(\chi^2)$ lies in the first quadrant and it is positively skewed curve, its tail on the right extends upto infinity, as given in the below diagram



$p = \Pr[X \geq \chi^2]$ is denoted by α and $\chi^2 = \chi^2_{n,\alpha}$

6) The total area under the Chi-square curve is 1.

7) $p(\chi^2 > c) = area\ under\ the\ curve\ y = f(\chi^2)\ to\ the\ right\ of\ \chi^2 = c$

8) For a chi square variate with n degrees of freedom, if $\boldsymbol{p(\chi^2 > c) = \alpha, then\ c\ is\ denoted\ by\ \chi^2_{n,\alpha}}$ $i.e\ p(\chi^2 > \chi^2_{n,\alpha}) = \alpha$

9) $\chi^2_{n,\alpha}$ is called α probability point of chi square distribution with n degrees of freedom.

Eg: 1) If a random variable X follows chi square distribution with 10 degrees of freedom find i) x_0 ii) x_1& iii) α such that $p(X > x_0) = 0.95, p(X \leq x_1) = 0.01$ & $p(X > 18.3) = \alpha$

Solution: n = 10 – degree of freedom

i) to find x_0 such that $p(X > x_0) = 0.95 \Rightarrow x_0 = \chi^2_{10,0.95} = 3.9403$

ii) to find x_1 such that $p(X \leq x_1) = 0.01$

$p(X \leq x_1) = 1 - p(X > x_1) \Rightarrow p(X > x_1) = 1 - 0.01 = 0.99$

$\Rightarrow p(X > x_1) = 0.99 \Rightarrow x_1 = \chi^2_{10,0.99} = 2.5582$

iii) $p(X > 18.3) = \alpha \Rightarrow \chi^2_{10,\alpha} = 18.3 \Rightarrow \alpha = 0.05$

Eg: 2) A random variable Y follows chi square distribution with S.D 4, Find y_0 if $p(Y \leq y_0) = 0.05$.

Solution: $S.D = 4 \Rightarrow var = 16 \Rightarrow 2n = 16 \Rightarrow n = 8$

$$p(Y \leq y_0) = 0.05 \Rightarrow p(Y \leq y_0) = 1 - p(Y > y_0) \Rightarrow p(Y > y_0)$$
$$= 1 - 0.05 = 0.95$$
$$p(Y > y_0) = 0.95 \Rightarrow y_0 = \chi^2_{8,0.95} \Rightarrow y_0 = 2.7326$$

Eg: 3) If a random variable X follows chi square distribution with S.D 4. Find mean and mode.

Solution: $S.D = 4 \Rightarrow Var = 16, but\ variance\ of\ \chi^2 = 2n \Rightarrow n = 8 mean\ of\ \chi^2 = n = 8\ \&\ mode\ of\ \chi^2 = n - 2 = 6$

11.3 THE CHI SQUARE TEST FOR GOODNESS OF FIT

When we come across some observations on a random variable, our curiosity may tempt us to investigate whether it can be considered to be a random variable following a certain specified probability law. A technique to test whether a given frequency distribution of a random variable follows a certain specified distribution (known as theoretical distribution) was proposed by Karl Pearson. It involves a test statistic that can be shown to follow Chi-square distribution under certain assumptions and the test is known as *Chi-square test for goodness of fit.*

Suppose we have an observed frequency distribution with n classes having observed frequencies
$O_1, O_2, O_3, \cdots\cdots, O_n\ with\ \sum_{i=1}^{i=n} O_i = N$. (Here the classes may correspond to the discrete values of a variable or groups of values of a variable or even to the groups corresponding to an attribute).

Further, suppose that according to our assumption to be called the null hypothesis H_0, the expected frequencies are $E_1, E_2, E_3, \cdots\cdots, E_n\ such\ that\ \sum_{i=1}^{i=n} E_i = N$.

Then the *test statistic* proposed by Karl Pearson is given by $\chi^2 = \sum_{i=1}^{i=n} \frac{(O_i - E_i)^2}{E_i}$

It can be shown that under the assumptions

1) observations are drawn independently and at random.

2) null hypothesis H_0 is true i.e the expected frequencies are $E_1, E_2, E_3, \cdots\cdots, E_n$ respectively.

3) total number of observations made = N is large
4) observed frequencies O_i's are large
5) expected frequencies E_i's are large so that the terms $\frac{O_i - E_i}{E_i}$ can be considered to be negligible.

The test statistic $\chi^2 = \sum_{i=1}^{i=n} \frac{(O_i - E_i)^2}{E_i}$ follow Chi-square distribution with (n-1) degrees of freedom.

For all practical purposes, the assumptions regarding the distribution of test statistic χ^2 can be considered to be reasonable if

a) the number of observations = N ≥ 50

b) the expected frequencies E_i's ≥ 5.

In case when the expected frequencies is less than 5 we require to combine more than one neighboring classes so that the expected frequency for such class is not less than 5.

11.3.1 DECISION CRITERION

While performing a test for goodness of fit, we shall use the following decision criteria:

$$Reject\ H_0\ if\ \chi^2 = \sum_{i=1}^{i=n} \frac{(O_i\ -\ E_i)^2}{E_i} > \chi^2_{(n-1),\alpha}$$

$$Do\ not\ reject\ H_0\ i.e\ accept\ H_0\ if\ \chi^2 \leq\ \chi^2_{(n-1),\alpha}$$

Eg: 1) The following data represents the last digit of the cars passing at a certain traffic signal observed during last 30 minutes for 180 cars.

last digit	0	1	2	3	4	5	6	7	8	9
frequency	12	20	14	12	21	18	17	26	19	21

Can we retain at 5% level of significance that all the digits are equally likely to occur?

Solution: we want to test whether all digits are equally likely to occur. If all digits are equally likely, $p = \frac{1}{10}. E_i = Np = 180 \times \frac{1}{10} = 18$

H_0: *All digits are equally likely to occur*

H_1: *not* H_0 *i.e logical negation to* H_0(*i.e all digits are not equally likely*)

$LOS = 5\%\ i.e\ \alpha = 0.05, N = 180 \geq 50\ and\ E_i = 18 \leq 5, n = 10 \chi^2$

$= \sum_{i=0}^{n} \frac{(O_i - E_i)^2}{E_i}\ follows\ chiq\ square\ distribution\ with\ 9\ degrees\ of\ freedom$

Decision criteria is given by

$reject\ H_0\ if\ \chi^2 > \chi^2_{9,0.05}$where $\chi^2_{9,0.05} = 16.9190$

$do\ not\ reject\ H_0\ if\ \chi^2 \leq \chi^2_{9,0.05}, \chi^2 = \sum_{i=0}^{n} \frac{(O_i - E_i)^2}{E_i}$

Digit	**observed freq. (O_i)**	**Exp. freq. (E_i)**	**O_i - E_i**	$\frac{(O_i - E_i)^2}{E_i}$
0	12	18	-6	36/18
1	20	18	2	4/18
2	14	18	-4	16/18
3	12	18	-6	36/18
4	21	18	3	9/18
5	18	18	0	0
6	17	18	-1	1/18
7	26	18	8	64/18
8	19	18	1	1/18
9	21	18	3	9/18
			total	176/18 = 9.7778

$$\chi^2 = \sum_{i=0}^{n} \frac{(O_i - E_i)^2}{E_i} = 9.7778 < 16.9190 = \chi^2_{9,0.05}$$

do not reject H_0 i.e accept H_0.

Eg: 2) As per Mendel's theory according to the shape and color, certain variety of pea that can be classified into four categories Round and yellow, Round and green, Angular and yellow, Angular and green occur in the proportion of 9:3:3:1. To test this a sample of N = 128 peas was taken and the following were the observed frequencies

RY – 66	RG – 28
AY – 29	AG – 5

Perform the chi square test for goodness of fit.

Solution: N = 128, n =4
The probability of occurrence and Expected frequencies are given by

category	p_i	$E_i = Np_i$
RY	9/16	128 X 9/16 = 72
RG	3/16	128 X 3/16 = 24
AY	3/16	128 X 3/16 = 24
AG	1/16	128 X1/16 = 8

$\therefore H_0$: $The\ four\ categories\ of\ peas\ i.e\ RY, RG, AY, AG$ have expected frequencies 72, 24, 24, 8 resp.

H_1: not H_0

LOS = 5%, $\alpha = 0.05$

Decision criteria is given by $Reject\ H_0\ if\ \chi^2 > \chi^2_{3,0.05} = 7.8147$

Do not reject H_0 if $\chi^2 \leq 7.8147, where\ \chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$

category of peas	obs. freq.(O_i)	Exp. freq. (E_i)	$O_i - E_i$	$\frac{(O_i - E_i)^2}{E_i}$
RY	66	72	-6	0.5
RG	28	24	4	0.6667
AY	29	24	5	1.0417
AG	5	8	-3	1.125
			total	3.3334

$$\chi^2 = 3.3334 < \chi^2_{3,0.05} = 7.8147$$

$$\Rightarrow Accept\ H_0\ at\ 5\%\ LOS.$$

Eg: 3) Four identical coins are tossed 100 times and the following results are obtained.

no. of heads (x)	0	1	2	3	4
frequency	8	29	40	19	4

Are there sufficient evidences to conclude that the coins are biased at 5% LOS.

Solution: let p denote the probability of getting a head with each of the four coins,

X: no. of heads follows binomial distribution with n= 4, p

X ~ B(n, p) i.e X ~ B(4,p)

H_0: the coins are unbiased i.e p =1/2

H_1 : the coins are baised i.e p ≠ ½

LOS = 5 %, $\alpha = 0.05$

$$X \sim B(4,p) \Rightarrow p(x) = \binom{4}{x} p^x q^{4-x}$$

$$p(0) = p^0 q^4 = \frac{1}{16}, p(1) = 4pq^3 = \frac{4}{16}, p(2) = 6p^2q^2 = \frac{6}{16}$$

$$p(3) = 4p^3q = \frac{4}{16}, \qquad p(4) = p^4q^0 = \frac{1}{16}$$

Decision criteria Reject $H_0\ if\ \chi^2 > \chi^2_{4,0.05} = 9.4877$

Expected frequencies $E_i = N \cdot p(x)$

x	obs. freq. O_i	Exp. freq. E_i	$O_i - E_i$	$\frac{(O_i - E_i)^2}{E_i}$
0	8	100p(0) = 6.25	1.75	0.49
1	29	100p(1) = 25	4	0.64
2	40	100p(2) = 37.5	2.5	0.1667
3	19	100p(3) = 25	-6	1.44
4	4	100p(4) = 6.25	-2.25	0.81
			total	3.5467

$$\chi^2 = 3.5467 < \chi^2_{4,0.05} = 9.4877$$
$$Accept\ H_0 at\ 5\%\ LOS \Rightarrow coins\ are\ unbiased$$

Eg: 4) The random variable X denotes the number of street accidents per week.

X	0	1	2	3	4	5	6	7
obs. freq.	15	30	28	14	8	4	0	1
exp. freq.	14	27	27	18	9	4	1	0

Test whether the random variable X follows Poisson distribution with parameter m =2 at 1% level of significance.

Solution: As the expected frequency is less than 5 for X = 5, 6 & 7 we combine them into one class

X	0	1	2	3	4	5-7
obs. freq.	15	30	28	14	8	5
exp. freq.	14	27	27	18	9	5

H_0: X follows Poisson distributions with parameter m =2

H_1: not H_0

LOS = 1%, $\alpha = 0.01$

Decision criteria is given by $Reject\ H_0\ if\ \chi^2 > \chi^2_{5,0.01} = 15.086$

Do not reject H_0 if $\chi^2 \leq \chi^2_{5,0.01}, where\ \chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$

i	X	obs. freq. O_i	exp. freq. E_i	$(O_i - E_i)$	$\frac{(O_i - E_i)^2}{E_i}$
1	0	15	14	1	0.0714
2	1	30	27	3	0.3333

3	2	28	27	1	0.0370
4	3	14	18	-4	0.8889
5	4	8	9	-1	0.1111
6	5-7	5	5	0	0
				total	1.4417

$$\chi^2 = 1.4417 < \chi^2_{5,0.01} = 15.086$$

Accept H_0 i.e at 1% level of significance the hypothesis that the variable X follows Poisson distribution with parameter m =2 is retainable.

11.4 TEST FOR INDEPENDENCE OF ATTRIBUTES

In this section, we consider the presence of two attributes among units from single population and the interest is centered around the possible dependence or independence of the attributes. In other words, on the basis of data regarding two attributes for some units from the population, we shall investigate whether the observed data provide sufficient reasons to reject the claim that the two attributes are independent of each other for the population under consideration. Such a test is called **test for independence of attributes.**

To understand the mechanism of the test, we consider the following table representing data for two attributes known as **2 × 2 contingency table.** It is called a contingency table as it represents the information which can be attributed to chance as the information is regarding randomly selected persons.

2 × 2 contingency table is given as below:

		Attribute B		total
		B1	B2	
Attribute A	A1	a	b	a+b
	A2	c	d	c+d
		a+c	b+d	a+b+c+d= N

With the help of the test statistic $\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$ we can perform a test for testing the hypothesis H_0: the attributes A and B under the consideration are independent against the logical alternative H_1: not H_0i.e the attributes A and B are dependent subject to the conditions $N \geq 50$ and each of the observed frequencies a, b, c, $d \geq 5$.

The decision criterion at level of significance = α is given by

$$Reject\ H_0 : if\ \chi^2 > \chi^2_{1,\alpha}, do\ not\ reject\ H_0 : if\ \chi^2 \leq \chi^2_{1,\alpha}$$

where $\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$ where a,b,c,d are the observed frequencies with a+b+c+d = N.

Eg: 1) The following results are obtained at the end of six months of a kind of psychotherapy given to a group of 120 patients and also for another group of 120 patients who were not given the psychotherapy.

	psychotherapy	
	given	not given
condition improved	71	42
condition did not improved	49	78

Can we conclude at 5% LOS that the psychotherapy is effective**?**

Solution: H_0: Psychotherapy is not effective

H_1: psychotherapy is effective

LOS = 5% ,$\alpha = 0.05$

Decision criteria Reject H_0 iff $\chi^2 > \chi^2_{1,0.05} = 3.8415$

N = 240, a =71, b = 42, c = 49, d = 78

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)} = \frac{240(71\times 78 - 49\times 42)^2}{113\cdot 120\cdot 120\cdot 127}$$

$$\chi^2 = 240 \times \frac{12110400}{206654400} = 14.0645$$

$\Rightarrow \chi^2 = 14.0645 > \chi^2_{1,0.05} = 3.8415$

$\Rightarrow$ Reject H_0 at 5% level of significance, we may say that the Psychotherapy is effective at 5% LOS.

11.5 YATE'S CORRECTION FOR CONTINUITY

When the cell frequencies a, b, c, d as observed in case of the four classes corresponding to two attributes are small, we cannot use the test statistic χ^2 as defined in previous section i.e for when the assumption a, b, c, d are greater than or equal to 5 does not hold, the distribution of $\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$ cannot be considered to be Chi-square with one degree of freedom.

In such a case i.e if the cell frequencies a, b, c, d are not all greater than or equal to 5, we make the following adjustment called **Yate's correction.**

- if ad < bc add ½ to a and b and subtract ½ from both b and c both.
- if ad > bc add ½ to b and c and subtract ½ from both a and b both.

With this adjustments, we get the test statistic as $\chi^2 = \frac{N\left(|ad-bc|-\frac{N}{2}\right)^2}{(a+b)(a+c)(b+d)(c+d)}$

The decision criterion at level of significance = α is given by

$$Reject\ H_0: if\ \chi^2 > \chi^2_{1,\alpha}, do\ not\ reject\ H_0 : if\ \chi^2 \leq \chi^2_{1,\alpha}$$

where $\chi^2 = \frac{N\left(|ad-bc|-\frac{N}{2}\right)^2}{(a+b)(a+c)(b+d)(b+c)}$

Eg: 1) In an experiment on immunization of cattle from tuberculosis the following results were obtained

	affected	unaffected
Inoculated	11	31
Not inoculated	14	4

Examine the effect of vaccine in controlling the incidence of the disease at 1% LOS.

Solution: H_0: the attributes are independent
H1: the attributes are not independent
LOS = 1%, $\alpha = 0.01$
Decision criteria Reject $H_0\ iff\ \chi^2 > \chi^2_{1,0.01} = 6.6349$

N = 60, a = 11, b = 31, c = 14, d = 4

$$\chi^2 = \frac{N\left(|ad-bc|-\frac{N}{2}\right)^2}{(a+b)(a+c)(b+d)(c+d)} = \frac{60(|11\times 4 - 31\times 14| - 30)^2}{42\cdot 18\cdot 25\cdot 35}$$

$$\chi^2 = \frac{60(390-30)^2}{661500} = \frac{60\times 360\times 360}{661500} = 11.7551$$

$\chi^2 = 11.7551 > \chi^2_{1,0.01} = 6.6349$

$\Rightarrow reject\ H_0\ at\ 1\%\ LOs$

We can say at 1% level of significance that the Inoculation and affection due to disease are dependent.

11.6 TEST IN r × c CONTINGENCY TABLE

If we have two attributes A and B classified into r and c classes respectively denoted by $A_1, A_2, \cdots\cdots A_r$ & $B_1, B_2, \cdots\cdots B_c$ then the observed frequency can be put in a tabular form with r rows and c columns called r × c contingency table.

If we use O_{ij} (i = 1,2,3,……,r , j = 1,2,3,……..,c)to denote the observed frequency for attribute class A_iB_j, then r × c contingency table can be represented as shown below:

Attribute	B_1	B_2	B_3			B_C	TOTAL
A_1	O_{11}	O_{12}	O_{13}			O_{1C}	a_1
A_2	O_{21}	O_{22}	O_{23}			O_{2C}	a_2
A_3	O_{31}	O_{32}	O_{33}			O_{3C}	a_3
A_R	O_{R1}	O_{R2}	O_{R3}			O_{RC}	a_r
TOTAL	b_1	b_2	b_3			b_c	N

Here a_i's represent total observed frequencies for attribute classes A_i's and b_j's represent the same for classes B_j's, N being the overall total frequency.

Then with these notations, we can write down the test statistic as $\chi^2 = \sum_i \sum_j \frac{(O_{ij}-E_{ij})^2}{E_{ij}}, where\ E_{ij} = \frac{a_i b_j}{N}$ are the expected frequencies.

With the help of the test statistic $\chi^2 = \sum_i \sum_j \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ we can perform a test for testing the hypothesis H_0: the attributes A and B under the consideration are independent against the logical alternative H_1: not H_0i.e the attributes A and B are dependent subject to the conditions $N \geq 50$ and each of the observed frequencies $\geq 5(O_{ij} \geq 5)$.

The decision criterion at level of significance = α is given by

$$Reject\ H_0 : if\ \chi^2 > \chi^2_{(r-1)(c-1),\alpha}$$

$$do\ not\ reject\ H_0 : if\ \chi^2 \leq \chi^2_{(r-1)(c-1),\alpha}$$

Eg: 1) Using the data given in the following table decide whether we can conclude that standard of clothing of a salesman has significant effect on his performance in field selling at 5% LOS.

	Performance in Field Selling			
	Disappointing	Satisfactory	Excellent	Total
Poorly dressed	21	15	6	42
Well dressed	24	35	26	85
Very well dressed	35	80	58	173
Total	80	130	90	300

Solution: H_0: attributes are independent
LOS = 5%, $\alpha = 0.05$

Reject H_0 iff $\chi^2 > \chi^2_{4,0.05} = 9.4877, where\ \chi^2 = \sum_i \sum_j \frac{(O_{Ij} - E_{ij})^2}{E_{ij}}$

$$E_{11} = \frac{a_1 b_1}{N} = \frac{42 \cdot 80}{300} = 11.2, E_{12} = \frac{a_1 b_2}{N} = \frac{42 \cdot 130}{300} = 18.2$$

$$E_{13} = \frac{a_1 b_3}{N} = \frac{42 \cdot 90}{300} = 12.6, E_{21} = \frac{a_2 b_1}{N} = \frac{85 \cdot 80}{300} = 22.67$$

$$E_{31} = \frac{a_3 b_1}{N} = \frac{173 \cdot 80}{300} = 46.13, E_{32} = \frac{a_3 b_2}{N} = \frac{173 \cdot 130}{300} = 74.97$$

$$E_{33} = \frac{a_3 b_3}{N} = \frac{173 \cdot 90}{300} = 51.9$$

O_{ij}	E_{ij}	$O_{ij} - E_{ij}$	$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
21	11.2	9.8	8.575
15	18.2	-3.2	0.5626
6	12.6	-6.6	3.4571
24	22.67	1.33	0.078
35	36.83	-1.83	0.0909
26	25.5	0.5	0.0098
35	46.13	-11.13	2.6854
80	74.97	5.03	0.3375
58	51.9	6.1	0.7170
		total	16.5133

$\chi^2 = 16.5133 > \chi^2_{4,0.05} = 9.488$

$\Rightarrow reject\ H_0\ at\ 5\%\ LOS$

We decide to reject H_0 at 5% level of significance and conclude that the standard of clothing of a salesman has significant effect on his performance in field selling.

11.7 COEFFICIENT OF CONTINGENCY

A measure of the degree of relationship, association, or dependence of the classifications in a contingency table is given by

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

which is called the *coefficient of contingency.* The larger the C, the greater is the degree of association. The number of rows and columns in the contingency table determines the maximum value of C, which is never greater than 1. If the number of rows and columns of a contingency table is equal to k, the maximum value of C is given by $\sqrt{\frac{k-1}{k}}$.

Eg: 1) Using the data given in the following table decide whether we can conclude that standard of clothing of a salesman has significant effect on his performance in field selling at 5% LOS.

	Performance in Field Selling			
	Disappointing	Satisfactory	Excellent	Total
Poorly dressed	21	15	6	42
Well dressed	24	35	26	85
Very well dressed	35	80	58	173
Total	80	130	90	300

Also find the coefficient of contingency.

Solution: H_0: attributes are independent

LOS = 5%, $\alpha = 0.05$

Reject H_0 iff $\chi^2 > \chi^2_{4,0.05} = 9.4877, where\ \chi^2 = \sum_i \sum_j \frac{(O_{Ij} - E_{ij})^2}{E_{ij}}$

$$E_{11} = \frac{a_1 b_1}{N} = \frac{42 \cdot 80}{300} = 11.2, E_{12} = \frac{a_1 b_2}{N} = \frac{42 \cdot 130}{300} = 18.2$$

$$E_{13} = \frac{a_1 b_3}{N} = \frac{42 \cdot 90}{300} = 12.6, E_{21} = \frac{a_2 b_1}{N} = \frac{85 \cdot 80}{300} = 22.67$$

$$E_{31} = \frac{a_3 b_1}{N} = \frac{173 \cdot 80}{300} = 46.13, E_{32} = \frac{a_3 b_2}{N} = \frac{173 \cdot 130}{300} = 74.97$$

$$E_{33} = \frac{a_3 b_3}{N} = \frac{173 \cdot 90}{300} = 51.9$$

O_{ij}	E_{ij}	$O_{ij} - E_{ij}$	$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
21	11.2	9.8	8.575
15	18.2	-3.2	0.5626
6	12.6	-6.6	3.4571
24	22.67	1.33	0.078
35	36.83	-1.83	0.0909
26	25.5	0.5	0.0098
35	46.13	-11.13	2.6854
80	74.97	5.03	0.3375
58	51.9	6.1	0.7170
		total	16.5133

$\chi^2 = 16.5133 > \chi^2_{4,0.05} = 9.488 \Rightarrow reject\ H_0\ at\ 5\%\ LOS$

We decide to reject H_0 at 5% level of significance and conclude that the standard of clothing of a salesman has significant effect on his performance in field selling.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{16.5133}{16.5133 + 300}} = 0.2284$$

11.8 CORRELATION OF ATTRIBUTES

Because classifications in a contingency table often describe characteristics of individuals or objects, they are often referred to as *attributes*, and the degree of dependence, association, or relationship is called the *correlation of attributes*. For k X k tables, we define

$$r = \sqrt{\frac{\chi^2}{N(k\ -\ 1)}}$$

as the correlation coefficient between attributes(or classification). This coefficient lies between 0 and 1. For 2 × 2 tables in which k = 2, the correlation is often called *tetrachoric* correlation.

11.9 ADDITIVE PROPERTY OF CHI SQUARE VARIATE

Suppose that the results of repeated experiments yield sample values of χ^2 given by

$\chi_1^2, \chi_2^2, \chi_3^2, \cdots\cdots, \chi_n^2$ with $v_1, v_2, v_3, \cdots\cdots, v_n$ degrees of freedom, respectively. Then the result of all these experiment can be considered equivalent to χ^2 value given by $\chi_1^2 + \chi_2^2 + \chi_3^2 + \cdots\cdots + \chi_n^2$

with $v_1 + v_2 + v_3 + \cdots\cdots + v_n$ degrees of freedom.

11.10 SUMMARY

In this chapter we discussed about the chi-square test which is a non-parametric test that compares two or more variables from randomly selected data. It helps to find the relationship between two or more variables. The chi square distribution is a theoretical or mathematical distribution which has wide applicability in statistical work. We had also seen the various properties of Chi-square variate along with different tests to check whether the variable follows Chi-square distribution, along with Yate's correction.

11.11 EXERCISES

1) If X is a chi square variate with 17 degreed of freedom findx_0 , x_1 and α such that
$$p(X > x_0) = 0.01, p(X \leq x_1) = 0.95 \ \& \ p(X \leq 8.67) = \alpha.$$

2) The following table gives the result of investigation in the association of eye color and hair color. Can we deduce that the two attributes are independent?

	Hair color

	Brown	Black
Eye color Blue	75	25
Brown	65	35

3) The following data show the classification of individuals with respect to gender and literacy in a random sample of 200. Test the data for independence of attributes using chi square at 1% LOS.

	literate	illiterate
male	95	5
female	75	25

4) The eyesight of 100 randomly selected people from a town were tested with the following results:

	poor eyesight	good eyesight
male	200	350
female	200	250

Can we conclude at 5% level of significance that gender has no bearing on the quality of eyesight?

5) The following table shows results of inoculation against cholera

	not attacked	attacked
Inoculated	446	4
Not inoculated	291	9

Can we say at 5% LOS that inoculation is effective in controlling susceptibility of cholera?

6) The following are the results of the tests performed on two brands of tyres manufactured by a manufacturer.

	Brand A	Brand B
Lasted more than 30000km	27	38
Failed to last 30000 km	18	27

Use chi square at 5% LOS to test whether we can say that the two brands of tyres differ significantly or not as regards their lifespan.

7) Determine at 1% LOS whether vaccination can be regarded as a preventive measure for small pox on the basis of following report:
Out of 1482 persons in a locality exposed to small pox, 368 in all were attacked. Out of 1482 persons, 343 had been vaccinated and of these only 35 were attacked.

8) During a market research survey organized by ABC Ltd. the households were asked whether they used "Beauty Soap"(the price of which is Rs. 7.25 a piece) and whether their per capita monthly expenditure exceeded Rs. 800 with the following results.

		Monthly per capita expenditure	
		Exceeded Rs 800	Did not exceed Rs. 800
Whether they used	yes	6	20
"Beauty Soap"	no	4	30

Can we say the use of "Beauty Soap" depends on monthly per capita expenditure at 5% LOS.

9) In a household survey conducted in certain locality, the following information is collected.

	Whether exclusive Indoor toilet facility is available		
	yes	no	
owned house	9	4	
rented house	21	16	

Use chi square test for independence at 1% LOS and give your conclusions.

10) The following data refer to an investigation carried out to examine the effect of T.V as a medium of advertisement on the turnover of the company's manufacturing and selling consumer products. A random sample of 40 companies was selected. Analyze the data and comment on your findings.

	Annual turnover exceeding one crore rupees	
	Yes	No
T.V advertisement	7	3
No T.V advertisement	10	20

11) A random sample of students of Mumbai University was selected and asked their opinion about autonomous college. The results are given below. The same number of each gender was included within each class group. Test the hypothesis at 5% LOS that opinions are independent of the class groupings.

	favoring autonomous colleges	opposed to autonomous college
First Year	120	80
Second Year	130	70
Third year	70	30
Post Graduation	80	20

12) Test for independence between health and working capacity from one following data:

		Health		
		very good	good	fair
working	good	20	25	15
Capacity	bad	10	15	15

13) ABC Ltd. employ a large number of handicapped persons. The following is an account of the performance of 200 randomly chosen employees of the company:

	performance		
	above average	average	below average
handicapped	29	31	20
non handicapped	55	30	35

Can we retain at 5% LOS that handicapped employees are equally efficient as the non-handicapped employees of the company.

14) In a survey 100 couples are interviewed and they were asked to give an opinion on the importance of amiable nature of partner in selection of a bride or a groom. The ranks were given independently by them as I or II or III.

		Rankings by wives		
		I	II	III
Ranking	I	25	12	8
by	II	10	23	8
Husbands	III	5	5	4

Use chi square test for independence and comment at 5% LOS.

15) A socio economic survey conducted in 1981 in Mumbai revealed the following results:

	Monthly family income		
	Below 1200	1200 to 1800	1800 and above
No child	18	15	12
one child	31	34	25
Two or more children	81	51	63

Can we regard at 1% LOS that the number of children in the family has no association with monthly income?

16) The data in the table were collected on how individuals prepared their taxes and their education level. The null hypothesis is that the way people prepare their taxes (computer software or pen and paper) is independent of their education level. The following table is the contingency table.

	Education		
Tax Prepare	High School	Bachelors	Masters
Computer Software	23	35	42
Pen and Paper	45	30	25

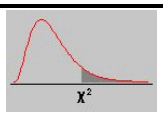
Find the coefficient of contingency.

11.12 SOLUTION TO EXERCISE

Q. No.	Solution	Q. No.	Solution

1	$x_0 = 33.4087, x_1 = 27.5871,$ $\alpha = 0.05$	2	2.38, accept H_0
3	15.686, reject H_0	4	3.84, donot reject H_0
5	3.552, accept H_0	6	0.026, accept
7	51.157 reject	8	0.6652 accept
9	0.2122, accept	10	2.762, accept
11	12.046 , reject	12	1.9097, accept
13	4.904, accept	14	10.8125, accept
15	4.305, accept	16	0.236

11.13 TABLE OF CHI-SQUARE DISTRIBUTION



df\area	0.995	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01	0.005
1	0.00004	0.00016	0.00098	0.00393	0.01579	0.10153	0.45494	1.3233	2.70554	3.84146	5.02389	6.6349	7.87944
2	0.01003	0.0201	0.05064	0.10259	0.21072	0.57536	1.38629	2.77259	4.60517	5.99146	7.37776	9.21034	10.59663
3	0.07172	0.11483	0.2158	0.35185	0.58437	1.21253	2.36597	4.10834	6.25139	7.81473	9.3484	11.34487	12.83816
4	0.20699	0.29711	0.48442	0.71072	1.06362	1.92256	3.35669	5.38527	7.77944	9.48773	11.14329	13.2767	14.86026
5	0.41174	0.5543	0.83121	1.14548	1.61031	2.6746	4.35146	6.62568	9.23636	11.0705	12.8325	15.08627	16.7496
6	0.67573	0.87209	1.23734	1.63538	2.20413	3.4546	5.34812	7.8408	10.64464	12.59159	14.44938	16.81189	18.54758
7	0.98926	1.23904	1.68987	2.16735	2.83311	4.25485	6.34581	9.03715	12.01704	14.06714	16.01276	18.47531	20.27774
8	1.34441	1.6465	2.17973	2.73264	3.48954	5.07064	7.34412	10.21885	13.36157	15.50731	17.53455	20.09024	21.95495
9	1.73493	2.0879	2.70039	3.32511	4.16816	5.89883	8.34283	11.38875	14.68366	16.91898	19.02277	21.66599	23.58935
10	2.15586	2.55821	3.24697	3.9403	4.86518	6.7372	9.34182	12.54886	15.98718	18.30704	20.48318	23.20925	25.18818
11	2.60322	3.05348	3.81575	4.57481	5.57778	7.58414	10.341	13.70069	17.27501	19.67514	21.92005	24.72497	26.75685
12	3.07382	3.57057	4.40379	5.22603	6.3038	8.43842	11.34032	14.8454	18.54935	21.02607	23.33666	26.21697	28.29952
13	3.56503	4.10692	5.00875	5.89186	7.0415	9.29907	12.33976	15.98391	19.81193	22.36203	24.7356	27.68825	29.81947
14	4.07467	4.66043	5.62873	6.57063	7.78953	10.16531	13.33927	17.11693	21.06414	23.68479	26.11895	29.14124	31.31935
15	4.60092	5.22935	6.26214	7.26094	8.54676	11.03654	14.33886	18.24509	22.30713	24.99579	27.48839	30.57791	32.80132
16	5.14221	5.81221	6.90766	7.96165	9.31224	11.91222	15.3385	19.36886	23.54183	26.29623	28.84535	31.99993	34.26719
17	5.69722	6.40776	7.56419	8.67176	10.08519	12.79193	16.33818	20.48868	24.76904	27.58711	30.19101	33.40866	35.71847
18	6.2648	7.01491	8.23075	9.39046	10.86494	13.67529	17.3379	21.60489	25.98942	28.8693	31.52638	34.80531	37.15645
19	6.84397	7.63273	8.90652	10.11701	11.65091	14.562	18.33765	22.71781	27.20357	30.14353	32.85233	36.19087	38.58226
20	7.43384	8.2604	9.59078	10.85081	12.44261	15.45177	19.33743	23.82769	28.41198	31.41043	34.16961	37.56623	39.99685
21	8.03365	8.8972	10.2829	11.59131	13.2396	16.34438	20.33723	24.93478	29.61509	32.67057	35.47888	38.93217	41.40106
22	8.64272	9.54249	10.98232	12.33801	14.04149	17.23962	21.33704	26.03927	30.81328	33.92444	36.78071	40.28936	42.79565
23	9.26042	10.19572	11.68855	13.09051	14.84796	18.1373	22.33688	27.14134	32.0069	35.17246	38.07563	41.6384	44.18128
24	9.88623	10.85636	12.40115	13.84843	15.65868	19.03725	23.33673	28.24115	33.19624	36.41503	39.36408	42.97982	45.55851
25	10.51965	11.52398	13.11972	14.61141	16.47341	19.93934	24.33659	29.33885	34.38159	37.65248	40.64647	44.3141	46.92789
26	11.16024	12.19815	13.8439	15.37916	17.29188	20.84343	25.33646	30.43457	35.56317	38.88514	41.92317	45.64168	48.28988
27	11.80759	12.8785	14.57338	16.1514	18.1139	21.7494	26.33634	31.52841	36.74122	40.11327	43.19451	46.96294	49.64492
28	12.46134	13.56471	15.30786	16.92788	18.93924	22.65716	27.33623	32.62049	37.91592	41.33714	44.46079	48.27824	50.99338
29	13.12115	14.25645	16.04707	17.70837	19.76774	23.56659	28.33613	33.71091	39.08747	42.55697	45.72229	49.58788	52.33562
30	13.78672	14.95346	16.79077	18.49266	20.59923	24.47761	29.33603	34.79974	40.25602	43.77297	46.97924	50.89218	53.67196

11.14 REFERENCE FOR FURTHER READING

Following books are recommended for further reading:

- Statistics by Murray R, Spiegel, Larry J. Stephens, Mcgraw Hill International Publisher, 4th edition

- Fundamental of Mathematical Statistics by S. C. Gupta and V. K. Kapoor, Sultan Chand and Sons publisher, 11th edition

*****

UNIT V

12

CURVE FITTING AND THE METHOD OF LEAST SQUARES

Unit Structure

12.0 OBJECTIVES

The main objective is to study the fitting of curves using the method of least square which minimizes the sum of the square of the errors.

12.1 INTRODUCTION

When we come across data for two variables and think about the relation between each other, two objects come to our mind:

1) To translate the possible relationship into a mathematical equation(called equation of curve)
2) To exploit the relationship between the variables for estimating the value of one variable corresponding to a given value of the other variable.

Curve fitting helps us in serving the first object directly, i.e., it leads to a mathematical equation describing the relationship between the variables. Further, using this equation one can estimate the value of the other variable.

12.2 RELATIONSHIP BETWEEN VARIABLES

Very often in practice a relationship is found to exist between two (or more) variables. For example, weights of adult males depend to some degree on their heights, the circumferences of circles depend on their radii, and the pressure of a given mass of gas depends on its temperature and volume.

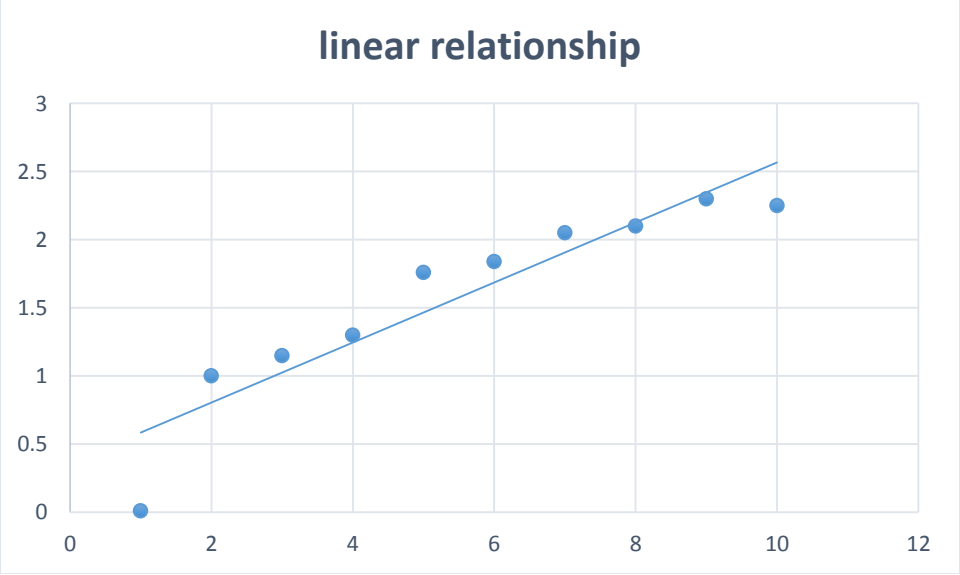
It is frequently desirable to express this relationship in mathematical form by determining an equation that connects the variables.

12.3 CURVE FITTING

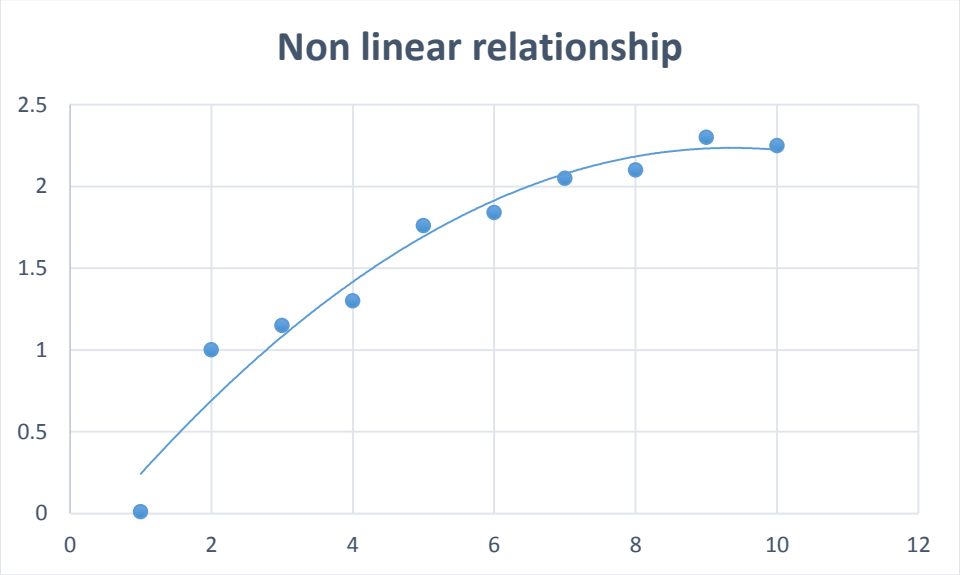
To determine an equation that connects variables, a first step is to collect data that show corresponding values of the variables under consideration. For example, suppose X and Y denote respectively, the height and weight of adult males; then a sample of N individuals would reveal the heights $X_1, X_2, X_3, \cdots\cdots, X_N$ and the corresponding weights $Y_1, Y_2, Y_3, \cdots\cdots Y_N$.

A next step is plot the points $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \cdots\cdots, (X_N, Y_N)$ on a rectangular coordinate system. The resulting set of points is some times called a **scatter diagram**. From the scatter diagram it is often possible to visualize a smooth curve that approximates the data. Such a curve is called an **approximating curve**.

For example, in the following diagram the data appear to be approximated by a straight line, and so we call that a linear relationship exists between the variables.



In the following diagram, however the relationship exists between the variables, it is not a linear relationship, and so we call it a non-linear relationship.



The general problem of finding equations of approximating curves that fit the given sets of data is called **curve fitting.**

12.4 EQUATIONS OF APPROXIMATING CURVES

Various common types of approximating curves and their equations are given below for reference. All letters other than *x* and *y* shall be treated as constants. The variables *x* and *y* are termed as independent variable and dependent variable respectively, roles of the variables can be interchanged as per the requirement.

Straight Line	$y = a + bx$
Parabola or quadratic curve	$y = a + bx + cx^2$
Cubic curve	$y = a + bx + cx^2 + dx^3$
Quadratic curve	$y = a + bx + cx^2 + dx^3 + ex^4$
nth degree curve	$y = a_0 + a_1x + a_2x^2 + \cdots\cdots + a_nx^n$

The above equations refer to the polynomial equations of the degree one, two, three, four and n respectively. The following are some examples of other equations most commonly used.

Hyperbola	$y = \frac{1}{a_0 + a_1x}$
Exponential curve	$y = a \cdot b^x$
Geometric curve	$y = a \cdot x^b$
Gompertz Curve	$y = pq^{b^x}$
Logistic Curve	$y = \frac{1}{ab^x + c}$

To decide which curve to be used, it is helpful to obtain scatter diagrams of the transformed variables. For example, if a scatter diagram of $\log y$versus x shows a linear relationship, then the equation is of the$y = a \cdot b^x$, while if $\log y$ versus $\log x$ shows a linear relationship, the equation is of the form $y = a \cdot x^b$.

12.5 FREEHAND METHOD OF CURVE FITTING

Individual judgment can often be used to draw an approximating curve to fit a set of data. This is called a freehand method of curve fitting. If the type of equation of this curve is known, it is possible to obtain the constants in the equation by choosing as many points on the curve as there are constants in the equation. For example, if the curve is a straight line, two points are necessary; if it is a parabola, three points are necessary. The method has the disadvantage that different observers will obtain different curves and equations.

12.6 THE STRAIGHT LINE METHOD

The simplest type of the approximating curve is a straight line, whose equation can be written

$$y = a + bx$$

Given any two points $(x_1, y_1)\&(x_2, y_2)$ on the line, the constants a and b can be determined. The resulting equation of the straight line can be written as

$$y - y_1 = m(x - x_1), m = \frac{y_2 - y_1}{x_2 - x_1}$$

where m is known as the slope of the line.

When the equation is written in the form of $y = a + bx$, the constant 'b' denotes the slope m. The constant 'a' is the value of y when x = 0 is called the y-intercept.

12.7 LEAST SQUARE CURVE FITTING

We can find the trend curve by fitting a mathematical equation. The method is more precise and can be used even for forecasting. We can fit either a straight line or a curve to the given data. We fit a straight line $y_c = a + bx$, where a and b are constants. We determine the constants a and b so that the following conditions are fulfilled:

i) The sum of the deviations of all the values of *y* from their trend values is zero, when, deviations above the line are given positive sign and deviations below, negative i.e if $y_c$is the trend value obtained from the trend line, and y is the actual value in the data.

$$\sum(y - y_c) = 0$$

ii) The sum of the squares of the deviations is the least, i.e.,

$\sum(y - y_c)^2$ is minimum.

The method gets the name 'least square method' because of this second property. This is also called *line of best fit*. In a sense this line is like arithmetic mean since arithmetic mean is a single value possessing the above two properties.

Remark: When the value of x is given in terms of years or value of x is big then we apply the following technique to get the values of x:

1) If the value of n is odd then we take the value of central most observation i.e value of $\frac{n+1}{2}$th observation as 0 and we add 1 as we move downward and we subtract 1 as we move upwards.

2) If the value if n is even then we take the value of $\frac{n}{2}$ th observation as -1 and we add 2 as we move downward and subtract 2 as we move upwards.

12.7.1 Straight Line:

Suppose o two variables *x* and *y*, *n* pairs of observations, say $(x_1, y_1), (x_2, y_2), \cdots\cdots, (x_n, y_n)$ are available and we want to fit a linear curve of the form $y = a + bx$ to the data. Then according to the least squares method we have to find *a* and *b* such that $\sum(y - y_i)^2$ is minimum where $y_i = a + bx_i$.

Thus, we have to minimize $D = \sum(y - a - bx)$. The first order conditions are $\frac{\delta D}{\delta a} = 0 \; and \; \frac{\delta D}{\delta b} = 0$

which are known as *normal equations* and they are given by

$$\sum y = an + b\sum x \; and \; \sum xy \; = a\sum x + b \sum x^2$$

It can be easily verified that the second order condition for minima are satisfied and therefore the best choice of *a* and *b* can be made by finding their values satisfying the normal equations. This equations will be used to fit a linear curve to the given data.

Eg: 1) Fit a straight line of the form y = a + bx using least square method

x	0	1	2	3	4
y	1	2.9	4.8	6.7	8.6

Solution: n = 5

	x	y	xy	x^2
	0	1	0	0
	1	2.9	2.9	1
	2	4.8	9.6	4
	3	6.7	20.1	9
	4	8.6	34.4	16
total	10	24.0	67.0	30

The straight line equation is given by $y = a + b$x. To find values of a and b we solve the following equations

$\sum y = an + b\sum x, \sum xy = a\sum x + b\sum x^2$ --- (1)

by substituting the values in eq (1) , we get

$$24 = 5a + 10b - - - -(2)$$

$$67 = 10a + 30b - - - -(3)$$

multiplying eq (2) by 2 we get $48 = 10a + 20b - - - -(4)$

subtracting eq (4) by eq (3), we get $19 = 10b \Rightarrow b = 1.9$

substituting $b = 1.9$ in eq (2) gives $a = 1$

$\therefore y = 1 + 1.9x$

is the required straight line.

Eg: 2) Fit a straight line of the form $y = a_0 + a_1x$

x	1	2	3	4	6	8
y	2.4	3.1	3.5	4.2	5	6

Solution: n = 6

	x	y	xy	x^2
	1	2.4	2.4	1
	2	3.1	6.2	4
	3	3.5	10.5	9
	4	4.2	16.8	16
	6	5	30	36
	8	6	48	64
total	24	24.2	113.9	130

The straight line equation is given by $y = a_0 + a_1$x. To find values of a_0 and a_1 we solve the following equations

$\sum y = a_0n + a_1\sum x, \sum xy = a_0\sum x + a_1\sum x^2$ - - - (1)

by substituting the values in eq (1) , we get

$$24.2 = 6a_0 + 24a_1 \; - - - (2)$$

$$113.9 = 24a_0 + 130a_1 \; - - - (3)$$

multiplying eq (2) by 4, we get $96.8 = 24a_0 + 96a_1 - - - (4)$

subtracting eq (4) from eq (3), we get $17.1 = 34a_1 \Rightarrow a_1 = 0.5029$

subs $a_1 = 0.5029$ in (2) gives $a_0 = 2.0217$

$$\therefore \; y = 2.0217 + 0.5029x$$

is the required straight line.

Eg: 3) Fit a straight line using least square method and estimate the exchange rate for the year 1993-94 and 1984-85.

year	1985-86	1986-87	1987-88	1988-89	1989-90	1990-91	1991-92
Exchange rate	12.24	12.78	12.97	14.48	16.65	17.94	24.47

Solution: n = 7

The straight line equation we take is $y = a + b$x. To find values of a and b we solve the following equations

$\sum y = an + b\sum x, \sum xy = a\sum x + b\sum x^2$ --- (1)

	year	ex. rate(y)	x	xy	x^2
	1985-86	12.24	-3	-36.72	9
	1986-87	12.78	-2	-25.56	4
	1987-88	12.97	-1	-12.97	1
	1988-89	14.48	0	0	0
	1989-90	16.65	1	16.65	1
	1990-91	17.94	2	35.88	4
	1991-92	24.47	3	73.41	9
total		111.53	0	50.69	28

by substituting the values in eq (1), we get

$$111.53 = 7a - - - -(2)$$

$$50.69 = 28b - - - -(3)$$

from eq (2), we get $a = \frac{111.53}{7} = 15.9328$

from eq (3), we get $b = \frac{50.69}{28} = 1.8104$

$$\therefore y = 15.9328 + 1.8104x \ - - - -(4)$$

is the required straight line.

To the get exchange rate for the year 1993-94 , we subs x = 5 in eq (4)

$$\therefore y = 15.9328 + 1.8104(5) = 24.9848$$

To the get the exchange rate for the year 1984-85, we subs x = -4 in eq (4)

$$\therefore y = 15.9328 + 1.8104(-4) = 8.6912$$

12.7.2 Parabola:

Suppose we want to fit a second degree curve called parabola of the form $y = a + bx + cx^2$ to n pairs of observations $(x_1, y_1), (x_2, y_2), \cdots\cdots, (x_n, y_n)$. Then according to the least squares method we have to minimize $D = \sum(y - a - bx - cx^2)^2$ with respect to a, b and c. The first order conditions are given by

$$\frac{\delta D}{\delta a} = 0, \frac{\delta D}{\delta b} = 0 \ and \ \frac{\delta D}{\delta c} = 0$$

Applying these we get,

$$\sum y = na + b\sum x + c\sum x^2$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

Eg: 1) Fit a parabola of the form $y = a + bx + cx^2$ using least square method

x	0	1	2
y	1	6	17

Solution: n=3

	x	y	x^2	x^3	x^4	xy	x^2y
	0	1	0	0	0	0	0
	1	6	1	1	1	6	6
	2	17	4	8	16	34	68
total	3	24	5	9	17	40	74

The equation of the parabola is given by $y = a + bx + cx^2$. To find values of $a, b \text{ and } c$ we solve the following equations

$$\sum y = na + b\sum x + c\sum x^2, \qquad \sum xy = a\sum x + b\sum x^2 + c\sum x^3,$$
$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4 \quad ----(*)$$

subs values in eq (*), we get

$$24 = 3a + 3b + 5c \quad ----(1)$$
$$40 = 3a + 5b + 9c \quad ----(2)$$
$$74 = 5a + 9b + 17c \quad ----(3)$$

Subtracting eq (1) by (2), we get $16 = 2b + 4c \Rightarrow 8 = b + 2c ---(4)$

$mul\ eq\ (2)\ by\ 5\ , we\ get\ 200 = 15a + 25b + 45c ---(5)$

$mul\ eq\ (3) by\ 3, we\ get\ 222 = 15a + 27b + 51c ---(6)$

$subtractting\ eq\ (5)\ by\ (6), we\ get\ 22 = 2b + 6c$

$\Rightarrow 11 = b + 3c ---(7)$

$subtracting\ eq\ (4)\ by\ \ (7), we\ get\ 3 = c$

$sub\ c = 3\ in\ eq\ (4)\ we\ get\ b = 2$

$sub\ b = 2, c = 3\ in\ (1), we\ get\ a = 1$

$\therefore y = 1 + 2x + 3x^2$ is the required equation of the parabola.

Eg: 2) Fit a parabola of the form $y = a + bx + cx^2$

year	1989	1990	1991	1992	1993	1994	1995	1996
no.of students	15	17	20	25	30	31	30	32

Also find the no. of students for the year 2000 & 1987.

Solution: n = 8

	year	no. of students(y)	x	x^2	x^3	x^4	xy	x^2y
	1989	15	-7	49	-343	2401	-105	735
	1990	17	-5	25	-125	625	-85	425
	1991	20	-3	9	-27	81	-60	180
	1992	25	-1	1	-1	1	-25	25

	1993	30	1	1	1	1	30	30
	1994	31	3	9	27	81	93	279
	1995	30	5	25	125	625	150	750
	1996	32	7	49	343	2401	224	1568
total		200	0	168	0	6216	222	3992

The equation of the parabola is given by $y = a + bx + cx^2$. To find values of $a, b \text{ and } c$ we solve the following equations

$$\sum y = na + b\sum x + c\sum x^2, \qquad \sum xy = a\sum x + b\sum x^2 + c\sum x^3,$$
$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4 \quad ----(*)$$

subs values in eq (*), we get

$$200 = 8a + 168c \quad ----(1)$$
$$222 = 168b \quad ----(2)$$
$$3992 = 168a + 6216c \quad ----(3)$$

$from\ (2), we\ get\ b = \frac{222}{168} = 1.3214$

$mul\ (1) by\ 21\ , we\ get\ 4200 = 168a + 3528c \ ---(4)$

$sub\ (4) by\ (3), we\ get\ 208 = -2688c \Rightarrow c = -\frac{208}{2688} = -0.0774$

$sub\ c = -0.0774\ in\ (1), we\ get\ a = \frac{200 - 168(-0.0774)}{8} = 26.6254$

$\therefore y = 26.6254 + 1.3214x - 0.0774x^2 ----(4)$

is the required equation of the parabola.

To get the no. of students for the year 2000 we substitute x = 15 in (4)

$$\therefore y = 26.6254 + 1.3214(15) - 0.0774(15)^2 = 29.0314$$

To get the no. of students for the year 1987 we substitute x = -11 in (4)

$$\therefore y = 26.6254 + 1.3214(-11) - 0.0774(-11)^2 = 2.7246$$

Eg: 3) Fit a parabola of the form $y = a + bx + cx^2$

year	1985	1986	1987	1988	1989
milk (in 100 litres)	20	25	27	35	38

Find the production of milk for the year 1982 and 1995.

`Solution: n =5

	year	milk (y)	x	x^2	x^3	x^4	xy	x^2y
	1985	20	-2	4	-8	16	-40	80
	1986	25	-1	1	-1	1	-25	25
	1987	27	0	0	0	0	0	0

	1988	35	1	1	1	1	35	35
	1989	38	2	4	8	16	76	152
total		145	0	10	0	34	46	292

The equation of the parabola is given by $y = a + bx + cx^2$. To find values of $a, b \text{ and } c$ we solve the following equations

$$\sum y = na + b\sum x + c\sum x^2, \qquad \sum xy = a\sum x + b\sum x^2 + c\sum x^3,$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4 \quad ----(*)$$

subs values in eq (*), we get

$145 = 5a + 10c --(1)$

$46 = 10b --(2) \Rightarrow b = 4.6$

$292 = 10a + 34c --(3)$

multiply eq (1) by 2, we get $290 = 10\ a + 20c --(4)$

subtracting eq (4) from eq (3), we get $2 = 14c \Rightarrow c = \frac{2}{14} = 0.1428$

substituting c = 0.1428 in eq (1), we get $a = 28.7144$

$\therefore y = 28.7144 + 4.6x + \ 0.1428x^2 --- (5)$

is the required equation of the parabola.

To get the production of milk for the year 1995, we put x = 8 in (5)

$\therefore y = 28.7144 + 4.6(8) + 0.1428(8)^2 \ = 74.6536$

To get the production of milk for the year 1982, we put x = -5 in (5)

$\therefore y = \ 28.7144 + 4.6(-5) + 0.1428(-5)^2 \ = 9.2844$

12.7.3 Non-Linear Relationship:

Fitting of a power Curve:

Suppose we want to fit a curve whose equation is of the form $y = a \cdot x^b$ to variables x and y on which n paired observations are available. Then we first rewrite the equation of the curve as

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

which can be written in the linear form $Y = A + bX$ where Y = log y, A = log a and X = log x. Therefore this curve can be fitted by the method of least squares solving the following normal equations.

$$\sum Y = nA + b\sum X \ , \sum XY \ = A\sum X + b\sum X^2$$

Once we get the value of A, we can find a = antilog (A). The method of fitting is explained in the following example.

Eg: 1) Fit a curve of the form $y = a \cdot x^b$ for the following data

x	1	2	4	5	7	10
y	2.1	4.9	20.8	32.7	60	131

Solution: n = 6

We can write the given curve in the form of $Y = A + bX$ where Y = log y, A = log a, X = log x

	x	y	X = log x	Y = log y	XY	X^2
	1	2.1	0	0.3222	0	0
	2	4.9	0.3010	0.6902	0.2078	0.0906
	4	20.8	0.6021	1.3181	0.7936	0.3625
	5	32.7	0.6990	1.5142	1.0586	0.4886
	7	60	0.8451	1.7782	1.5028	0.7142
	10	131	1	2.1173	2.1173	1
Total			3.4472	7.7405	5.6801	2.6559

The given curve is of the form $y = ax^b$, therefore the normal equations are given by

$$\sum Y = nA + b\sum X \,, \sum XY = A\sum X + b\sum X^2 \quad ---(1)$$

substituting the values in eq (1), we get

$$7.7405 = 6A + 3.4472b \qquad ---(2)$$

$$5.6801 = 3.4472A + 2.6559b \qquad ----(3)$$

multiplying eq (1) by 3.4472 and eq (2) by 6, then subtracting the two eq., we get

$$4.0522b = 7.3975 \Rightarrow b = \frac{7.3975}{4.0522} = 1.8256$$

substituting b = 1.8256 in eq (1), we get a = 002412

Now a = antilog(A) = antilog(0.2412) = 1.7426

$\therefore y = 1.7426x^{1.8256}$

is the required equation.

Fitting of the curve $y = ae^{bx}$

We first try to convert the equation $y = ae^{bx}$ to a linear equation by applying logarithms on both sides with respect to base 10. By applying logarithms on both sides, we get

$$\log y = \log a + bx\log_{10} e$$

Let Y = log y, A = log a, $\log_{10} e = 0.4343b$. Then the equation of the curve reduces to a linear equation of the form

$$Y = A + Bx$$

By using the least square method the constants A and b can be found by solving the normal equations

$$\sum Y = nA + B\sum x, \qquad \sum xY = A\sum x + B\sum x^2$$

Once A and B are found, we can find a = antilog (A) and b = $\frac{B}{0.4343}$

Eg: 1) Fit a curve of the form $y = ae^{bx}$ for the following data

x	0	2	4	6	8
y	3	55	1095	22000	442000

Solution : n =5

	x	y	Y = log y	xY	x^2
	0	3	0.4771	0	0
	2	55	1.7404	3.4808	4
	4	1095	3.0394	12.1576	16
	6	22000	4.3424	26.0544	36
	8	442000	5.6454	45.1632	64
Total	20		15.2447	86.8560	120

The given equation is of the form $y = ae^{bx}$, therefore the normal equations are given by

$$\sum Y = nA + B\sum x, \qquad \sum xY = A\sum x + B\sum x^2 \qquad ---(1)$$

substituting the values in eq (1), we get

$$15.2447 = 5A + 20B \ ----(2)$$

$$86.8560 = 20A + 120B \ ----(3)$$

multiplying eq (1) by 4 and subtracting it from eq (2), we get

$$25.8772 = 40B \Rightarrow B = \frac{25.8772}{40} = 0.6469$$

substituting B = 0.6469 in eq (1), we get

$$15.2447 = 5A + 20(0.6469) \Rightarrow A = \frac{2.3067}{5} = 0.4613$$

Now a = antilog (A) = antilog(0.4613) = 2.892

b = $\frac{B}{0.4343} = \frac{0.6469}{0.4343} = 1.4895$

$\therefore y = 2.892e^{1.4895x}$

is the required equation.

12.8 REGRESSION

Often, on the basis of sample data, we wish to estimate the value of a variable Y corresponding to a given value of a variable *x*. This can be accomplished by estimating the value of *y* from a least-squares curve that fits the sample data. The resulting curve is called a regression curve of *y* on *x*, since *y* is estimated from *x*.

If we want to estimate the value of *x* from a given value of *y*, we would use a regression curve of *x* on *y*, which amounts to interchanging the variables in the scatter diagram so that *x* is the dependent variable and *y* is the independent variable. This is equivalent to replacing the vertical deviations in the definition of the least-squares curve with horizontal deviations.

In general, the regression line or curve of *y* on *x* is not the same as the regression line or curve of *x* on *y*.

12.9 APPLICATIONS TO TIME SERIES

If the independent variable *x* is time, the data show the values of *y* at various times. Data arranged according to time are called time series. The regression line or curve of *y* on *x* in this case is often called atrend line or trend curve and is often used for purposes of estimation, prediction, or forecasting.

12.10 PROBLEMS INVOLVING MORE THAN TWO VARIABLES

Problems involving more than two variables can be treated in a manner analogous to that for two variables. For example, there may be a relationship between the three variables X, Y, and Z that can be described by the equation

$$z = a + bx + cy$$

which is called a linear equation in variables x, y and z.

In a three dimensional rectangular coordinate system this equation represents a plane, and the actual sample points $(x_1, y_1, z_1), (x_2, y_2, z_2), \cdots\cdots, (x_N, y_N, z_N)$ may scatter not too far from this plane which we call an *approximating plane.*

By extension of the least square method, we can talk about a least square plane approximating the data. If we are estimating *z* from the given values of *x* and *y*, this would be called a regression plane of *z* on *x* and *y*. The normal equations corresponding to the least square plane are given by

$$\sum z = aN + b\sum x + c\sum y$$
$$\sum xz = a\sum x + b\sum x^2 + c\sum xy$$
$$\sum yz = a\sum y + b\sum xy + c\sum y^2$$

Problems involving the estimation of a variable from two or more variables are called problems of multiple regression which is not there in the syllabus.

12.11 SUMMARY

In this chapter we learnt about curve fitting, various equations of approximating curves, the straight line method. We also learnt about the

fitting of linear and non-linear curves using the technique of least square which minimizes the sum of the square of the error.

12.12 EXERCISES

1) Fit a straight line of the form y = ax+ b using least square method

x	0	1	2	3
y	2	5	8	11

2) Fit a straight line of the form y = a + bx

x	10	12	13	16	17	20	25
y	19	22	24	27	29	33	37

3) Fit a straight line of the form y = a + bx

x	1	3	5	7	9	11	13	15	17
y	10	15	20	27	31	35	30	35	40

Also estimate y when x = 21.

4) Fit a parabola of the form $y = a + bx + cx^2$ using least square method

x	0	1	2	3	4
y	1	0	3	10	21

5) Fit a parabola of the form $y = a + bx + cx^2$

x	0	1	2	3	4
y	1	1.8	1.3	2.5	6.3

6) Fit a parabola using least square method $y = ax + bx^2$

x	1	2	3	4	5
y	1.8	5.1	8.9	14.1	19.8

7) Fit a curve of the form $y = ax^b$ for the following data

x	1	2	3	4
y	0.7	0.86	0.97	1.06

8) Fit a curve of the form $y = ae^{bx}$ for the following

x	0	2	4
y	5.012	10	31.62

9) Fit a curve of the form $y = ae^{bx}$ for the following

x	1	2	3	4	5
y	1.230	2.042	3.162	3.981	5.624

10) The number y of bacteria per unit volume present in a culture after x hours is given in the following table. Fit a of the form $y = ab^x$ using least square method for the following data.

x	0	1	2	3	4	5	6
y	30	45	60	90	130	190	275

11) Fit a curve of the form $y = ab^x$ for the following

x	1	2	3
y	8.3	15.4	33.1

12) The population of a state at ten yearly intervals is given below. Fit a curve of the form $y = ab^x$ using least square method and also estimate the population for the year 1961.

year	1881	1891	1901	1911	1921	1931	1941	1951
population in millions	3.9	5.3	7.3	9.6	12.9	17.1	23.2	30.5

13) Fit a straight line of the form y = a + bx for the following data

x	5	4	3	2	1
y	1	2	3	4	5

14) Fit a straight line of the form y = a + bx for the following data

x	3	5	7	9	11
y	2.3	2.6	2.8	3.2	3.5

15) Fit a straight line to the following data on production

Year	1996	1997	1998	1999	2000
Production	40	50	62	58	60

16) Fit a straight line to the following data on profit

Year	1992	1993	1994	1995	1996	1997	1998	1999
Profit	38	40	65	72	69	60	87	95

17) Fit a second degree parabolic equation of the form $y = a + bx + cx^2$

x	12	10	8	6	4	2
y	6	5	4	3	2	1

12.14 SOLUTION TO EXERCISES

Q. No.	Solution	Q. No.	Solution
1	2, 3	2	a = 7.7044, b = 1.213
3	a=11.43, b=1.73	4	1, -3, 2
5	1.42, -1.07, 0.55	6	a= 1.44 , b = 0.51
7	$y = 4.642e^{0.46x}$	8	$y = 4.642e^{0.46x}$
9	$y = 0.7762e^{0.472x}$	10	$y = 30(1.306)^{x}$
11	$y = 8.099(1.997)^{x}$	12	$y = 11.074(1.0298)^{x-1916}$
13	6, -1	14	1.83, 0.15
15	54, 4.8	16	62.0833, 7.3333
17	0, 0.5, 0		

12.15 LOGARITHM TABLES

LOGARITHMS

	0	1	2	3	4	5	6	7	8	9	Mean Difference 1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374	4	8	12	17	21	25	29	33	37
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	4	8	11	15	19	23	26	30	34
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	3	7	10	14	17	21	24	28	31
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	3	6	10	13	16	19	23	26	29
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3	6	9	12	15	18	21	24	27
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3	6	8	11	14	17	20	22	25
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3	5	8	11	13	16	18	21	24
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	2	5	7	10	12	15	17	20	22
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	2	5	7	9	12	14	16	19	21
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	2	4	7	9	11	13	16	18	20
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	2	4	6	8	11	13	15	17	19
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	6	8	10	12	14	16	18
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598	2	4	6	8	10	12	14	15	17
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784	2	4	6	7	9	11	13	15	17
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962	2	4	5	7	9	11	12	14	16
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	2	3	5	7	9	10	12	14	15
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298	2	3	5	7	8	10	11	13	15
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456	2	3	5	6	8	9	11	13	14
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	2	3	5	6	8	9	11	12	14
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757	1	3	4	6	7	9	10	12	13
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900	1	3	4	6	7	9	10	11	13
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038	1	3	4	6	7	8	10	11	12
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172	1	3	4	5	7	8	9	11	12
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302	1	3	4	5	6	8	9	10	12
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428	1	3	4	5	6	8	9	10	11
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	1	2	4	5	6	7	9	10	11
36	5563	5575	5587	5566	5611	5623	5635	5647	5658	5670	1	2	4	5	6	7	8	10	11
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786	1	2	3	5	6	7	8	9	10
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	1	2	3	5	6	7	8	9	10
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	1	2	3	4	5	7	8	9	10
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117	1	2	3	4	5	6	8	9	10
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	1	2	3	4	5	6	7	8	9
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	1	2	3	4	5	6	7	8	9
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	1	2	3	4	5	6	7	8	9
44	6435	6345	6454	6464	6474	6484	6493	6503	6513	6522	1	2	3	4	5	6	7	8	9
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618	1	2	3	4	5	6	7	8	9
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	1	2	3	4	5	6	7	7	8
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	1	2	3	4	5	5	6	7	8
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	1	2	3	4	4	5	6	7	8
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	1	2	3	4	4	5	6	7	8
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067	1	2	3	3	4	5	6	7	8
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152	1	2	3	3	4	5	6	7	8
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235	1	2	2	3	4	5	6	7	7
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316	1	2	2	3	4	5	6	6	7
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	1	2	2	3	4	5	6	6	7
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

ANTILOGARITHMS

	0	1	2	3	4	5	6	7	8	9	Mean Difference								
											1	2	3	4	5	6	7	8	9
.00	1000	1002	1005	1007	1009	1012	1014	1016	1019	1021	0	0	1	1	1	1	2	2	2
.01	1023	1026	1028	1030	1033	1035	1038	1040	1042	1045	0	0	1	1	1	1	2	2	2
.02	1047	1050	1052	1054	1057	1059	1062	1064	1067	1069	0	0	1	1	1	1	2	2	2
.03	1072	1074	1076	1079	1081	1084	1086	1089	1091	1094	0	0	1	1	1	1	2	2	2
.04	1096	1099	1102	1104	1107	1109	1112	1114	1117	1119	0	1	1	1	1	2	2	2	2
.05	1122	1125	1127	1130	1132	1135	1138	1140	1143	1146	0	1	1	1	1	2	2	2	2
.06	1148	1151	1153	1156	1159	1161	1164	1167	1169	1172	0	1	1	1	1	2	2	2	2
.07	1175	1178	1180	1183	1186	1189	1191	1194	1197	1199	0	1	1	1	1	2	2	2	2
.08	1202	1205	1208	1211	1213	1216	1219	1222	1225	1227	0	1	1	1	1	2	2	2	3
.09	1230	1233	1236	1239	1242	1245	1247	1250	1253	1256	0	1	1	1	1	2	2	2	3
.10	1259	1262	1265	1268	1271	1274	1276	1279	1282	1285	0	1	1	1	1	2	2	2	3
.11	1288	1291	1294	1297	1300	1303	1306	1309	1312	1315	0	1	1	1	2	2	2	2	3
.12	1318	1321	1324	1327	1330	1334	1337	1340	1343	1346	0	1	1	1	2	2	2	2	3
.13	1349	1352	1355	1358	1361	1365	1368	1371	1374	1377	0	1	1	1	2	2	2	3	3
.14	1380	1384	1387	1390	1393	1396	1400	1403	1406	1409	0	1	1	1	2	2	2	3	3
.15	1413	1416	1419	1422	1426	1429	1432	1435	1439	1442	0	1	1	1	2	2	2	3	3
.16	1445	1449	1452	1455	1459	1462	1466	1469	1472	1476	0	1	1	1	2	2	2	3	3
.17	1479	1483	1486	1489	1493	1496	1500	1503	1507	1510	0	1	1	1	2	2	2	3	3
.18	1514	1517	1521	1524	1528	1531	1535	1538	1542	1545	0	1	1	1	2	2	2	3	3
.19	1549	1552	1556	1560	1563	1567	1570	1574	1578	1581	0	1	1	1	2	2	3	3	3
.20	1585	1289	1592	1596	1600	1603	1607	1611	1614	1618	0	1	1	1	2	2	3	3	3
.21	1622	1626	1629	1633	1637	1641	1644	1648	1652	1656	0	1	1	2	2	2	3	3	3
.22	1660	1663	1667	1671	1675	1679	1683	1687	1690	1694	0	1	1	2	2	2	3	3	3
.23	1698	1702	1706	1710	1714	1718	1722	1726	1730	1734	0	1	1	2	2	2	3	3	4
.24	1738	1742	1746	1750	1754	1758	1762	1766	1770	1774	0	1	1	2	2	2	3	3	4
.25	1778	1782	1786	1791	1795	1799	1803	1807	1811	1816	0	1	1	2	2	2	3	3	4
.26	1820	1824	1828	1832	1837	1841	1845	1849	1854	1858	0	1	1	2	2	3	3	3	4
.27	1862	1866	1871	1875	1879	1884	1888	1892	1897	1901	0	1	1	2	2	3	3	3	4
.28	1905	1910	1914	1919	1923	1928	1932	1936	1941	1945	0	1	1	2	2	3	3	4	4
.29	1950	1954	1959	1963	1968	1972	1977	1982	1986	1991	0	1	1	2	2	3	3	4	4
.30	1995	2000	2004	2009	2014	2018	2023	2028	2032	2037	0	1	1	2	2	3	3	4	4
.31	2042	2046	2051	2056	2061	2065	2070	2075	2080	2084	0	1	1	2	2	3	3	4	4
.32	2089	2094	2099	2104	2109	2113	2118	2123	2128	2133	0	1	1	2	2	3	3	4	4
.33	2138	2143	2148	2153	2158	2163	2168	2173	2178	2183	0	1	1	2	2	3	3	4	4
.34	2188	2193	2198	2203	2208	2213	2218	2223	2228	2234	1	1	2	2	3	3	4	4	5
.35	2239	2244	2249	2254	2259	2265	2270	2275	2280	2286	1	1	2	2	3	3	4	4	5
.36	2291	2296	2301	2307	2312	2317	2323	2328	2333	2339	1	1	2	2	3	3	4	4	5
.37	2344	2350	2355	2360	2366	2371	2377	2382	2388	2393	1	1	2	2	3	3	4	4	5
.38	2399	2404	2410	2415	2421	2427	2432	2438	2443	2449	1	1	2	2	3	3	4	4	5
.39	2455	2460	2466	2472	2477	2483	2489	2495	2500	2506	1	1	2	2	3	3	4	5	5
.40	2512	2518	2523	2529	2535	2541	2547	2553	2559	2564	1	1	2	2	3	4	4	5	5
.41	2570	2576	2582	2588	2594	2600	2606	2612	2618	2624	1	1	2	2	3	4	4	5	5
.42	2630	2636	2642	2649	2655	2661	2667	2673	2679	2685	1	1	2	2	3	4	4	5	6
.43	2692	2698	2704	2710	2716	2723	2729	2735	2742	2748	1	1	2	3	3	4	4	5	6
.44	2754	2761	2767	2773	2780	2786	2793	2799	2805	2812	1	1	2	3	3	4	4	5	6
.45	2818	2825	2831	2838	2844	2851	2858	2864	2871	2877	1	1	2	3	3	4	5	5	6
.46	2884	2891	2897	2904	2911	2917	2924	2931	2938	2944	1	1	2	3	3	4	5	5	6
.47	2951	2958	2965	2972	2979	2985	2992	2999	3006	3013	1	1	2	3	3	4	5	5	6
.48	3020	3027	3034	3041	3048	3055	3062	3069	3076	3083	1	1	2	3	4	4	5	6	6
.49	3090	3097	3105	3112	3119	3126	3133	3141	3148	3155	1	1	2	3	4	4	5	6	6
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

ANTILOGARITHMS

	0	1	2	3	4	5	6	7	8	9	Mean Difference								
											1	2	3	4	5	6	7	8	9
.50	3162	3170	3177	3184	3192	3199	3206	3214	3221	3228	1	1	2	3	4	4	5	6	7
.51	3236	3243	3251	3258	3266	3273	3281	3289	3296	3304	1	2	2	3	4	5	5	6	7
.52	3311	3319	3327	3334	3342	3350	3357	3365	3373	3381	1	2	2	3	4	5	5	6	7
.53	3388	3396	3404	3412	3420	3428	3436	3443	3451	3459	1	2	2	3	4	5	6	6	7
.54	3467	3475	3483	3491	3499	3508	3516	3524	3532	3540	1	2	2	3	4	5	6	6	7
.55	3548	3556	3565	3573	3581	3589	3597	3606	3614	3622	1	2	2	3	4	5	6	7	7
.56	3631	3639	3648	3656	3664	3673	3681	3690	3698	3707	1	2	3	3	4	5	6	7	8
.57	3715	3724	3733	3741	3750	3758	3767	3776	3784	3793	1	2	3	3	4	5	6	7	8
.58	3802	3811	3819	3828	3837	3846	3855	3864	3873	3882	1	2	3	4	4	5	6	7	8
.59	3890	3899	3908	3917	3926	3936	3945	3954	3963	3972	1	2	3	4	5	5	6	7	8
.60	3981	3990	3999	4009	4018	4027	4036	4046	4055	4064	1	2	3	4	5	6	6	7	8
.61	4074	4083	4093	4102	4111	4121	4130	4140	4150	4159	1	2	3	4	5	6	7	8	9
.62	4169	4178	4188	4198	4207	4217	4227	4236	4246	4256	1	2	3	4	5	6	7	8	9
.63	4266	4276	4285	4295	4305	4315	4325	4335	4345	4355	1	2	3	4	5	6	7	8	9
.64	4365	4375	4385	4395	4406	4416	4426	4436	4446	4457	1	2	3	4	5	6	7	8	9
.65	4467	4477	4487	4498	4508	4519	4529	4539	4550	4560	1	2	3	4	5	6	7	8	9
.66	4571	4581	4592	4603	4613	4624	4634	4645	4656	4667	1	2	3	4	5	6	7	9	10
.67	4677	4688	4699	4710	4721	4732	4742	4753	4764	4775	1	2	3	4	5	7	8	9	10
.68	4786	4797	4808	4819	4831	4842	4853	4864	4875	4887	1	2	3	4	6	7	8	9	10
.69	4898	4909	4920	4932	4943	4955	4966	4977	4989	5000	1	2	3	5	6	7	8	9	10
.70	5012	5023	5035	5047	5058	5070	5082	5093	5105	5117	1	2	4	5	6	7	8	9	11
.71	5129	5140	5152	5164	5176	5188	5200	5212	5224	5236	1	2	4	5	6	7	8	10	11
.72	5248	5260	5272	5284	5297	5309	5321	5333	5346	5358	1	2	4	5	6	7	9	10	11
.73	5370	5383	5395	5408	5420	5433	5445	5458	5470	5483	1	3	4	5	6	8	9	10	11
.74	5495	5508	5521	5534	5546	5559	5572	5585	5598	5610	1	3	4	5	6	8	9	10	12
.75	5623	5636	5649	5662	5675	5689	5702	5715	5728	5741	1	3	4	5	7	8	9	10	12
.76	5754	5768	5781	5794	5808	5821	5834	5848	5861	5875	1	3	4	5	7	8	9	11	12
.77	5888	5902	5916	5929	5943	5957	5970	5984	5998	6012	1	3	4	5	7	8	10	11	12
.78	6026	6039	6053	6067	6081	6095	6109	6124	6138	6152	1	3	4	6	7	8	10	11	13
.79	6166	6180	6194	6209	6223	6237	6252	6266	6281	6295	1	3	4	6	7	9	10	11	13
.80	6310	6324	6339	6353	6368	6383	6397	6412	6427	6442	1	3	4	6	7	9	10	12	13
.81	6457	6471	6486	6501	6516	6531	6546	6561	6577	6592	2	3	5	6	8	9	11	12	14
.82	6607	6622	6637	6653	6668	6683	6699	6715	6730	6745	2	3	5	6	8	9	11	12	14
.83	6761	6776	6792	6808	6823	6839	6855	6871	6887	6902	2	3	5	6	8	9	11	13	14
.84	6918	6934	6950	6966	6982	6998	7015	7031	7047	7063	2	3	5	6	8	10	11	13	15
.85	7079	7096	7112	7129	7145	7161	7178	7194	7211	7228	2	3	5	7	8	10	12	13	15
.86	7244	7261	7278	7295	7311	7328	7345	7362	7379	7396	2	3	5	7	8	10	12	13	15
.87	7413	7430	7447	7464	7482	7499	7516	7534	7551	7568	2	3	5	7	9	10	12	14	16
.88	7586	7603	7621	7638	7656	7674	7691	7709	7727	7745	2	4	5	7	9	11	12	14	16
.89	7762	7780	7798	7816	7834	7852	7870	7889	7907	7925	2	4	5	7	9	11	12	14	16
.90	7943	7962	7980	7998	8017	8035	8054	8072	8091	8110	2	4	6	7	9	11	13	15	17
.91	8128	8147	8166	8185	8204	8222	8241	8260	8279	8299	2	4	6	8	9	11	13	15	17
.92	8318	8337	8356	8375	8395	8414	8433	8453	8472	8492	2	4	6	8	10	12	14	15	17
.93	8511	8531	8551	8570	8590	8610	8630	8650	8670	8690	2	4	6	8	10	12	14	16	18
.94	8710	8730	8750	8770	8790	8810	8831	8851	8872	8892	2	4	6	8	10	12	14	16	18
.95	8913	8933	8954	8974	8995	9016	9036	9057	9078	9099	2	4	6	8	10	12	15	17	19
.96	9120	9141	9162	9183	9204	9226	9247	9268	9290	9311	2	4	6	8	11	13	15	17	19
.97	9333	9354	9376	9397	9419	9441	9462	9484	9506	9528	2	4	7	9	11	13	15	17	20
.98	9550	9572	9594	9616	9638	9661	9683	9705	9727	9750	2	4	7	9	11	13	16	18	20
.99	9772	9795	9817	9840	9863	9886	9908	9931	9954	9977	2	5	7	9	11	14	16	18	20
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

Note: The logarithmic tables are directly taken from internet and inserted in picture format.

12.13 REFERENCE FOR FURTHER READING

Following books are recommended for further reading:

- Statistics by Murray R, Spiegel, Larry J. Stephens, Mcgraw Hill International Publisher, 4th edition
- Fundamental of Mathematical Statistics by S. C. Gupta and V. K. Kapoor, Sultan Chand and Sons publisher, 11th edition
- Descriptive Statistics by R. J. Shah

*****

13

CORRELATION THEORY

Unit Structure

13.0 OBJECTIVES

This Chapter would make you understand about the following concepts about correlation and regression:

- Scatter Diagram
- Linear Correlation
- Least square regression
- Explained and unexplained variation
- Regression Lines
- Correlation of Time series and Attributes

13.1 INTRODUCTION

By now we know how to find averages and dispersion of a distribution. Involving one variable. These measures give a complete idea about the structure of a distribution.

Sometimes it is necessary to know the relationship between two variables.For instance, a businessman would like to know the effect of production on price. If the two are related, he would like to know the

nature of the relationship and to use that knowledge to his benefit. If we get heights and weights of a group of students we can see that in general taller students will be heavier, though there will be some exceptions. As a man's income increases he spends more. That is, there will be some relation between income and expenditure of any person. For such data we would like to find answer for the following questions:

i) Are the two variables related?

ii) If they are related, how?

iii) To what extent they are related?

Correlation helps to find out answers to these questions.

If two variables vary together in the same direction or in opposite directions, they are said to be correlated. That is if as X increases, Y increases consistently, we say that X and Y are positively relatedi.e the variables are directly related with each other. In this case the values of X and Y for a particular individual have roughly the same relative position among their respective distributions, i.e if X is far above mean of X, then corresponding Y will be above mean of Y. If X is near to mean of X then Y will be near to mean of Y. We know that if the weight of the parents is above average then the son/daughter also will be having more weight.

There are some variables which are negatively correlated where, as X increases, Y decreases and as X decreases, Y increases i.e the variables are inversely proportional to each other, e.g Price increases as the supply decreases, that is, as the commodity becomes scare, the price increases.

If the change in one variable is proportional to the change in the other, the two variables are said to perfectly correlated.

If the two variables are not related to each other, we say the two variables have zero correlation, e.g length of the hair of an individual and the I.Q level of the same individual.

13.2 CORRELATION

The various method of finding whether two given variables are realted or not are:

- Scatter diagram
- Coefficient of Correlation

13.2.1 Scatter Diagram:

Scatter Diagram can be obtained by simply plotting the points on a graph where the two variables, say x and y are taken along x – axis and y – axis respectively.

Eg: A manager of a firm may want to appoint salesman for promoting his sales. When he gets a number of applications, he conducts an aptitude test and selects on the basis of their results. But after employing them,

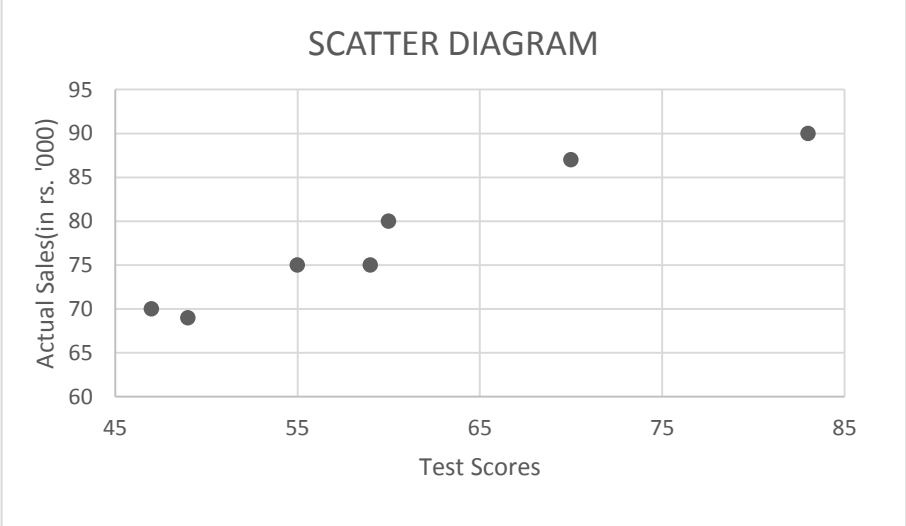
he wants to know whether there is any relation between the actual sales and the marks in the aptitude test.

The data for seven salesman are as follows:

Salesman	1	2	3	4	5	6	7
Aptitude test score	47	49	60	55	59	70	83
Actual sales in ‘000 in Rs.	70	69	80	75	75	87	90

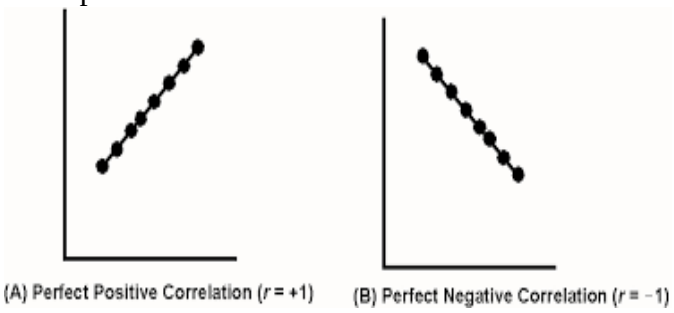
Solution:

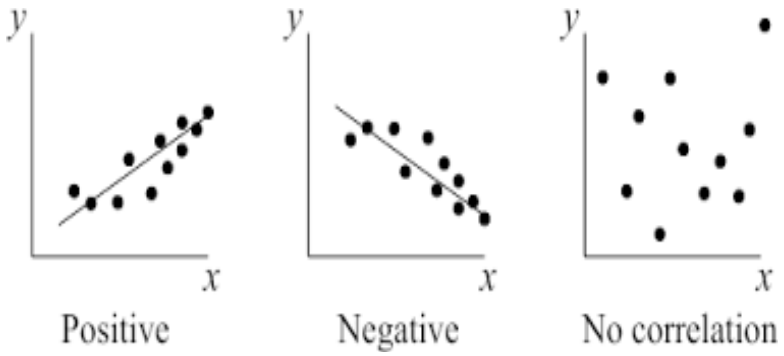
Here the test scores are plotted on the X –axis and actual sales on the Y –axis.



Here we can see that as X increases, Y also increases therefore the two variables are positively related. The businessman knows by this that it is useful to select salesman on the basis of the aptitude test conducted in this case.

Various possible patterns are shown here below





Types of Correlation:

Perfect Positive Correlation: When all plotted values lie exactly on a straight line and the line runs from lower left to upper right corner it is called perfect positive correlation. Here r = 1

Perfect Negative Correlation: When all plotted values lie exactly on a straight line and the line runs from upper left to lower right corner it is called perfect negative correlation. Here r = -1

Positive Correlation: If the value of two variables deviate in the same direction. It is known as Positive correlation or direct correlation.

Negative Correlation: If the value of two variables deviate in the opposite direction. It is also known as the Negative correlation or inverse correlation.

No Correlation: The variables are independent i.e there is no relation between the two variables i.e r = 0.

13.2.2 Coefficient of Correlation:

The things discussed in the previous section gives us the direction of existence of correlation. But we also require to find exact numerical measurement for the degree or extent of correlation. It is useful to have a numerical measure, which is independent of units of the original data, so that the two variables can be compared. For this we calculate the coefficient of correlation. It is denoted by r.

Definition: The coefficient of Correlation denoted by r and name after **Karl Pearson** is defined as

$$r = r_{x,y} = \frac{\sum\big((x - \bar{x})(y - \bar{y})\big)}{N\sigma_x\sigma_y}$$

$$\sigma_x = \sqrt{\left(\frac{\sum(x-\bar{x})^2}{N}\right)}, \sigma_y = \sqrt{\left(\frac{\sum y - \bar{y}}{N}\right)}$$

where there are N pairs.

This is also called **Product Moment Coefficient of Correlation.**

Covariance of x and y is defined as $cov(x,y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{N}$

$$\therefore r = \frac{cov(x,y)}{\sigma_x\sigma_y}$$

The formula of r can be simplified as $r = \frac{\sum xy - N\bar{x}\bar{y}}{\sqrt{\sum x^2 - N\bar{x}^2}\sqrt{\sum y^2 - N\bar{y}^2}}$

$$= \frac{\sum xy - \frac{\sum x \sum y}{N}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{N}}\sqrt{\sum y^2 - \frac{(\sum y)^2}{N}}}$$

Properties of Correlation:

1) $-1 \leq r \leq 1$

2) $r = 1, perfect + vecorrelation$

3) $r = -1, perfect - vecorrelation$

4) If 0 < r < 1, the correlation is positive

5) If -1 < r < 0, the correlation is negative

6) r = 0, no correlation

7) r is a pure number and is not affected by change of origin and scale in magnitude.

$i.eifu = \frac{x-a}{b}$, $v = \frac{y-c}{d}$ $thenr_{xy} = \frac{bd}{|b||d|} r_{uv}$

a) If b and d are of same sign then $r_{xy} = r_{uv}$

b) If b and d are of opposite signs then $r_{xy} = -r_{uv}$

8) If y = ax + b, $r_{xy} = 1\ if a > 0\ and r_{xy} = -1, if a < 0$

Eg 1: $r_{xy} = 0.6$, find r_{uv}if 2u – 3x + 4 = 0, 4v – 16y + 11 = 0

solution: $2u - 3x + 4 = 0 \Rightarrow u = \frac{3x-4}{2} = \frac{x-\frac{4}{3}}{\frac{2}{3}} \Rightarrow b = \frac{2}{3}$

$$4v - 16y + 11 = 0 \Rightarrow v = \frac{16y - 11}{4} = \frac{y - \frac{11}{16}}{\frac{4}{16}} \Rightarrow d = \frac{4}{16}$$

$$b\ and\ d\ are\ of\ same\ signs \therefore r_{uv} = r_{xy} = 0.6$$

Eg 2: $r_{xy} = 0.6$, find r_{uv}if 2u + 3x + 4 = 0, 4v – 16y + 11 = 0

solution: $2u + 3x + 4 = 0 \Rightarrow u = -\frac{x+\frac{4}{3}}{\frac{2}{3}} \Rightarrow b = -\frac{2}{3}$,

$$4v - 16y + 11 = 0 \Rightarrow v = \frac{y - \frac{11}{16}}{\frac{4}{16}} \Rightarrow d = \frac{4}{16}$$

$$b\ and\ d\ are\ of\ opposite\ signs \therefore r_{uv} = -r_{xy} = -0.6$$

Eg 3: If 2x + y = 3 what is the value of $r_{xy.}$

solution: $2x + y = 3 \Rightarrow y = -2x + 3 \Rightarrow a = -2 < 0 \Rightarrow r_{xy} = -1$

Eg4: Calculate the karlpearson's coefficient of correlation given $cov(x, y) = -15, \sigma_x = 5, \sigma_y = 4.$

solution: $r = \frac{cov(x,y)}{\sigma_x \sigma_y} = -\frac{15}{5\times4} = -\frac{15}{20} = -\frac{3}{4} = -0.75$

Eg 5: Calculate the coefficient of correlation for the following:

x	-2	-1	0	1	2
y	4	1	0	1	4

solution: $N = 5, \sum x = 0, \sum y = 10, \sum(x - \bar{x})(y - \bar{y}) = 0, \sigma_x = \sqrt{\frac{10}{5}} = \sqrt{2}, \sigma_y = \sqrt{\frac{14}{5}} = \sqrt{2.8}$

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{N\sigma_x\sigma_y} = \frac{0}{5 \cdot \sqrt{2} \cdot \sqrt{2.8}} = 0$$

Remark: If there is no correlation between the two variables, r = 0 but the converse is not true. In this example the values are related by the equation $y = x^2$ but the observed value of coefficient of correlation is zero.

Eg 6: Calculate the coefficient of correlation for the following:

x	1	2	3	4	5	6	7	8	9	10
y	2	4	9	7	10	5	14	16	2	20

solution: N = 10

x	1	2	3	4	5	6	7	8	9	10	Total 55
y	2	4	9	7	10	5	14	16	2	20	89
xy	2	8	27	28	50	30	98	128	18	200	589
x^2	1	4	9	16	25	36	49	64	81	100	385
y^2	4	16	81	49	100	25	196	256	4	400	1131

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}\sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}} = \frac{\left(589 - \frac{55\times89}{10}\right)}{\sqrt{385 - \frac{(55)^2}{10}}\sqrt{1131 - \frac{(89)^2}{10}}}$$

$$r = \frac{(589 - 489.5)}{\sqrt{82.5} \times \sqrt{338.9}} = \frac{99.5}{(9.08)(18.41)} = 0.6$$

Eg 7: Find the coefficient of correlation for the following:

x	60	50	45	47	53	70	75	57	73	48
y	30	29	29	28	29	35	40	32	35	28

solution: N = 10, we take a = 60, c = 35(assumed mean)

											Total
x	60	50	45	47	53	70	75	57	73	48	
y	30	29	29	28	29	35	40	32	35	28	
u = x – 60	0	-10	-15	-13	-7	10	15	-3	13	-12	-22
v = y – 35	-5	-6	-6	-7	-6	0	5	-3	0	-7	-35
uv	0	60	90	91	42	0	75	9	0	84	451
u^2	0	100	125	169	49	100	225	9	169	144	1190
v^2	25	36	36	49	36	0	25	9	0	49	265

here b and d are same i.e 1∴ $r_{xy} = r_{uv} = \frac{\sum uv - \frac{\sum u \sum v}{N}}{\sqrt{\sum u^2 - \frac{(\sum u)^2}{N}}\sqrt{\sum v^2 - \frac{\sum v^2}{N}}}$

$$= \frac{451 - \frac{(-22)(-35)}{10}}{\sqrt{1190 - \frac{(-22)^2}{10}}\sqrt{265 - \frac{(-35)^2}{10}}} = \frac{374}{\sqrt{1141.6}\sqrt{142.5}} = 0.927$$

Eg 8: Calculate the coefficient of correlation for the following:

x	5	10	10	15	15	20	25	30
y	15	17	17	19	21	21	19	17

solution: N = 8 , we take a = 20, c = 19, b = 5, d = 2

									Total
x	5	10	10	15	15	20	25	30	
y	15	17	17	19	21	21	19	17	
$u = \frac{x-20}{5}$	-3	-2	-2	-1	-1	0	1	2	-6
$v = \frac{y-19}{2}$	-2	-1	-1	0	1	1	0	-1	-3
uv	6	2	2	0	-1	0	0	-2	7
u^2	9	4	4	1	1	0	1	4	24
v^2	4	1	1	0	1	1	0	1	9

Since b and d are of same signs, $r_{xy} = r_{uv} = \frac{\sum uv - \frac{\sum u \sum v}{N}}{\sqrt{\sum u^2 - \frac{(\sum u)^2}{N}}\sqrt{\sum v^2 - \frac{\sum v^2}{N}}}$

$$r_{xy} = r_{uv} = \frac{7 - \frac{(-6)(-3)}{8}}{\sqrt{24 - \frac{(-6)^2}{8}}\sqrt{9 - \frac{(-3)^2}{8}}} = \frac{7 - 2.25}{\sqrt{19.5}\sqrt{7.87}} = 0.38$$

Eg 9: From the data given below find the number of items n, r = 0.5,$\sum(x-\bar{x})(y-\bar{y}) = 120,\ \sigma_y = 8, \sum(x-\bar{x})^2 = 90.$

solution:$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{n\sigma_x\sigma_y}$, $\sigma_x = \sqrt{\left(\frac{\sum(x-\bar{x})^2}{n}\right)}$, $\sigma_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n}}$

$$0.5 = \frac{120}{n \times \sqrt{\left(\frac{90}{n}\right) \times 8}} \Rightarrow 0.5 = \frac{120}{\sqrt{n^2 \times \frac{90}{n} \times 8}}$$

$$0.5 = \frac{120}{\sqrt{90n} \times 8} \Rightarrow \sqrt{90n} = \frac{120}{0.5 \times 8} = 30$$

$$squaring\ we\ get, 90n = 900 \Rightarrow n = 10$$

Eg 10: Calculate the coefficient of correlation between x and y.

	x	y
no.of observations	15	15
arithmetic mean	25	18
standard deviation	3.01	3.03
sum of squares of the deviation from arithmetic mean	136	138
$\sum(x-\bar{x})(y-\bar{y}) = 122$		

solution: $r = \frac{\sum(x-\bar{x})(y-\bar{y})}{n\sigma_x\sigma_y}$, $\sigma_x = \sqrt{\left(\frac{\sum(x-\bar{x})^2}{n}\right)}$, $\sigma_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n}}$

$$r = \frac{122}{15 \times \sqrt{\frac{136}{15}} \times \sqrt{\frac{138}{15}}} = \frac{122}{15 \times 3.01 \times 3.03} = 0.89$$

Eg 11:n = 20, r = 0.3, $\bar{x} = 15, \bar{y} = 20, \sigma_x = 4, \sigma_y = 5$. One pair (27, 30) was wrongly taken as (17, 35). Find corrected value of r.

solution: $\bar{x} = 15, \frac{\sum x}{n} = 15 \Rightarrow \sum x = 15 \times 20 = 300$

corrected value $\bar{x} = \frac{300-17+27}{20} = \frac{310}{20} = 15.5$

$\bar{y} = 20, \frac{\sum y}{n} = 20 \Rightarrow \sum y = 20 \times 20 = 400$

corrected value $\bar{y} = \frac{400-35+30}{20} = \frac{395}{20} = 19.75$

$\sigma_x = 4 \Rightarrow (\sigma_x)^2 = 16 \Rightarrow \frac{\sum x^2}{n} - (\bar{x})^2 = 16$

$\sum x^2 = (16 + 225) \times 20 = 4820$

corrected value of $\sum x^2 = 4820 + 27 \times 27 - 17 \times 17 = 5260$

corrected value of $\sigma_x = \sqrt{\frac{5260}{20} - (15.5)^2} = \sqrt{22.75} = 4.77$

$\sigma_y = 5 \Rightarrow (\sigma_y)^2 = 25 \Rightarrow \frac{\sum y^2}{n} - (\bar{y})^2 = 25$

$\sum y^2 = (25 + 400) \times 20 = 8500$

corrected value of $\sum y^2 = 8500 + 30 \times 30 - 35 \times 35 = 8175$

corrected value of $\sigma_y = \sqrt{\frac{8500}{20} - (19.75)^2} = 5.91$

$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{n\sigma_x\sigma_y} \Rightarrow 0.3 = \frac{\sum(x - \bar{x})(y - \bar{y})}{20 \times 4 \times 5}$

$\Rightarrow \sum(x - \bar{x})(y - \bar{y}) = 120$

corrected value of

$$\sum(x - \bar{x})(y - \bar{y}) = 120 + (27 - 17.5)(30 - 19.75) - (17 - 15)(35 - 20)$$

$$= 120 + 9.5 \times 10.25 - (2) \times (15) = 187.375 = 187.38$$

$$corrected\ value\ of\ r = \frac{187.38}{20 \times 4.77 \times 5.91} = 0.33$$

13.2.3 Coefficient of Rank Correlation:

In certain types of characteristics it is not possible to get numerical measurements, but we can rank the individuals in order according to our own judgment, e.g. beauty, smartness. If two persons rank a given group of individuals and we have to find how far the two judges agree with each other, the technique of rank correlation can be used. In some cases though actual measurements are available we may be interested in only the ranks, that is the relative position of an individual in the group. Here also the rank correlation is used.

The formula for **Spearman's rank correlation coefficient** is

$R = 1 - \frac{6\sum d^2}{N(N^2 - 1)}$ where d = difference between the ranks of the same individual, N = number of individuals.

Remark: R follows the same property as r

Eg 1: The ranks according to judges in a beauty contest are

R_1	1	2	3	4	5	6
R_2	4	1	2	3	6	5

Find the coefficient of rank correlation.

Solution: N = 6

							Total
$R_1(d_1)$	1	2	3	4	5	6	
$R_2(d_2)$	4	1	2	3	6	5	
$d = d_1 - d_2$	-3	1	1	1	-1	1	
d^2	9	1	1	1	1	1	14

$$R = 1 - \frac{6\sum d^2}{N(N^2-1)} = 1 - 6 \times \frac{14}{6(36-1)} = 1 - \frac{14}{35} = \frac{21}{35} = \frac{3}{5}$$
$$= 0.6.$$

Eg 2: Find spearman's rank correlation coefficient between cost and sales for the following

cost	39	65	62	90	82	75	25	98	36	78
sales	47	53	58	86	62	68	60	91	51	84

Solution: N = 10

											Total
cost (X)	39	65	62	90	82	75	25	98	36	78	
sales (Y)	47	53	58	86	62	68	60	91	51	84	
d_1	8	6	7	2	3	5	10	1	9	4	
d_2	10	8	7	2	5	4	6	1	9	3	
$d = d_1 - d_2$	-2	-2	0	0	-2	1	4	0	0	1	
d^2	4	4	0	0	4	1	16	0	0	1	30

$$R = 1 - \frac{6\sum d^2}{N(N^2-1)} = 1 - 6 \times \frac{30}{10(100-1)} = 1 - \frac{18}{99} = \frac{81}{99} = \frac{9}{11}$$
$$= 0.8182$$

Coefficient of Rank correlation when ranks are repeated

In the above example, the ranks were different for all the but in some cases, two or more items may have the same numerical values and ranks should be the same for these values. Suppose we give ranks 1,2,3 and the next two values have to given same rank. In this case next two ranks are 4 and 5.These are to be distributed equally. Therefore both the individuals will get the rank $\frac{4+5}{2}$and the next one will get the rank as 6.

When there are groups getting the same rank, there is some adjustment in the formula also. If $m_1, m_2, m_3, \ldots$ denotes the number of times the same rank appear, the coefficient of Rank Correlation will be

$R = 1 - \frac{6\{\sum d^2 + CF\}}{N(N^2-1)}$, where$= \frac{1}{12}[(m_1^3 - m_1) + (m_2^3 - m_2) +$

$(m_3^2 - m_3) + \cdots\cdots]$, CF – Correction Factor

Eg 1: Marks of 10 students in two I.Q tests carried out in two successive months. Find the coefficient of correlation.

marks test 1	75	60	60	73	55	57	53	72	65	69
marks test 2	65	65	64	70	58	60	58	68	63	65

Solution: N = 10

											Total
marks test 1	75	60	60	73	55	57	53	72	65	69	
marks test 2	65	65	64	70	58	60	58	68	63	65	
d_1	1	6.5 **{6}**	6.5 **{7}**	2	9	8	10	3	5	4	
d_2	4 **{3}**	4 **{4}**	6	1	9.5 **{9}**	8	9.5 **{10}**	2	7	4 **{5}**	
$d = d_1 - d_2$	-3	2.5	0.5	1	-0.5	0	0.5	1	2	0	
d^2	9	6.25	0.25	1	0.25	0	0.25	1	4	0	22

$m_1 = 2$, {no. of times marks 60 is repeated in test 1}
$m_2 = 3$, {no. of times marks 65 is repeated in test 2}
$m_3 = 2$, {no. of times marks 58 is repeated in test 2}

$$CF = \frac{1}{12}\{(m_1^3 - m_1) + (m_2^3 - m_2) + (m_3^3 - m_3)\} = \frac{1}{12}\{6 + 24 + 6\}$$

$$= \frac{1}{12}\{36\} = 3$$

$$R = 1 - \frac{6\{\sum d^2 + CF\}}{n(n^2 - 1)} = 1 - \frac{6\{22 + 3\}}{10(99)} = 1 - 6 \times \frac{25}{990} = 1 - \frac{15}{99}$$

$$= \frac{84}{99} = 0.8485$$

Eg 2: The coefficient of Rank correlation for certain data is found to 0.6. If the sum of the squares of the differences is given to be 66, find the number of observations.

Solution: Given R = 0.6, $\sum d^2 = 66$, To find number of observations i.e to find N

$$R = 1 - \frac{6\sum d^2}{N(N^2 - 1)} \Rightarrow 0.6 = 1 - \frac{6 \times 66}{N(N^2 - 1)}$$

$$\Rightarrow \frac{6 \times 66}{N(N^2 - 1)} = 1 - 0.6$$

$$\Rightarrow \frac{6 \times 66}{N(N^2-1)} = 0.4 \Rightarrow N(N^2-1) = \frac{6 \times 66}{0.4}$$

$$\Rightarrow (N-1)N(N+1) = 990$$

$$\Rightarrow (N-1)N(N+1) = 9 \times 10 \times 11 \Rightarrow N = 10$$

13.3 THE LEAST SQUARE REGRESSION LINES

When we know that two given variables are correlated we try to establish some relation between the two so that we can estimate the value of one of the variables given the value of other e.g If we know that there is positive correlation between the heights and weights of a group of individuals, we can find an equation between the height and the weight. We can estimate the weight of an individual belonging to the same population given his height.

Correlation coefficient only determines whether the variables are related and if so, how strong is the relationship. But it is not useful for prediction. The equations used for prediction or estimation are known as **regression equations.**With the help of regression analysis we establish a model which expresses the functional relationship between the two variables. These are also known as estimation equations.

13.3.1 Regression Lines:

We can find two straight lines which will be useful for estimating Y when X is given. It is known as regression of Yon X. Here X is considered as independent variable. For estimating X when Y is given. It is known as regression of X on Y. Here Y is considered as independent variable. We fit a straight line for the set of points given in the bivariate data.

13.3.2 Least Square Method:

We know that a degree one equation represents a straight line. Therefore we take the equation of the regression line of Y on X as $Y = a + bX$where a and b are constants. The constant 'a' determines the point where the line cuts the Y – axis and constant 'b' determines the slope of the line. The method of *least squares*is used to determine the constants and we get a and b by solving the following two equations simultaneously which are known as *normal equations.*

$$\sum Y = Na + b\sum X\,, \qquad \sum XY = a\sum X + b\sum X^2$$

Similarly, we take the regression of X on Y as $X = c + dY$where c and d are constants which can be determined by solving the following normal equations.

$$\sum X = Nc + d\sum Y\,, \qquad \sum XY = c\sum Y + d\sum Y^2$$

Eg 1: Find the regression of Y on X and X on Y for the following data. Also estimate Y when X = 7 and estimate X when Y = 16

X	1	2	3	4	5
Y	10	12	15	14	15

Solution: N = 5

X	1	2	3	4	5	Total 15
Y	10	12	15	14	15	66
XY	10	24	45	56	75	210
X^2	1	4	9	16	25	55
Y^2	100	144	225	196	225	890

Normal Equation for regression of Y on X are given by

$$\sum y = Na + b\sum X \;\; i.e \;\; 66 = 5a + 15b$$

$$\sum XY = a\sum X + b\sum X^2 \;\; i.e \;\; 210 = 15a + 55b$$

Solving the two equations simultaneously we get a = 9.6 and b = 1.2
∴ The regression line of Y on X is given by $Y = 9.6 + 1.2X$ --- (1)
For estimating value of Y when X = 7, we put X = 7 in eq (1)
∴ The estimate will be Y = 9.6 + 1.2(7) = 18

Normal equations for regression of X on Y are given by

$$\sum X = Nc + d\sum Y \qquad i.e \;\; 15 = 5c + 66d$$

$$\sum XY = c\sum Y + d\sum Y^2 \qquad i.e \; 210 = 66c + 890d$$

Solving the two equations simultaneously we get c = -5.4255, d = 0.6383
∴ The regression line of X on Y is given by $X = -5.4255 + 0.6383Y$ --- (2)
For estimating value of X when Y = 16, we put Y = 16 in eq (2)
∴ The estimate will be X = -5.4255 + 0.6383(16) = 4.7873

13.3.3 Regression Lines and Regression Coefficients:

The regression of Y on X is given by

$$Y - \bar{Y} = b_{YX}(X - \bar{X}), b_{YX} = r\frac{\sigma_Y}{\sigma_X}$$

The regression of X on Y is given by

$$X - \bar{X} = b_{XY}(Y - \bar{Y}), b_{XY} = r\frac{\sigma_X}{\sigma_Y}$$

where $b_{XY}\ and\ b_{YX}$ are known as regression coefficients.

Properties:

1) $b_{YX} = r\frac{\sigma_Y}{\sigma_X}$, $b_{XY} = r\frac{\sigma_X}{\sigma_Y}$

$r^2 = b_{YX} \times b_{XY} \Rightarrow r = \pm\sqrt{b_{YX} \times b_{XY}}$

$r > 0\ if\ b_{XY} \& b_{YX} > 0$

$r < 0\ if\ b_{XY} \& b_{YX} < 0$

2) $if\ u = X - a, v = Y - b\ then\ b_{XY} = b_{uv} \& b_{YX} = b_{vu}$

3) $b_{YX} = r\frac{\sigma_Y}{\sigma_X} = \frac{cov(X,Y)}{\sigma_X^2} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sum X^2 - \frac{(\sum X)^2}{N}}$

4) $b_{xy} = r\frac{\sigma_X}{\sigma_Y} = \frac{cov(X,Y)}{\sigma_Y^2} = \frac{\left(\sum XY - \frac{\sum X \sum Y}{N}\right)}{\sum Y^2 - \frac{(\sum Y)^2}{N}}$

Eg: 1) The following data are given about the expenditure on clothes and expenditure on entertainment. Average expenditure on clothes Rs. 300, average expenditure on entertainment Rs. 100, S.D of expenditure on clothes Rs. 20, S.D of expenditure on entertainment Rs. 15, coefficient of correlation 0.78. Find the two regression equations.

Solution: Let the expenditure on clothes be denoted by x and the expenditure on entertainment be denoted by y.

Given $\bar{x} = 300,\ \bar{y} = 100, \sigma_x = 20, \sigma_y = 15, r = 0.78$

$\therefore b_{xy} = \frac{r\sigma_x}{\sigma_y} = \frac{0.78 \times 20}{15} = 0.78 \times \frac{4}{3} = 1.04$

$\& \ b_{yx} = \frac{r\sigma_y}{\sigma_x} = \frac{0.78 \times 15}{20} = 0.78 \times \frac{3}{4} = 0.585$

$Regression\ equation\ of\ y\ on\ x\ is\ given\ by$

$(y - \bar{y}) = b_{yx}(x - \bar{x}) \Rightarrow y - 100 = 0.585(x - 300)$

$-0.585x + y = 100 - 175.5 \Rightarrow y = 0.585x - 75.5$

$Regression\ equation\ of\ x\ on\ y\ is\ given\ by$

$(x - \bar{x}) = b_{xy}(y - \bar{y}) \Rightarrow x - 300 = 1.04(y - 100)$

$x = 1.04y + 300 - 104 \Rightarrow x = 1.04y + 196$

Eg 2: If $\sum x = 37, \sum y = 71, \sum xy = 563, \sum x^2 = 297, \sum y^2 = 1079, n = 5$. Find the two regression equations.

Solution: $\bar{x} = \frac{37}{5} = 7.4\ , \bar{y} = \frac{71}{5} = 14.2,\ b_{yx} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{\left(563 - \frac{37 \times 71}{5}\right)}{297 - \frac{(37)^2}{5}}$

$$b_{yx} = \frac{563 \times 5 - 37 \times 71}{297 \times 5 - 37 \times 37} = \frac{2815 - 2627}{1485 - 1369} = \frac{188}{116} = 1.62$$

$$b_{xy} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum y^2 - \frac{(\sum y)^2}{n}} = \frac{\left(563 - \frac{37 \times 71}{5}\right)}{1079 - \frac{(71)^2}{5}}$$

$$b_{xy} = \frac{563 \times 5 - 37 \times 71}{1079 \times 5 - 71 \times 71} = \frac{188}{354} = 0.53$$

$regression\ equation\ of\ y\ on\ x$

$y - \bar{y} = b_{yx}(x\ - \bar{x}) \Rightarrow y - 14.2 = 1.62 \times (x - 7.4)$

$y = 1.62x + 2.21$

$regression\ equation\ of\ x\ on\ y$

$x - \bar{x} = b_{xy}(y - \bar{y}) \Rightarrow x - 7.4 = 0.53 \times (y - 14.2)$

$x = 0.53y - 0.126$

Eg 3: Find the two regression lines for the following data

Price in Rs.	100	120	110	110	160	150
Demand(in units)	40	38	43	45	37	23

Also estimate the demand when price is 130 and the price when demand is 30 units.

Solution:N=6

							Total
Price in Rs. (X)	100	120	110	110	160	150	750
Demand(in units) (Y)	40	38	43	45	37	23	226
XY	4000	4560	4730	4950	5920	3450	27610
X^2	10000	14400	12100	12100	25600	22500	96700
Y^2	1600	1444	1849	2025	1369	529	8816

$$\bar{x} = \frac{\sum x}{n} = \frac{750}{6} = 125, \bar{y} = \frac{\sum y}{n} = \frac{226}{6} = 37.67$$

$$b_{yx} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{\left(27610 - \frac{750 \times 226}{6}\right)}{96700 - 750 \times \frac{750}{6}} = -\frac{3840}{17700} = -0.22$$

$$b_{xy} = \frac{\left(\sum xy - \frac{\sum x \sum y}{n}\right)}{\sum y^2 - \frac{(\sum y)^2}{n}} = -\frac{3840}{1820} = -2.11$$

$regression\ equation\ of\ y\ on\ x$

$y - \bar{y} = b_{yx}(x - \bar{x}) \Rightarrow y - 37.67 = -0.22(x - 125)$

$y = -0.22x + 65.17 \quad - - - \ -(1)$

$regression\ equation\ of\ x\ on\ y$

$x - \bar{x} = b_{xy}(y - \bar{y}) \Rightarrow x - 125 = -2.11(y - 37.67)$

$x = -2.11y + 204.48 \quad - - - - - (2)$

$demand = ?\, when\ price = 130, we\ put\ x = 130\ in\ (1)$

$y = -0.22 \times 130 + 65.17 = 36.57$

$price = ?\, when\ demand = 30, we\ put\ y = 30\ in\ (2)$

$x = -2.11 \times 30 + 204.48 = 141.18$

Eg 4: The two regression lines are given by $x + 2y = 5,\ 2x + 3y - 8 = 0, (\sigma_x)^2 = 12$. Find the values of $\bar{x}, \bar{y}, (\sigma_y)^2$.

Solution: We solve the two regression equations simultaneously to find $\bar{x}$ & $\bar{y}$

$x + 2y = 5 \ --(1)$

$2x + 3y - 8 = 0 \ -\ -(2)$

$Mul\ eq\ (1)\ by\ 2\ and\ subtract\ from\ eq\ (2)\ gives\ \bar{y} = 2$

$substitute\ y = 2\ in\ eq\ (1)\ gives\ \bar{x} = 1$,

$x + 2y = 5 \Rightarrow x = -2y + 5, b_{xy} = -2$

$2x + 3y - 8 = 0 \Rightarrow y = \left(-\frac{2}{3}\right)x + 4, \qquad b_{yx} = -\frac{2}{3}$

$b_{xy} \times b_{yx} = \frac{4}{3} > 1\ not\ possible$

$regression\ of\ y\ on\ x\ is\ given\ by\ x + 2y = 5\ and\ x\ on\ y\ is$

$given\ by\ 2x + 3y - 8 = 0$

$x + 2y = 5 \Rightarrow y = -\frac{x}{2} + \frac{5}{2}, b_{yx} = -\frac{1}{2}$

$2x + 3y = 8 \Rightarrow x = \left(-\frac{3}{2}\right)y + 4, \qquad b_{xy} = -\frac{3}{2}$

$r^2 = b_{yx} \times b_{xy} \Rightarrow r^2 = \frac{3}{4} = 0.75 \Rightarrow r = -\sqrt{\frac{3}{4}} = -0.87$

$r^2 = (b_{yx})^2 \times \frac{(\sigma_y)^2}{(\sigma_x)^2} \Rightarrow r^2 = \frac{1}{4} \times \frac{(\sigma_y)^2}{12} \Rightarrow \frac{3}{4} = \frac{1}{4} \times \frac{(\sigma_y)^2}{12}$

$\Rightarrow (\sigma_y)^2 = 0.75 \times 48 = 36$

13.4 STANDARD ERROR OF ESTIMATE

Ley y_1 be the value of y for given value of x, a measure of the scatter about the regression line y on x is given by

$$S_{y,x} = \sqrt{\frac{\sum(y - y_1)^2}{n}}$$

which is called the standard error of estimate of y on x.
Standard error of estimate of x on y

$$S_{x,y} = \sqrt{\frac{\sum(x - x_1)^2}{n}}$$

In general, $S_{y,x} \neq S_{x,y}$

Eg: 1) If the regression line of y on x is $y = a_0 + a_1x$. Show that

$S_{y,x}^2 = \frac{\sum y^2 - a_0\sum y - a_1\sum xy}{n}$.

solution: $y_1 = a_0 + a_1x$

$$S_{y,x}^2 = \frac{\sum(y - y_1)^2}{n} = \frac{\sum(y - y_1)^2}{n} = \frac{\sum(y^2 - 2yy_1 + y_1^2)}{n}$$

$$= \frac{\sum(y^2 - 2y(a_0 + a_1x) + (a_0 + a_1x)^2)}{n}$$

$$= \frac{\sum(y^2 - 2a_0y - 2a_1xy + a_0^2 + 2a_0a_1x + a_1^2x^2)}{n}$$

$$= \frac{\sum(y^2 - a_0y - a_1xy - a_0y - a_1xy + a_0^2 + a_0a_1x + a_1^2x^2 + a_0a_1x)}{n}$$

$$= \frac{\{\sum y(y - a_0 - a_1x) - a_0\sum(y - a_1x - a_0) - a_1\sum x(y - a_1x - a_0)\}}{n}$$

$y = a_0 + a_1x$ *by least square method to find* a_0 *and* a_1,

we solve $\sum y = a_0n + a_1\sum x$ *and* $\sum xy = a_0\sum x + a_1\sum x^2$

$\Rightarrow \sum(y - a_0 - a_1x) = 0, \sum(xy - a_0x - a_1x^2) = 0$

$$S_{y,x}^2 = \frac{\sum y(y - a_0 - a_1x)}{n} = \frac{\sum y^2 - a_0\sum y - a_1\sum xy}{n}$$

Eg: 2) If the regression line of y on x is y = 35.82 + 0.476x and values of x and y are

x	65	63	67	64	68	62	70
y	68	66	68	65	69	66	68

Find $S_{y,x}^2$.

solution: y = 35.82 + 0.476x

x	y	y_1= 35.82+0.476x	$(y - y_1)$	$(y - y_1)^2$
65	68	66.76	1.24	1.538
63	66	65.808	0.192	0.037
67	68	67.712	0.288	0.083
64	65	66.284	-1.284	1.649
68	69	68.188	0.812	0.659
62	66	65.332	0.668	0.446

70	68	69.14	-1.14	1.30
total				5.712

$$S_{y,x}^2 = \frac{\sum(y-y_1)^2}{n} = \frac{5.712}{7} = 0.816$$

Eg: 3) Show that $\sum(y-\bar{y})^2 = \sum(y-y_1)^2 + \sum(y_1-\bar{y})^2$

solution: $lhs = \sum(y-\bar{y})^2 = \sum(y-y_1+y_1-\bar{y})^2$

$= \sum\{(y-y_1)^2 + 2(y-y_1)(y_1-\bar{y}) + (y_1-\bar{y})^2\}$

$= \sum(y-y_1)^2 + \sum(y_1-\bar{y})^2 + 2\sum(y-y_1)(y_1-\bar{y})$

$nowtoprove, \quad \sum(y-y_1)(y_1-\bar{y}) = 0$

i. e $\sum(y-y_1)(y_1-\bar{y}) = \sum(y-a_0-a_1x)(a_0+a_1x-\bar{y})$

$= a_0\sum(y-a_0-a_1x) + a_1\sum x(y-a_0-a_1x) - \bar{y}\sum(y-a_0-a_1x) = 0$

$y = a_0 + a_1x, usingleastsquaremethod$ we $finda_0anda_1,$

$wesolve \sum y = a_0n + a_1\sum xand \sum xy = a_0\sum x + a_1\sum x^2$

$\Rightarrow \sum(y-a_0-a_1x) = 0, \sum(xy-a_0x-a_1x^2) = 0$

$\Rightarrow \sum(y-\bar{y})^2 = \sum(y-y_1)^2 + \sum(y_1-\bar{y})^2$

13.5 EXPLAINED AND UNEXPLAINED VARIATION

In the previous section we had proved,

$\sum(y-\bar{y})^2 = \sum(y-y_1)^2 + \sum(y_1-\bar{y})^2$

$\sum(y-y_1)^2$ is called unexplained variation as the deviation $(y-y_1)$ behave randomly and have an unpredictable manner.

$\sum(y_1-\bar{y})^2$is called explained variation as the deviation$(y_1-\bar{y})$have a definite pattern.

$\sum(y-\bar{y})^2$is called as total variation.

Eg: 1) For the following calculate total variation, explained variation and unexplained variation, given the regression of on x as y = 35.82 + 0.476x

x	65	63	67	64	68	62	70
y	68	66	68	65	69	66	68

Solution: y = 35.82 + 0.476x, $\bar{y} = \sum\frac{y}{n} = 67.143$

x	y	y_1= 35.82+0.476x	(y –y_1)	$(y-y_1)^2$	$(y_1-\bar{y})^2$
65	68	66.76	1.24	1.538	0.147
63	66	65.808	0.192	0.037	1.306
67	68	67.712	0.288	0.083	0.734
64	65	66.284	-1.284	1.649	4.592
68	69	68.188	0.812	0.659	3.448

62	66	65.332	0.668	0.446	1.306
70	68	69.14	-1.14	1.30	0.734
total				5.712	12.267

Unexplained variation= $\sum(y - y_1)^2 = 5.712$

Explained variation =$\sum(y_1 - \bar{y})^2 = 12.267$

Total variation = $\sum(y - \bar{y})^2 = \sum(y - y_1)^2 + \sum(y_1 - \bar{y})^2 = 17.979$

13.6 COEFFICIENT OF DETERMINATION

The ratio of the explained variation to the total variation is called the coefficient of determination.
It is given by $\frac{\sum(y_1 - \bar{y})^2}{\sum(y - \bar{y})^2}$.

If there is zero explained variation (i.e the total variation is same as unexplained variation) then coefficient of determination will be 0. If there is zero unexplained variation (i.e the total variation is same as explained variation) then coefficient of determination will be 1. Since the ratio is always non negative, we denote it by r^2. $r$called as coefficient of correlation is given by

$$r = \pm\sqrt{\frac{explained\ variation}{total\ variation}} = \pm\sqrt{\frac{\sum(y_1 - \bar{y})^2}{\sum(y - \bar{y})^2}}$$

varies from between -1 and +1.The + and – signs indicates positive and negative linear correlation respectively. Note $r$is a dimensionless quantity, i.e it does not depend on the units.
For the case of linear correlation, the quantity r is the same regardless of whether X or Y is considered the independent variables. Thus r is a good measure of the linear correlation between two variables.

Eg : 1) Show that $s_{y,x}^2 = s_y^2(1 - r^2)$.

Solution: $r = \pm\sqrt{\frac{explained\ variation}{total\ variation}} = \pm\sqrt{\frac{\sum(y_1 - \bar{y})^2}{\sum(y - \bar{y})^2}}$

$$r^2 = \left(\sqrt{\frac{\sum(y_1 - \bar{y})^2}{\sum(y - \bar{y})^2}}\right)^2 = \left(\sqrt{\frac{\sum(y - \bar{y})^2 - \sum(y - y_1)^2}{\sum(y - \bar{y})^2}}\right)^2$$

$$\Rightarrow r^2 = \left(\sqrt{1 - \left(\frac{\sum(y - y_1)^2}{\sum(y - \bar{y})^2}\right)}\right)^2 = \left(\sqrt{1 - \frac{s_{y,x}^2}{S_y^2}}\right)^2$$

$$\Rightarrow s_{y,x}^2 = s_y^2(1 - r^2)$$

Eg: 2) Given the regression of y on x as y = 35.82 + 0.476x. Find the coefficient of determination and coefficient of correlation for the following data

x	65	63	67	64	68	62	70
y	68	66	68	65	69	66	68

Solution: y = 35.82 + 0.476x, $\bar{y} = \sum \frac{y}{n} = 67.143$

x	y	y_1= 35.82+0.476x	(y –y_1)	$(y - y_1)^2$	$(y_1 - \bar{y})^2$
65	68	66.76	1.24	1.538	0.147
63	66	65.808	0.192	0.037	1.306
67	68	67.712	0.288	0.083	0.734
64	65	66.284	-1.284	1.649	4.592
68	69	68.188	0.812	0.659	3.448
62	66	65.332	0.668	0.446	1.306
70	68	69.14	-1.14	1.30	0.734
total				5.712	12.267

Unexplained variation $= \sum(y - y_1)^2 = 5.712$

Explained variation $= \sum(y_1 - \bar{y})^2 = 12.267$

Total variation $= \sum(y - \bar{y})^2 = \sum(y - y_1)^2 + \sum(y_1 - \bar{y})^2 = 17.979$

Coefficient of determination $= \frac{\sum(y_1 - \bar{y})^2}{\sum(y - \bar{y})^2} = \frac{12.267}{17.979} = 0.682$

Coefficient of correlation $r = \pm\sqrt{\frac{explained\ variation}{total\ variation}} = \pm\sqrt{0.682}$

13.7 SUMMARY

In this chapter we have learnt about the scatter diagram which helps us to find the nature and extent of relationship between the variables, Coefficient of correlation which is a numerical measure of nature and extent of relationship between two given variables whose values lies between +1 and -1. We had also learnt about the coefficient of Rank correlation which is used in cases where it is not possible to get numerical measurements, but we can rank the individuals in order according to our judgment. This chapter also deals with the least square regression lines of y on x as well as x on y. This chapter also explains about the standard error of estimate, Explained and unexplained variation and correlation of time series.

13.8 EXERCISES

1) Following table gives the us the marks obtained by 6 students in mid term exam and final semester exam. Find the coefficient of correlation.

Mid term Exam	12	14	23	18	10	19
Final semester Exam	68	78	85	75	70	74

2) Given covariance = 27, r = 0.6, variance of y = 25. Find variance of x.

3) Find the coefficient of correlation between the heights of male students and female students

height of male students	65	66	67	68	69	70	71
height of female students	67	68	66	69	72	72	69

4) If$r = 0.38, cov(x, y) = 10.2,\ \sigma_x = 16. Find\ \sigma_y$

5) Find r, if $cov(x, y) = 6, \sigma_x = 2.45, \sigma_y = 3.41.$

6) Find r, given$\sum(x - \bar{x})(y - \bar{y}) = 29, \sum(x - \bar{x})^2 = 28, \sum(y - \bar{y})^2 = 42.$

7) Ten competitors in Miss Universe are ranked by three judges in the following order:

J1	1	6	5	10	3	2	4	9	7	8
J2	3	5	8	4	7	10	2	1	6	9
J3	6	4	9	8	1	2	3	10	5	7

Using rank correlation coefficient determine which pair of judges has the nearest approach to common tastes in beauty.

8) The coefficient of rank correlation for certain data is found to be 0.6. If the sum of the squares of the differences is given to be 66. Find the number of items in the group.

9) The ranks of 16 students in the subject of DBMS and CN are given as follows. Calculate the rank coefficient of correlation.

Rank in DBMS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Rank in CN	1	10	3	4	5	7	2	6	8	11	15	9	14	12	16	13

10) The coefficient of rank correlation of marks obtained by 9 students was calculated to be 0.4. It was later discovered that the value of the difference between ranks for one student was written wrongly as 6 instead of 8. Find the correct value of coefficient of rank correlation.

11) Find the two regression equations given $\bar{x} = 8, \bar{y} = 2000, \sigma_x = 2, \sigma_y = 80, r = 0.7.$

Also estimate y when x = 10 and estimate x when y = 2500.

10

12) Find the two regression equations for the following data: $\sum(x-\bar{x})(y-\bar{y}) = 135, \sum(x-\bar{x})^2 = 96, \sum(y-\bar{y})^2 = 206, \sum x = 120,$
$\sum y = 180, n = 5$

13) n = 50, regression equation of marks in mathematics (Y) on the marks in English(X) was 4Y- 5X = 8.Mean marks in English are 40. The ratio of the two standard deviation $\sigma_y : \sigma_x = 5:2$. Find the average marks in mathematics and coefficient of correlation between the marks in the two subjects.

14) The two regression lines between x and y are given below$2x + 3y = 61$,
$x + y = 25$. Find$\bar{x}, \bar{y}$ and r.

15) Find the regression line of profits on output from the following data using least square method.

Ouput(100 tons)	5	7	9	11	13	15
Profit per unit(Rs.)	1.7	2.4	2.8	3.4	3.7	4.4

16) For the following find total variation given the regression of y on x as y = 35.82 + 0.476x

x	65	63	67	64	68	62	70	66	68	67	69	71
y	68	66	68	65	69	66	68	65	71	67	68	70

13.9 SOLUTION TO EXERCISE

Q. No.	Solution	Q. No.	Solution
1	r = 0.81	2	variance of x = 81
3	r = 0.67	4	1.68
5	r = 0.72	6	r = 0.85
7	R_{12} = -0.212, R_{13} = 0.636, R_{23} = -0.297	8	n= 10
9	R = 0.8	10	R = 0.17
11	y = 28x +1776, x = 0.02y – 32, y = 2056, x = 18	12	y = 1.41x + 2.16, x = 0.66y +0.24
13	52, r = 0.5	14	14, 11, -0.82
15	y = 0.26x + 0.5	16	38.917

13.10 REFERENCES

Following books are recommended for further reading:

- Statistics by Murray R, Spiegel, Larry J. Stephens, Mcgraw Hill International Publisher, 4th edition
- Fundamental of Mathematical Statistics by S. C. Gupta and V. K. Kapoor, Sultan Chand and Sons publisher, 11th edition
- Mathematical Statistics by J. N. Kapur and H. C. Saxena, S. Chand publisher, 12th edition

*****